TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Informatik XIX

Automated Semantic Analysis, Legal Assessment, and Summarization of Standard Form Contracts

Daniel Braun

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		apl. Prof. Dr. Georg Groh
Prüfer der Dissertation:		
	1.	Prof. Dr. Florian Matthes
	2.	Prof. Dr. Burkhard Schäfer

Die Dissertation wurde am 14.01.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 25.05.2021 angenommen.

Abstract

Consumers are confronted with standard form contracts on a daily basis, for example, when shopping online, registering for online platforms, or opening bank accounts. With expected revenue of more than 343 billion Euro in 2020, e-commerce is an ever more important branch of the European economy. Accepting standard form contracts often is a prerequisite to access products or services, and consumers frequently do so without reading, let alone understanding, them. Consumer protection organizations can advise and represent consumers in such situations of power imbalance. However, with increasing demand, limited budgets, and ever more complex regulations, they struggle to provide the necessary support.

This thesis investigates techniques for the automated semantic analysis, legal assessment, and summarization of standard form contracts in German and English, which can be used to support consumers and those who protect them. We focus on Terms and Conditions from the fast growing market of European e-commerce, but also show that the developed techniques can in parts be applied to other types of standard form contracts.

We elicited requirements from consumers and consumer advocates to understand their needs, identified the most relevant clause topics, and analyzed the processes in consumer protection organizations concerning the handling of standard form contracts. Based on these insights, a pipeline for the automated semantic analysis, legal assessment, and summarization of standard form contracts was developed. The components of this pipeline can automatically identify and extract standard form contracts from the internet and hierarchically structure them into their individual clauses. Clause topics can be automatically identified, and relevant information can be extracted. Clauses can then be legally assessed, either using a knowledge-base we constructed or through binary classification by a transformer model. This information is then used to create summaries that are tailored to the needs of the different user groups. For each step of the pipeline, different approaches were developed and compared, from classical rule-based systems to deep learning techniques. Each approach was evaluated on German and English corpora containing more than 10,000 clauses, which were annotated as part of this thesis. The developed pipeline was prototypically implemented as part of a web-based tool to support consumer advocates in analyzing and assessing standard form contracts. The implementation was evaluated with experts from two German consumer protection organizations with questionnaires and task-based evaluations.

The results of the evaluation show that our system can identify over 50 different types of clauses, which cover more than 90% of the clauses typically occurring in Terms and Conditions from online shops, with an accuracy of 0.80 to 0.84. The system can also automatically extract 21 relevant data points from these clauses with a precision of 0.91 and a recall of 0.86. On a corpus of more than 200 German clauses, the system was also able to assess the legality of clauses with an accuracy of 0.90. The expert evaluation has shown that the system is indeed able to support consumer advocates in their daily work by reducing the time they need to analyze and assess clauses in standard form contracts.

Zusammenfassung

Verbraucher sind täglich mit Allgemeinen Geschäftsbedingungen (AGB) konfrontiert, zum Beispiel beim Online-Shopping, beim Registrieren auf Online-Plattformen oder beim Eröffnen eines Kontos. Mit einem erwarteten Umsatz von 343 Mrd. Euro in 2020, ist der Onlinehandel ein immer wichtigerer Wirtschaftszweig in Europa. Das Akzeptieren von AGB ist hier meist Voraussetzung für den Zugang zu Waren und Dienstleistungen. Häufig tun Verbraucher dies ohne diese zu verstehen oder auch nur zu lesen. Verbraucherschutzorganisationen können Verbraucher, in solchen Situationen des Kräfteungleichgewichts, vertreten und beraten. Wegen steigender Bedarfe, limitierten Budgets und immer komplexeren Regulierungen, haben sie aber häufig Probleme die notwendige Unterstützung zu bieten.

Diese Dissertation untersucht Techniken zur automatischen semantischen Analyse, rechtlichen Bewertung und Zusammenfassung von AGB in Deutsch und Englisch, die genutzt werden können um Verbraucher und Verbraucherschützer zu unterstützen. Der Fokus liegt dabei auf AGB im schnell wachsenden Markt des europäischen Onlinehandels, aber wir zeigen auch, dass die entwickelten Techniken teilweise auf andere Arten von AGB übertragen werden können.

Um die Bedürfnisse von Verbrauchern und Verbraucherschützern besser zu verstehen, haben wir deren Anforderungen erhoben, relevante Klauseln identifiziert und die Prozesse innerhalb der Verbraucherzentralen in Bezug auf AGB analysiert. Basierend darauf haben wir eine Pipeline für die automatische semantische Analyse, rechtliche Bewertung und Zusammenfassung von AGB entwickelt. Die Komponenten dieser Pipeline können AGB automatisch identifizieren, extrahieren und die einzelnen Klauseln hierarchisch anordnen. Die Themen dieser Klauseln können dann automatisch erkannt werden und die Klauseln rechtlich bewertet werden, entweder mithilfe einer Wissensdatenbank oder durch binäre Klassifikation mit einem Transformer Model. Diese Informationen werden dann genutzt, um Zusammenfassungen zu erstellen, die auf die Bedürfnisse der jeweiligen Nutzergruppen angepasst sind. Für jede Stufe der Pipeline wurden verschiedene Ansätze, von klassischen regelbasierten Systemen bis hin zu Deep Learning Technologien, entwickelt und verglichen. Jeder der Ansätze wurde auf deutschen und englischen Korpora mit mehr als 10.000 Klauseln evaluiert, die im Rahmen dieser Arbeit annotiert wurden. Die entwickelte Pipeline wurde prototypisch implementiert als Teil eines Web-basierten Tools zur Unterstützung von Verbraucherschützern bei der Analyse und Bewertung von AGB. Die Implementierung wurde mit Experten von zwei deutschen Verbraucherzentralen mit einer aufgabenbasierten Evaluation und einem Fragebogen bewertet.

Die Ergebnisse zeigen, dass unser System über 50 verschiedene Klauseltypen, welche mehr als 90% der üblichen Klauseln in AGB von Onlineshops abdecken, mit einer Accuracy von 0,80 bis 0,84 erkennen kann. Das System kann aus diesen Klauseln außerdem 21 Datenpunkte, mit einer Precision von 0,91 und einem Recall von 0,86, automatisch extrahieren. Auf einem Corpus mit mehr als 200 deutschen Klauseln ist das System in der Lage, die Gültigkeit von Klauseln mit einer Accuracy von 0,90 zu erkennen. Die Expertenevaluation hat gezeigt, dass das System in der Lage ist Verbraucherschützer in ihrer täglichen Arbeit zu unterstützen, indem es die Zeit verringert, die diese zur Analyse und Bewertung von Klauseln in AGB benötigen.

Acknowledgment

The research presented in this thesis was enabled by generous funding from the German Federal Ministry of Education and Research (BMBF) grant A-SUM 01IS17049 and the Federal Ministry of Justice and Consumer Protection (BMJV) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support program. The Leibniz Supercomputing Centre generously provided computing time on its Linux-Cluster.

I want to thank my supervisor Prof. Dr. Florian Matthes, for providing both the freedom and the guidance I needed, for the trust he put in me and my research, and for the open and interdisciplinary research environment, he created at his chair, which enabled this thesis. I would also like to thank Prof. Dr. Burkhard Schäfer for agreeing to be the second examiner of this thesis and the impulses he provided.

Research is a team effort, and for the last four years, I was lucky enough to be part of a great team at the chair of Software Engineering for Business Information Systems. I want to thank my (former) colleagues Dr. Bernhard Waltl, Patrick Holl, and Elena Scepankova, with whom I developed the idea to work on consumer protection, and Oleksandra Klymenko and Nektarios Machner for their support in the writing process. I would also like to thank the former students who chose to work with me and contributed to my research, especially Kira Klimt, Jan Robin Geibel, and, again, Oleksandra Klymenko. I am also very grateful to our collaborators from the consumer protection agencies in Hamburg and Brandenburg, who made this research possible by sharing their expertise and knowledge with me.

Four years can feel like a lifetime, especially if they are accompanied by tragedy, from a global to a personal level. I want to thank my former colleagues Adrian Hernandez Mendez, Dr. Anne Faber, Klim Shumaiev, and Dr. Manoj Bhat, who, against my best efforts, enriched not just my research.

It was a long way from when my grandfather gave me my first computer and taught me how to program to this point. A way that I could only make because I was accompanied and supported by my family on every step. Above everything else, I would like to thank my parents, siblings, and grandparents for their unconditional love and support. There are no words for the gratitude I feel, nor for the sadness I feel, that not all of you are with me anymore for the rest of the way.

Table of Contents

List of Figures

List of Tables

Listings

CHAPTER 1

Introduction

Most aspects of our daily life are governed by law, from the energy consumption of the coffee machine we use to brew our morning coffee to the text on the label of the bed sheet we sleep in at night. These laws are drafted by local, state, national, and sometimes even supranational lawmakers, who are sworn in to act in the best interest of the people they represent. The oldest known law code is the “Code of Ur-Nammu” which is believed to have been created between 2100 BC and 2050 BC (Diller, 2012). Today, as then, there are still many aspects of our daily living together which are not regulated by any laws. To manage and codify inter-personal (and inter-organizational) relationships of all kinds, humans use contracts. The first contract known today precedes the Code Of Ur-Nammu by more than 200 years and is believed to have been drafted in Mesopotamia around 2300 BC (Molina-Jimenez et al., 2004).

The idealized notion of a contract most people have in mind is an agreement negotiated between two equal parties. In reality, however, most contracts consumers are facing today offer Hobson’s choice: sink or swim. So-called “standard form contracts” trace back to the 19th century. In the age of industrialization, entering into contracts has been accompanied by the unilateral use of pre-formulated rules tailored to one party’s interests, resulting in an imbalance of powers between the contracting parties. In the modern consumer society, standard form contracts are used by online shops, banks, insurances, telecommunication service providers, and many other businesses. (Zerres, 2014)

In acknowledgment of the imbalance of power such contracts introduce, many legislators have limited the creative leeway for companies when it comes to drafting standard form consumer contracts. An example of such legislation is the Council Directive 93/13/EEC of 5 April 1993 on unfair terms in consumer contracts from the Council of the European Union (1993).[1] Despite the existing guiding principles, standard form contracts still regularly overwhelm consumers in

[1] See Section 2.3 for more details on the regulatory frameworks in Germany and the European Union.

their daily life. In 2018, the Federation of German Consumer Organisations (Bundesverband der Verbraucherzentrale (vzbv)[2]) has sent a cease-and-desist order to the online payment service provider PayPal, because its German T&C consisted of more than 20,000 words or 1,000 sentences, the longest of which was constituted of 111 words. According to the vzbv, an average reader would need about 80 minutes to read these T&C. Moreover, the consumer advocates called the text "formally incomprehensible". (Verbraucherzentrale Bundesverband e.V., 2018a) The case was later brought to court but both, the regional court (Landgericht Köln, 2018) and the higher regional court of Cologne (Oberlandesgericht Köln, 2019), ruled that the T&C were lawful. This is just one example that, despite legislation to protect consumers and their rights, companies still can draft one-sided standard form contracts legally in a way which is, arguably, very difficult if not impossible to understand for consumers.

Another aspect that is fostering the imbalance of power between consumers and corporations is the unevenly distributed access to legal consultation. While most large corporations have in-house legal counsels (Association of Corporate Counsel, 2019), consumers often do not have direct access to legal advice and might shy away from consulting legal experts due to potentially high costs, especially in cases with small litigious values. An alternative, often more affordable, source of legal advice for consumers are Non Governmental Organisations (NGOs) for consumer protection. These organizations often struggle to cope with the growing demand from consumers due to insufficient funding. Digitization and automation could help to support such organizations in their work.

Many aspects of our lives have been influenced, and arguably democratized, by the digital revolution. Access to knowledge is no longer restricted to those who can afford 32 volumes of Encyclopædia Britannica or have access to university libraries. It is instead available to everyone with access to the internet. The access to the fine arts, but also to once expensive services like translation, has been opened up to new classes of citizens by digitization. Digitization and artificial intelligence are providing "opportunities for improving and augmenting the capabilities of individuals and society at large" (Floridi et al., 2018) in many different domains. For a long time, the legal profession was arguably one of the biggest resisters to digitization efforts and, in some aspects, still struggles to catch up with other industries and areas of life. In recent years, digitization has entered the legal profession as so-called "LegalTech". The term is a portmanteau word consisting of "legal services" and "technology". It is widely used as a description for the support or automation of legal processes with software or, more broadly, technology.

Unlike in the areas described before, digitization in the legal domain, so far, almost exclusively benefits big corporations and law firms. Most existing LegalTech tools are made for companies and law firms, rather than consumers (see Section 3.4). Thereby, LegalTech tools are not only missing the opportunity to democratize access to legal advice by making it more affordable and available, but they are also actively supporting the current imbalance of power between companies and consumers by providing companies with even more advantages over consumers.

In this thesis, we present research on the automated analysis, legal assessment, and summarization of standard form contracts in German and English. While some of the results of our research are also applicable in business-to-business contexts, the goal of this thesis is to further

[2] See Section 2.3.1 for more information on consumer protection organizations in Germany.

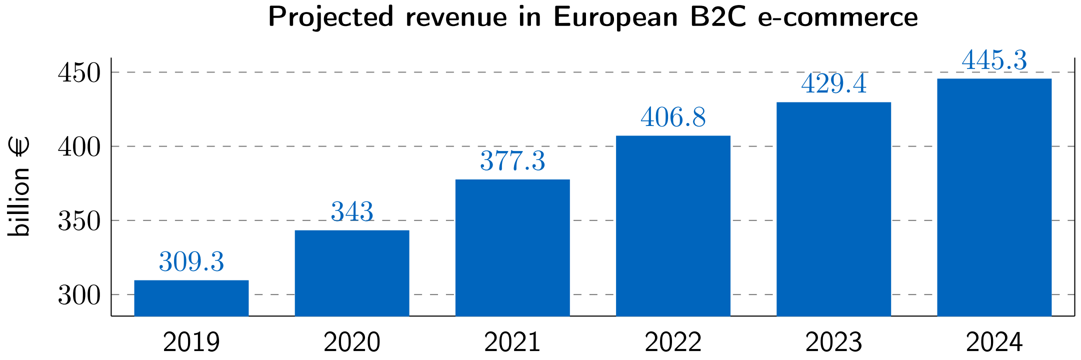


Figure 1.1.: Projected revenue in European B2C e-commerce in billion Euro; data source: statista (2020)

consumer protection by supporting consumers and those who represent their interests by applying Artificial Intelligence (AI), NLP, and Natural Language Generation (NLG) techniques to solve problems that consumer are facing with regard to standard form contracts.

For most parts of this thesis, we will focus on one particular kind of standard form contract, the so-called T&C (or Terms of Service (ToS)) from online shops. On the one hand, B2C e-commerce is one of the economically most important areas of consumer standard form contracts with an estimated net volume of 57.8 billion Euro in 2019 in Germany (Institut für Handelsforschung, 2019) and 309.3 billion in Europe (statista, 2020) and expected to grow further (see Figure 1.1). On the other hand, values at stake for individual consumers are regularly very small, which increases the hesitation to seek professional legal advice and therefore increases the need for automated solutions. Additionally, T&Cs are publicly available in sufficiently large amounts (in a machine-readable format and different languages), which is helpful, e.g., in order to train Machine Learning (ML) classifier on them. This is not the case for other forms of consumer standard form contracts, like loan or lease agreements.

In German legislation, the term equivalent to T&C is defined by the German civil code (Bürgerliches Gesetzbuch (BGB)) as Allgemeine Geschäftsbedingungen (AGB) (§ 305 Abs. 1 BGB)[3]. In the Anglo-American legal sphere, the two terms, Terms and Conditions and Terms of Service, can be used interchangeably. The general convention is to use the term ToS for contracts applying to software, or more generally digital products, and T&C in the other cases. For the ease of reading, throughout the remainder of the thesis, Terms and Conditions (T&C) will be used to address all three Forms T&C, ToS, and the German AGB. This usage of the term can be subsumed under the definition given by § 305 Abs. 1 BGB:

Definition 1. *Terms and Conditions (T&C) are all contract terms pre-formulated for a multitude of contracts that are posed by one party to the other when entering a contract.*[4]

While the general focus will be on T&C from online shops, we will show at various points of

[3] A note on citations: When referring to German laws, we will use the citation format that is commonly used in legal literature, using the paragraph and the abbreviation of the law.

[4] Analogous translation by the author.

this thesis that the findings can also be generalized to other types of T&C and standard form contracts, e.g., from banks.

1.1. Problem Description

The main problem that motivated this thesis is the previously mentioned existing imbalance of power between consumers and corporations when entering a contractual relationship based on a standard form contract. In this section, we will describe the existing problems that consumers (Section 1.1.1) and consumer advocates (Section 1.1.2) are currently facing more concretely.

Standard form contracts inherently have the characteristic of being drafted unilaterally and are regularly putting consumers in a Hobson's choice situation. One could, therefore, argue that they are "flawed beyond repair". However, our modern economy is unimaginable without them: Netflix cannot negotiate individual terms with each of their 182 million clients or Amazon with their 150 million Prime members. We will, therefore, focus on problems that consumers are facing when they want to make informed decisions about T&C or exercise their rights against possibly illegal standard form contracts and problems that consumer advocates face when they support consumers in this process.

1.1.1. Consumer Perspective

The model of the consumer set by the European Union (EU) originates from the idea of a "reasonably well-informed and reasonably observant and circumspect" (European Court of Justice, 2014) consumer. However, evidence suggests that the assumption of a consumer who is "reasonably well-informed" made by the European Court of Justice (ECJ) is not applicable in the context of T&C of online shops because the average consumer is not well or sometimes at all informed about the T&C they agree to. Even among law students, which do arguably not represent average consumers, only 3.5% (Plaut and Bartlett III, 2012) to 4% (Hillman, 2005) *claim* to regularly read T&C and 16.5% at least *claim* to skim them (Plaut and Bartlett III, 2012). In an experiment with real customers of software online shops, Bakos et al. (2014) found that only 0.11% of all buyers *opened* the T&C before proceeding to the checkout page. The average time (mean) these customers spend with the T&C open was 84 seconds (60 seconds median). Since the T&C in the experiment had an average length of 2,277 words, an average reader with an assumed reading speed of 250 to 300 words per minute would have needed eight to ten minutes to actually read the T&C. Therefore, the first problem consumers face is that they do not know what T&C they agree to because they either do not read them at all or do not read them completely.

Legal language has its own rules and pragmatics, whether it is English (Marmor, 2008) or German (Felder and Vogel, 2017). Therefore, legal language is often hard to understand for people without legal training. This leads to the second problem that consumers are facing: Even if they do read the T&C, they often have a hard time understanding them. Plaut and Bartlett III (2012) report that people who actually read T&C have only a slightly higher (and absolutely still low) confidence about knowing their content than people who did not even bother

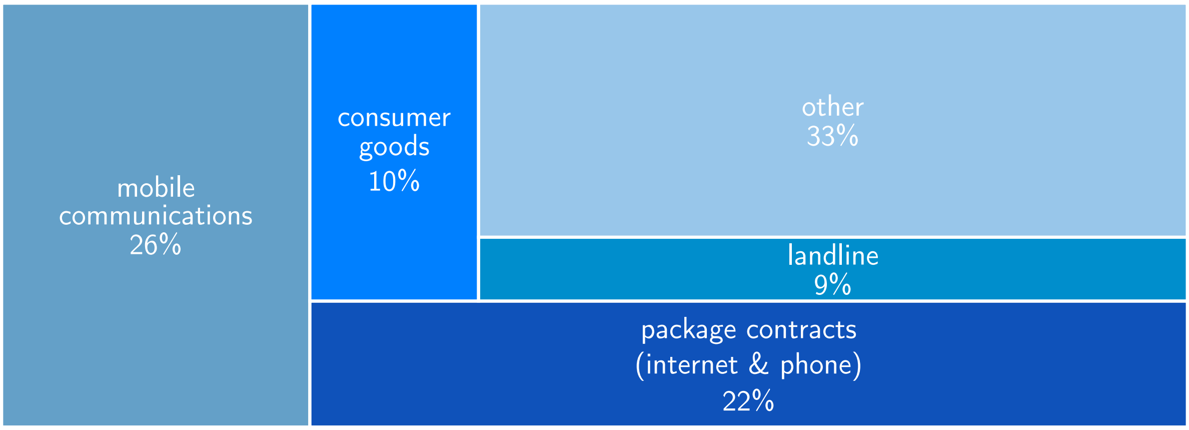


Figure 1.2.: Areas of complaints to consumer protection agencies in Germany about digital offerings in 2018; data source: Verbraucherzentrale Bundesverband e.V. (2019a)

reading them. Obar and Oeldorf-Hirsch (2020) found in their lab experiment that, although 26% of participants have looked at the T&C, only 2 % noticed that there was a clause hidden that participants agree to share all their data with the NSA.

Both problems lead to consumers making uninformed decisions with potentially negative consequences for them. Once consumers get into such a situation, they often need professional legal advice, which leads to a third problem consumers are facing. Initial consultation with a solicitor can cost up to 190 Euro for consumers in Germany (§ 34 Rechtsanwaltsvergütungsgesetz (RVG)) which is limiting access to legal advice. Many consumers, therefore, turn to consumer advocates for advice; however, these organizations are facing problems themselves.

1.1.2. Consumer Advocate Perspective

In many countries, consumer advocates and consumer protection agencies are chronically underfunded. In 2019, the government-funded Verbraucherzentrale Bremen e.V. (consumer protection agency Bremen) even had to file for bankruptcy (Der Spiegel, 2019). With their limited financial means, consumer advocates all over Europe struggle to keep up with the demand generated by the increasing importance of digital offerings. In 2018, the consumer protection agencies ("Verbraucherzentralen") in Germany received in total 184,579 complaints from consumers. 65,370 of these complaints (more than 35%) were related to digital offerings (see Figure 1.2). In comparison, only 36,945 complaints (20%) were received about products and services from the financial industry (Verbraucherzentrale Bundesverband e.V., 2019a). The problem consumer advocates are facing is that they struggle to keep up with the demand for individual counseling from consumers. In addition, consumer advocates increasingly try to monitor (digital) markets proactively and react to negative developments before consumers are harmed. In Germany, the consumer protection agencies have pooled these efforts under the "Marktwächter" (market watchman) program (Hörmann, 2017). Monitoring markets as big as eCommerce and proactively act against void clauses in standard form contracts is, at scale, simply not possible without automation of the underlying processes.

1.2. Research Questions

Based on the problem description and the experience with AI technology in other fields of application, we formulated the following research hypothesis which guided us during our research:

> ***Research Hypothesis:*** AI can help consumers and consumer advocates in finding, analyzing, and assessing standard form contracts.

The term "AI" might seem broad and unspecific in this context, however, in the course of this thesis, we will use a plethora of approaches from fields like NLP, NLG, expert systems, and knowledge representation and a range of technologies from basic Boolean logic to deep neural networks, that can only be summarized under the admittedly broad term of AI.

In order to test this hypothesis and address the previously described problems, we formulated eight research questions. These questions guided our research and will be answered in this thesis.

> ***Research Question RQ1:*** What are existing technical approaches to the semantic analysis, legal assessment, and summarization of legal documents?

In order to align our research with the existing body of knowledge and show where our work adds to it, the related scientific literature was analyzed. In addition, we also analyzed the market for commercial software for contract analysis in order to answer this research question.

> ***Research Question RQ2:*** Which clauses in standard form contracts are most relevant from a consumer and consumer advocate perspective?

Since it is virtually impossible to cover all legal aspects of standard form contracts, one has to define a subset of clauses on which to focus. Therefore, we wanted to identify which types of clauses in standard form contracts are most relevant from a consumer and from a consumer advocate perspective. In this context, relevance can be influenced by different factors, e.g., whether consumers themselves specifically care about a certain clause, whether a clause appears in many different types of standard form contracts or whether a clause is known to be void in many cases.

> ***Research Question RQ3:*** How do consumer advocates work with standard form contracts?

Before we can find out how to best support consumer advocates in their work (RQ4), we first have to analyze their current modus operandi by investigating typical tasks and workflows. The integration into existing workflows and the automation of presumably simple tasks are often perceived as very helpful by users, therefore, the thorough analysis of existing workflows is very important to identify such potentials.

> ***Research Question RQ4:*** How should software be designed to support consumer advocates in their efforts to protect consumers from void clauses in standard form contracts?

Based on the insights gained from the previous three research questions, a software is designed that is able to support the work of consumer advocates. As part of this question, the requirements of the prospective users are analyzed in interviews and workshops. Additionally, a software architecture that fosters reusability of the different components is drafted. Software development, and the development of research prototypes even more so, is an iterative, or “agile”, process. The final design of the software is, therefore, the result of many iterations involving various stakeholders.

Although we want to support consumers and consumer advocates in the same task (understanding the content of a standard form contract), due to their different background knowledge and contexts, they have very different requirements. Therefore, we analyze separately how consumers can be supported by software.

> ***Research Question RQ5:*** Which methods can be used to semantically analyze standard form contracts in German and English?

The core task, from a technological point of view that has to be achieved in order to provide the desired support and automation features, is the semantic analysis of the standard form contracts in German and English. It is, therefore, crucial to identify suitable methods for this task. Different methods have different advantages and disadvantages and decisions are always a trade-off between them. Different types of clauses and different types of situations may also benefit from different methods. The answer to this research question will therefore not be a single method, but a set of methods for different tasks, from low-level pre-processing tasks, like content extraction, to high-level semantic analysis tasks like clause topic classification.

> ***Research Question RQ6:*** How can the body of law governing the drafting of standard form contracts in Germany and the EU be formalized and represented in order to enable the automated legal assessment of standard form contracts?

In order to assess the lawfulness of a contract, a software does not only need to “understand” its content, but it also needs an understanding of the law that is governing the clauses of standard form contracts and their drafting. It is important to stress that we do not try to find a universal language or approach to the formalization of the law, a goal that is pursued by legal scholars for centuries. We rather look for a lightweight solution that allows us to formalize the applicable regulations for standard form contracts in Germany and the EU (see Section 2.3.1.1 and 2.3.2.1) in a way that supports the automated legal assessments of their clauses.

> ***Research Question RQ7:*** Which methods can be used to automatically summarize German and English standard form contracts?

The automatic summarization of texts is a sub-field of NLP with a large body of existing research (see Section 3.3). Depending on the approach, automatic summarization does not necessarily depend on the semantic analysis of a text and can, therefore, be separate from RQ5. Legal texts in general and standard form contracts specifically pose unique challenges to summarization approaches. We evaluate which of them is most suitable for this specific type of document. The requirements for a summarization also depend heavily on the recipient. For a consumer advocate, a very condensed concatenation of facts can be a suitable summary. In contrast, a consumer might need a simplified version of the original text in order to be able to understand all aspects of it.

> ***Research Question RQ8:*** Can AI-based automation and support tools improve the way consumers and consumer advocates work with standard form contracts?

The research presented in this thesis is mainly driven by the existing challenges that consumers and consumer advocates are facing every day. The success of the developed solutions can, therefore, not solely be evaluated with scientific metrics. Even if a system achieves perfect F1-scores and accuracy in certain tasks, e.g., the classification of clause topics, that does not necessarily imply that a practical utility arises from such a system. In order to test the research hypothesis we initially formulated, this practical utility has to be evaluated.

While scientific evaluation methods are standardized and well-defined, it is still an open research question how to evaluate the practical value of any NLP-based automation or support tool. We conducted questionnaire-based as well as task-based evaluations.

1.3. Methodology

The research presented in this thesis is multidisciplinary and combines research in the area of software engineering, (computational) linguistics, and law. Since this thesis is motivated by real-world problems, we mainly apply empirical research methods.

From a methodological perspective, the thesis can be divided into three parts, according to the disciplines they cover: law (Chapter 2, 4, and 7), computational linguistic (Chapter 6 to 8), and software engineering (Chapter 5, 9, and 10).

Our research on the legal aspects of the topic is mainly based on qualitative methods. We use expert interviews and literature research to gather the necessary external expertise. For the linguistic aspects, we mainly apply quantitative methods. We conducted a multitude of controlled experiments using existing and newly created corpora, which are based on legal documents.

For the engineering of our research prototype, we followed the approach of Action Research (Hult and Lennung, 1980). The term Action Research was initially coined by Lewin in his 1946 article "Action research and minority problems" and was developed to be a research methodology for the social sciences. Over the years, it has become popular in other domains like education (McNiff, 1993) and healthcare (Koshy et al., 2010). In software engineering, Action Research only slowly gained traction. In a survey dos Santos and Travassos (2009) conducted, they found that only 16 papers had been published in major software engineering conferences and journals using Action Research by 2009.

O'Leary defines Action Research as:

> *"Research strategies that tackle real-world problems in participatory and collaborative ways in order to produce action and knowledge in an integrated fashion through a cyclical process. In action research, process, outcome and application are inextricably linked."* (O'Leary, 2017, p. 350)

In the context of software engineering and information systems, Avison et al. describe the approach as follows:

> *"In action research, the researcher wants to try out a theory with practitioners in real situations, gain feedback from this experience, modify the theory as a result of this feedback, and try it again. Each iteration of the action research process adds to the theory [...] so it is more likely to be appropriate for a variety of situations."* (Avison et al., 1999, p. 95)

According to O'Leary (2017), Action Research can be broken down into four key elements:

- *Addresses Real-World Problems*
  "Action research is grounded in real problems and real-life situations. It generally begins with the identification of practical problems in a specific real-world context."
- *Pursues Action and Knowledge*
  "Action research rejects the two-stage process of 'knowledge first, change second', and suggests that they are highly integrated. Action research practitioners believe that enacting change should not just be seen as the end product of knowledge; rather it should be valued as a source of knowledge itself."
- *Participation*
  "The notion of research as the domain of the expert is rejected, with action research calling for participation of, and collaboration between, researchers, practitioners and any other interested stakeholders."
- *Cycles of Learning and Action*
  "Action research is a cyclical process that takes shape as knowledge emerges. The premise here is that you learn, you do, you reflect, you learn how to do better, you do it better, you learn from that, do it better still, and so on. You work through a series of continuous improvement cycles that converge towards better situation understanding and improved action."

(O'Leary, 2017)

We address all of these key elements as part of the software engineering process for our research prototype: During the requirement elicitation, we show that we *address a real-world problem.* By providing experts with tool support and analyze how they use it, we *pursue action and knowledge* in parallel. During the entire process, we *work closely with practitioners* from German consumer protection agencies. And last but not least, modern software development is always an iterative process, representing a *cycle of learning and action.*

On an operational level, Stringer (2014) describes a basic Action Research routine as follows:

Look

- *Gather relevant information*
- *Describe the situation*

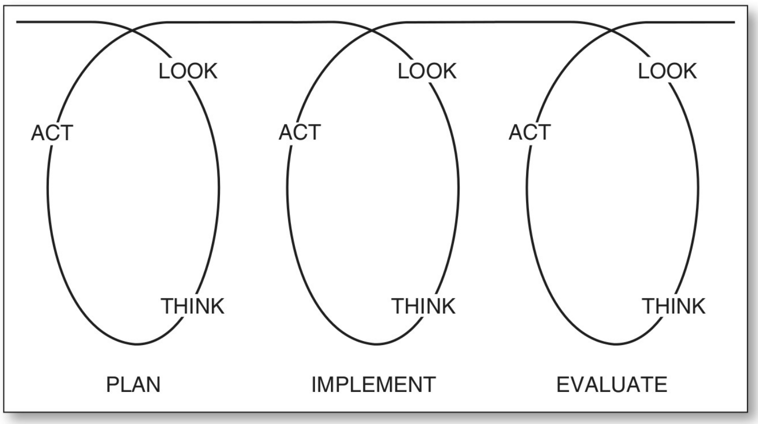


Figure 1.3.: Action research interacting spiral from Stringer (2014)

Think

- *Explore and analyze: What is happening here? (Analyze)*
- *Interpret and explain: How or why are things as they are? (Theorize)*

Act

- *Plan: Define a course of action based on analysis and interpretation.*
- *Implement: Implement the plan.*
- *Evaluate: Assess the effectiveness of actions taken.*

(Stringer, 2014, p. 7)

The three main activities, look, think, and act, are repeatedly execute in cycles or spirals. Stringer (2014) develops a model which consists of three main repetition phases: plan, implement, and evaluate (see Figure 1.3). We take this model and adapt it to reflect an iterative software engineering process. Although there are plenty of models for the process of software development, most of them contain three main steps in some form: requirements analysis, implementation, and testing or evaluation. In the classical waterfall model (Petersen et al., 2009), each of these steps is conducted once. In an iterative software development process, however, each of these steps is repeatedly executed for each iteration. One first plans the changes, then implement them and finally test them. These iterations usually happen in relatively short time frames of a few weeks.

In addition, there is usually a more elaborate requirements elicitation and also evaluation process at the beginning and end of a project or major release of a software. In those instances, each of these activities can be conducted for multiple weeks instead of just a few weeks for a whole iteration including all activities. Based on these steps, we developed an activity model for software-driven action research, which is shown in Figure 1.4.

In this thesis, the first planning-cycle is described in Chapter 5. The implementation cycle is described in Chapter 9, and the final evaluation cycle is described in Chapter 10.

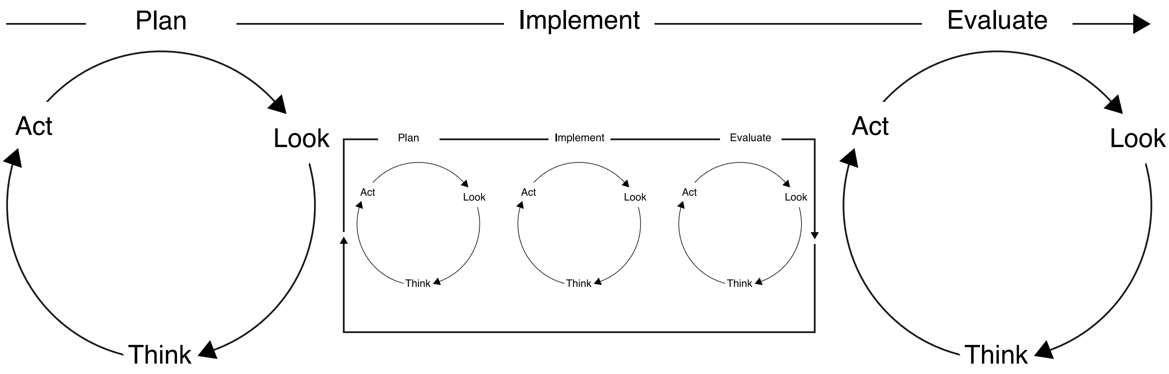


Figure 1.4.: Activity model for software-driven action research

1.4. Contributions

The results and insights gained in the course of this thesis contribute to science but also to the practice of consumer advocacy. Our contributions come from three different categories: We contribute to the scientific body of knowledge, provide linguistic resources to scientists and practitioners and develop software artifacts in the form of libraries and research prototypes. The contributions of this thesis, which are described in detail in the course of the following chapters, are:

Knowledge

- A taxonomy of clause topics and subtopics in standard form contracts.
- A list of the most relevant clauses in standard form contracts from a consumer and consumer advocate perspective.
- An analysis of the processes consumer advocates conduct to battle void clauses in standard form contracts.
- A literature review on the state-of-the-art in the automatic analysis, assessment, and summarization of legal texts (Klymenko et al., 2020), as well as an overview of commercial tools in the area of contract analysis and T&C generation.
- An evaluation of the performance of popular Natural Language Understanding (NLU) services. (Braun et al., 2017a)
- A comparison of NLP-techniques for the semantic analysis, legal assessment, and summarisation of standard form contracts in German and English. (Braun et al., 2017b, 2018c, 2019d,c; Braun and Matthes, 2020)

Linguistic Resources

- An English corpus of more than 450 sentences annotated with intents and entities for the evaluation of NLU services. (Braun et al., 2017a; ISLRN[5]165-571-578-116-6)
- A German lexicon for surface realisation with more than 100,000 lemmata based on Wiktionary. (Klimt et al., 2020; ISLRN 206-939-257-359-6)
- A list of 16,000 pages from more than 5,000 German and British online shops annotated by their page type (T&C, landing or other page).
- A corpus of 142 German and 30 English T&C from online shops, containing 5,020 German and 1,040 English clauses that were manually annotated with their topics and subtopics, based on the taxonomy we developed.
- A corpus of 1,185 German and 193 English clauses, that were manually annotated with a legal assessment made by two consumer advocates.

Software Artefacts

- A set of tools for the automated evaluation of NLU services. (Braun et al., 2017a)
- A knowledge-base containing formalised representations of relevant German laws regulating the drafting of standard form contracts.
- A surface realiser for German, based on SimpleNLG. (Braun et al., 2019b)
- A complete NLP-pipeline for the automated detection, semantic analysis, legal assessment, and summarization of standard form contracts.
- A web-based research prototype to support consumer advocates in analyzing and assessing void clauses in standard form contracts.

In order to increase the impact of our research and make the results transparent and replicable, we publish as much of our results as possible under open licenses. Unfortunately, we can not publish the corpora containing clauses from T&C, because of intellectual property and liability concerns. Appendix C presents a list of the results we published under open licenses.

[5] The International Standard Language Resource Number (ISLRN) is a unique identifier for language resources (comparable to ISBN numbers for books) supported by the European Language Resources Association (ELRA).

1.5. Outline

This thesis consists of eleven chapters. Table 1.1 shows a brief overview of these chapters and how they relate to the aforementioned research questions and contributions, as well as the key publications related to these chapters. This section briefly outlines the content of the following chapters.

Chapter 2: Foundations
Chapter 2 outlines relevant NLP, statistic classification, and legal foundations. It introduces the statistical classifiers and some of the relevant NLP concepts used in the later parts of this work. This thesis is by no means a legal treatise and does not aspire to be. However, one can not talk about the detection of void clauses in standard form contracts without having at least a basic understanding of the regulations that govern the drafting of such contracts. The focus here is on the situation at the federal level in Germany and EU legislation. In addition, the legal foundations for the work of consumer protection organization in Germany and the EU are described.

Chapter 3: Related Work
In this chapter, we present relevant related work from other researchers and position our own work in relation to existing research. We focus on three related research areas: the semantic analysis of legal texts, knowledge representation in the legal domain, and text summarization. In addition, we give an overview of the market for commercial contract analysis tools and introduce two commercial T&C generators.

Chapter 4: Empirical Relevance of Standard Form Contracts
We show the empirical relevance of the investigated topic by analyzing the relevance of standard form contracts to the economy, consumers, and consumer advocates. We investigate which types of clauses are used most commonly in actual T&C and derive a taxonomy of relevant clause topics and subtopics from different sources. Based on this taxonomy, we analyze which clauses are most important to consumers and consumer advocates and why. Finally, we discuss the ethical implications a software system like the one described in this thesis can have.

Chapter 5: Requirements Identification
In Chapter 5, we report on the requirement elicitation process we conducted with two consumer protection agencies and ten consumers. We analyzed their present workflow and elicited their requirements towards tool support in a series of interviews over the course of two years.

Chapter 6: Semantic Analysis of Standard Form Contracts
In the core NLP part of this thesis, we investigate linguistic and technical issues in semantically analyzing (Chapter 6), legally assessing (Chapter 7), and summarizing (Chapter 8) standard form contracts. In the sixth chapter, we compare different techniques, from simple rule-based approaches to “classic” machine learning and deep learning, for the semantic analysis of standard form contracts. We report which techniques are most suitable for the different classes of problems we are tackling: automatic detection, segmentation (including

content extract, sentence segmentation, and paragraph segmentation), topic classification, and information extraction.

Chapter 7: Legal Assessment of Standard Form Contracts
In Chapter 7, we describe how the relevant governing law regarding standard form contracts can be formalized in a knowledge-base and how this knowledge-base can subsequently be used to assess the lawfulness of a clause after its semantic analysis. We compare this classical expert system approach with a black-box binary classification approach using transformer models.

Chapter 8: Summarization of Standard Form Contracts
In Chapter 8, we present an approach to the automated generation of summaries of standard form contracts. We investigate two target audiences for these summaries: consumer advocates, i.e., domain experts with in-depth knowledge, and consumers, i.e., laypersons without any special training or knowledge in the legal domain. We evaluate two different versions of the summaries, adapted for their respective target audience and report the results of this evaluation.

Chapter 9: Tool-Supported Legal Assessment of Standard Form Contracts
In Chapter 9, we present the research prototype of a tool that we build to support consumer advocates in working with standard form contracts and detecting void clauses, based on the requirements identified in Chapter 5. The tool is based on the NLP methods described in the previous chapters. This chapter focuses on the overarching architecture, the underlying data models, the connecting APIs, and the user interface of the application.

Chapter 10: Evaluation
We evaluated the tool described in the previous chapter based on a questionnaire and a task-based evaluation, in which we provide experts with real-life scenarios and observe whether the automation and support features we developed help them to complete these tasks.

Chapter 11: Conclusion
The final chapter of this thesis summarizes the results, critically reflects on them and revisits the research questions. Limitations of the presented results are analyzed and open questions, as well as future research directions, are outlined.

Automated Semantic Analysis, Legal Assessment, and Summarization of Standard Form Contracts			
Chapters	**Contributions**	**RQs**	**Publications**
Introduction			
Foundations	NLU Evaluation NLU Corpus NLU Scripts		Braun et al. (2017a) Braun et al. (2018a)
Related Work	Literature Review	RQ1	Klymenko et al. (2020)
Empirical Relevance of Standard Form Contracts	Relevant Clauses	RQ2	
Requirements Identification	Process Model	RQ3, RQ4	
Semantic Analysis of Standard Form Contracts	Technology Comparison NLP Pipeline T&C Corpus	RQ5	Braun and Matthes (2020) Braun et al. (2019c) Braun et al. (2019d)
Legal Assessment of Standard Form Contracts	Knowledge-base	RQ6	Braun et al. (2018c)
Summarization of Standard Form Contracts	Lexicon Surface Realizer	RQ7	Braun et al. (2019b) Braun et al. (2017b) Klimt et al. (2020)
Tool-Supported Legal Assessment of Standard Form Contracts	Web-based Prototype	RQ4	
Evaluation	Evaluation Technique Tool Evaluation	RQ8	
Conclusion			Braun et al. (2020)

Table 1.1.: Organisation of the thesis, including contributions (Knowledge, Linguistic Resources, Software Artifacts), research questions, and main publications

CHAPTER 2

Foundations

In this chapter, we introduce technical and legal foundations on which the remainder of this thesis is based. We start in Section 2.1 with the statistical classification techniques which will be used in this work. In Section 2.2, relevant NLP foundations follow and Section 2.3 focuses on the legal foundations of the drafting of standard form contracts and the special position of consumer protection organizations within the legal system.

2.1. Statistical Classification

Statistical classification, from a computer science perspective, is an instance of supervised ML, where the goal is to assign each input with a label from a set of classes. In the most simple case, this can be an exclusive binary classification, like the classification of whether a given HTML document is a T&C page or not we present in Section 6.1. If the set of classes contains more than two labels, it is considered to be a multi-class classification task. If more than one label can be applied per input, the task is considered to be a multi-label classification task. The topic classification task we present in Section 6.3, for example, is a multi-label multi-class classification task.

In this section, we will shortly introduce the different statistical classifiers we are going to use in this thesis. This is not meant to be a comprehensive introduction into ML; instead, we will focus on the differences between the introduced classifiers, which will be relevant for the analysis and interpretation of our results. We will also introduce the hyper-parameters for the different classifiers which can be optimized to improve the results that can be achieved.

2.1.1. Logistic Regression

Logistic regression is one of the most basic approaches to statistical classification. It works much like linear regression, but instead of fitting a linear function to the data, it fits a sigmoid function. It also does not assume normal distribution of the dependent variable and can therefore be used on categorical dependent variables. One of the advantages of logistic regression is that it provides a probability for each classification. While it only supports binary classification natively, by combining multiple logistic regression classifiers, we can perform multi-label and multi-class classification. In order to avoid overfitting, especially in cases with little but high-dimensional data, logistic regression, like support vector machines and other classifiers, use a regularization-parameter that penalizes large values of parameters. The inverse of the regularization strength ($C = 1/\lambda$) is a hyper-parameter that can be optimized in order to achieve optimal results.

2.1.2. Decision Trees

Decision trees are tree structures in which each inner node represents a binary decision rule. Based on the input, we either follow the left or right branch. Each leaf represents a classification. By following the tree from the root to a leaf, each input is assigned a class. One advantage of decision trees is their interpretability. Thanks to the explicit individual decisions, the elements that contribute to a decision can be easily identified and analyzed. Unlike logistic regression, decision trees also inherently support multi-label classification because a leave can represent more than one class. However, decision trees are known to be more prone to overfitting and also less efficient.

In order to mitigate the tendency for overfitting and improve the classification quality, instead of using just one decision tree, multiple decision trees can be generated with the help of a randomization element. Each individual decision tree then performs a classification, and in the end, a majority vote decides on the final classification. The approach is called random forest. The number of trees, the maximum depth of each individual tree, the minimum number of samples per internal node that is needed to perform a decision, and the minimum number of samples per leaf are hyper-parameters that can be optimized in order to improve classification performance.

2.1.3. Multilayer Perceptrons

A Multilayer Perceptron (MLP) is the most basic manifestation of a feed-forward artificial neural network, which consists of an input layer, an output layer, and one or multiple hidden layers, each of which can contain one or multiple neurons with non-linear activation functions. Thanks to these non-linear activation functions, MLPs are universal function approximators (Cybenko, 1989).

Many hyper-parameters are involved in training an MLP, which makes their optimization resource-intensive. In addition to the number of hidden layers and the number of neurons in each of these layers, a dropout coefficient can be set, which helps to prevent overfitting. The number of training epochs and the batch size are hyper-parameters that can be optimized, as well as the optimizer function, the learning rate, and the activation functions that are used.

2.1.4. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a different kind of artificial neural network. Their main area of application in computer vision tasks (i.e., image and video processing). In an MLP, layers are fully connected, i.e., each neuron of the previous layer is connected to each neuron of the next layer. All these connections increase the complexity of the network and the number of parameters. Layers in CNNs, on the other hand, are not fully connected. They operate under the assumption that spatially close inputs are correlated and can, therefore, share their weights. This assumption is true for images, where each pixel is correlated with its neighbors, but also for text, where each word is correlated with its neighbors. Additionally, CNNs also use so-called pooling-layers, which can be used to further reduce the complexity of the input by *pooling* spatially close inputs. Thanks to these two concepts, sharing of weights and pooling, CNNs are computationally much more efficient than MLPs. CNNs can also be used in connection with fully-connected layers of neurons, which are usually placed at the end of the network after the complexity of the input has already been reduced by the convolutional and pooling layers.

The hyper-parameters for CNNs that can be optimized correspond with the hyper-parameters of MLPs, however, instead of hidden layers, we have convolution layers, and instead of neurons, we have filters. Since filters, unlike neurons, can take multiple inputs (e.g., multiple pixels or multiple words), the size of the input is also a fixed parameter, the so-called kernel size.

2.1.5. Recurrent Neural Networks

CNNs are particularly good at working with two-dimensional input data. In NLP, however, we usually deal with sequential data, which often has a less strong correlation between neighboring inputs: while removing every second pixel in a high-resolution image will most likely still allow us to understand the content of the picture, removing every second word of a text, will most likely make that impossible.

Recurrent Neural Networks (RNNs) are artificial neural networks that are optimized for dealing with sequential data. Like MLPs, their layers are fully connected. While MLPs treat each input isolated, RNNs have a “memory” of previous input. The activation function of each neuron is updated with every new input, based on the previous input. In this way, in the case of a word sequence, the previous words of a sequence influence the analysis of the current word, while an MLP would treat each word isolated. In practice, RNNs deal with the problem of “vanishing gradients” during training, i.e., the gradient that is used becomes so small that weights are not updated and the network is “stuck”. To mitigate this problem, Hochreiter and Schmidhuber (1997) suggest an RNN architecture they called Long short-term memory (LSTM), which increases the memory of the network and helps to keep the gradient sufficiently large. RNNs use the same hyper-parameters as MLPs, but in addition, we have to set the length of the sequence as a fixed parameter.

2.1.6. Transformers

So-called transformer models (or just transformers) are the latest development in NLP. They were introduced by Vaswani et al. (2017) and have since shown to outperform previous approaches in many NLP tasks. In their core, transformer models rely on a stack of feed-forward neural networks, however, they add so-called attention layers. Instead of simply feeding individual words into the network, the attention layer calculates an inter-dependency between the input word and the rest of the sentence, based on their co-occurrence. In order to be able to calculate these inter-dependencies, the models have to be trained on huge data sets. Initially, transformer models were designed for Machine Translation. The following example might help to understand why the idea of attention can help to improve performance on this task. If we look at the two English sentences "The apple is red." and "The car is red." and their German translations, "Der Apfel ist rot." and "Das Auto ist rot.", we see that the same word "the" is once translated as "der" and once translated as "das". Since the rest of the sentence is translated the same, whether "the" is translated as "der" or "das" seems to be terminated by the following word. Expressed in an attention model, we would say that there is high attention between "the" and "apple" and "the" and "car", but low attention between, e.g., "the" and "red". By giving the feed-forward neural network not just a single word but also the weights from the attention layer, we do not look at words isolated anymore (like in an MLP) or based on the previous words in the sequence (like in an RNN), but look at each word in relation to the whole sequence.

Training a transformer model is a massive effort with regard to the necessary data and computational power. Therefore, the standard approach is to use a pre-trained transformer model and fine-tune it on a specific task. One of the largest and best-known transformer models is the Bidirectional Encoder Representations from Transformers (BERT) from Devlin et al. (2019). In order to fine-tune the model, an additional hidden layer is added, which is trained on the specific task (in our case, for example, classification), based on the output of the transformer model. For this fine-tuning, the batch size, number of epochs, and the learning rate are hyper-parameters that can be optimized; the other parameters are usually defined by the pre-trained model.

2.2. Natural Language Processing Foundations

In the following sections, we will introduce some foundations from the domain of NLP on which this thesis is based. We will start by introducing so-called NLU services, based on a comparison study we conducted and published (Braun et al., 2017a). Afterwards, we will introduce two basic methods to parse the structure of sentences.

2.2.1. Natural Language Understanding Services

NLU is the subarea of NLP, which is concerned with understanding natural language, i.e., processing its meaning. That includes tasks like relation extraction, sentiment analysis, and question answering, while tasks like part of speech tagging or named entity recognition are generally seen as NLP but not NLU tasks.

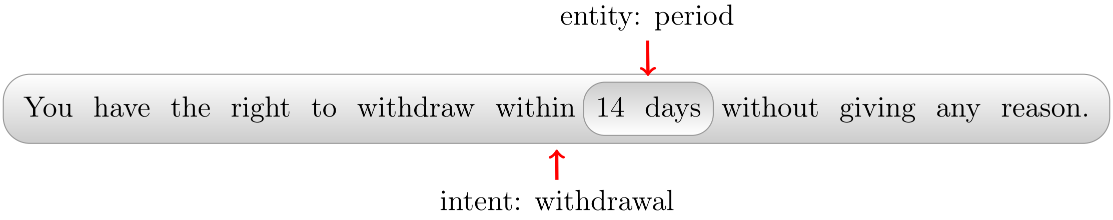

entity: period
You have the right to withdraw within 14 days without giving any reason.
intent: withdrawal

Figure 2.1.: Withdrawal clause with intent and entity annotations

In recent years, in alignment with the general trend towards Software as a Service (SaaS), big tech companies like IBM, Amazon, Google, and Microsoft started to offer AI and ML cloud services. By providing ML through APIs over the internet, such services have become very popular with scientists and practitioners. They enable their users to rapidly build prototypes or products without having to set-up (and sometimes even without having to train) ML environments. These services achieve this by moving algorithms, their parameters and models, and their execution from a local machine to the cloud.

One class of such services that has emerged are NLU cloud services, like Microsoft LUIS[1], IBM Watson Conversation[2] and Dialog Flow[3] (formerly known as API.ai). Despite their very general categorization as NLU services, all of these services offer a very specific set of features. They can be used for the extraction of structured, semantic information from unstructured natural language input, like chat messages, by attaching user-defined labels to whole sentences or parts of them. We will use this approach as one mean to semantically analyze clauses from standard form contracts (see Chapter 6). The classification of one or multiple complete sentences is called *intent extraction* by these services because the goal is to extract the objective or intention of a given text segment. The more fine-grained labeling of one or multiple words is called entity extraction because the goal is to label key elements of a text. In both cases, the labels are pre-defined by the user.

In the context of the semantic analysis of standard form contracts, intent types could, for example, be the different clause types, like a withdrawal clause or a warranty clause. Entity types could be the withdrawal period or the warranty period, respectively. Figure 2.1 shows how a withdrawal clause could look like after it has been annotated by an NLU service.

One of the limitations these services have is that they do not support multi-label classification for intents; each text unit can only belong to one intent. In standard form contracts, however, a clause can regulate multiple aspects and hence belong to multiple types of intents. We will therefore only use the entity extraction of NLU services in our work and use different approaches to the classification of intents.

[1] https://www.luis.ai

[2] https://www.ibm.com/watson/services/conversation/

[3] https://www.dialogflow.com

2.2.1.1. Corpus

Because the functionality all of these NLU services offer are so similar, we conducted a structured comparison of them in order to find out which of them shows the best classification performance (Braun et al., 2017a). Since, at the time, there was no corpus available that was sufficiently annotated to conduct such a comparison, we build a corpus for the task. Since then, the systematic evaluation of NLU services has gained more attention, and a number of researchers have used our corpora to evaluate new NLU services and libraries (Coucke et al., 2018; Chandna and Iyer, 2018; Huijzer, 2019; Schuurmans et al., 2020; Shridhar et al., 2019; Sergio and Lee, 2020) or introduced new corpora, wit reference to our work, covering other languages and domains (Bellomaria et al., 2019; Larson et al., 2019; Liu et al., 2019).

The corpus we built consists of three sub-corpora which cover different domains: The “Chatbot Corpus” is based on questions gathered by a Telegram chatbot that answers questions about public transport in Munich. “ask ubuntu”[4] corpus is based on data from the StackExchange platform with the same name and so is the “Web Applications”[5] corpus. All three corpora are available under the Creative Commons CC BY-SA 3.0 license.[6]

The Chatbot Corpus consists of 206 questions, which were manually labeled with two different intents (Departure Time, Find Connection; see Table 2.1) and five different entity types (StationStart, StationDest, Criterion, Vehicle, Line; see Table 2.2). The general language of the questions was English, however, mixed with German street and station names. For the evaluation, the corpus was split into a training data set with 100 entries and a test data set with 106 entries.

For the collection of the StackExchange corpus, we used the StackExchange Data Explorer[7]. We choose the most popular questions (i.e., questions with the highest scores and most views) from the two StackExchange platforms *ask ubuntu* and *Web Applications*, because they are likely to have better quality and a higher relevance, compared to lower ranked questions. We also only used questions with an accepted answer. Although we did not need the answers for the evaluation, we included them in the corpus to create a corpus that is also applicable to question answering tasks. In total, we gathered 290 questions and answers. 100 from *Web Applications* and 190 from *ask ubuntu*. The corpus was labeled with intents and entities using Amazon Mechanical Turk (AMT). Each question was labeled by five different workers, summing up to nearly 1,500 data points. For each source, we created a list of candidates for intents, which were extracted from tags assigned to the questions by StackExchange users. For each question, the AMT workers were asked to chose one of these intent candidates or “None”, if they think no candidate is fitting. For *ask ubuntu*, the possible intents were: “Make Update”, “Setup Printer”, “Shutdown Computer”, and “Software Recommendation”. For *Web Applications*, the candidates were: “Change Password”, “Delete Account”, “Download Video”, “Export Data”, “Filter Spam”, “Find Alternative”, and “Sync Accounts”. Similarly, a set of entity types were given. By highlighting parts of the questions, workers could assign these entity types to words (or characters) within the questions. For *Web Applications* the possible entity types were:

[4] `https://www.askubuntu.com`

[5] `https://webapps.stackexchange.com`

[6] `https://github.com/sebischair/NLU-Evaluation-Corpora`

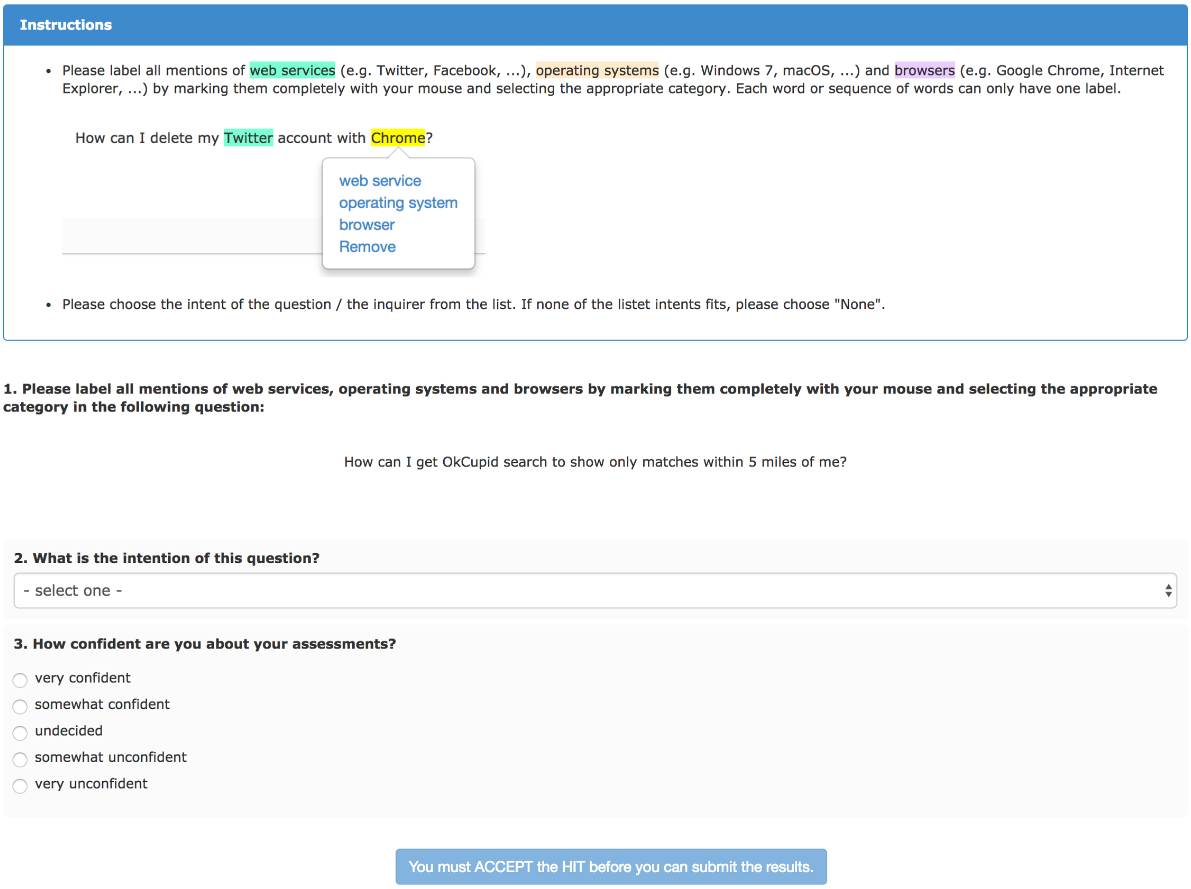
[7] `https://data.stackexchange.com`

Corpus	Intent	Training	Test	Σ
web apps	ChangePassword	2	6	8
	DeleteAccount	7	10	17
	DownloadVideo	1	0	1
	ExportData	2	3	5
	FilterSpam	6	14	20
	FindAlternative	7	16	23
	SyncAccounts	3	6	9
	None	2	4	6
	Σ	30	59	89
ask ubuntu	MakeUpdate	10	37	47
	SetupPrinter	10	13	23
	ShutdownComputer	13	14	27
	S.Recommendation	17	40	57
	None	3	5	8
	Σ	53	109	162
chatbot	DepartureTime	43	35	78
	FindConnection	57	71	128
	Σ	100	106	206

Table 2.1.: Intents within the three corpora

Corpus	Entity Type	Training	Test	Σ
web apps	WebService	33	64	97
	OS	1	0	1
	Browser	1	0	1
	Σ	35	64	99
ubuntu	Printer	8	12	20
	Software	3	4	7
	Version	24	78	102
	Σ	35	94	129
chatbot	StationStart	91	102	193
	StationDest	57	71	128
	Criterion	48	34	82
	Vehicle	50	35	85
	Line	4	2	6
	Σ	250	244	494

Table 2.2.: Entity types within the three corpora



Instructions

- Please label all mentions of web services (e.g. Twitter, Facebook, ...), operating systems (e.g. Windows 7, macOS, ...) and browsers (e.g. Google Chrome, Internet Explorer, ...) by marking them completely with your mouse and selecting the appropriate category. Each word or sequence of words can only have one label.

  How can I delete my Twitter account with Chrome?

- Please choose the intent of the question / the inquirer from the list. If none of the listet intents fits, please choose "None".

1. Please label all mentions of web services, operating systems and browsers by marking them completely with your mouse and selecting the appropriate category in the following question:

How can I get OkCupid search to show only matches within 5 miles of me?

2. What is the intention of this question?

- select one -

3. How confident are you about your assessments?

○ very confident
○ somewhat confident
○ undecided
○ somewhat unconfident
○ very unconfident

Figure 2.2.: Mechanical Turk interface for workers

"WebService", "OperatingSystem" and "Browser". For *ask ubuntu*, "SoftwareName", "Printer", and "UbuntuVersion" were given. Moreover, workers were asked to state how confident they are in their assessment on a five-time Likert-scale: very confident, somewhat confident, undecided, somewhat unconfident, or very unconfident. Figure 2.2 shows the interface which was shown to AMT workers.

For the final corpus, only submissions with a confidence level of "undecided" or higher were taken into account. A label, no matter if intent or entity, was only added to the corpus if the inter-annotator agreement between confident annotators was 60% or higher. If no intent was satisfying these criteria for a question, it was not added to the corpus. The final corpus consists of 251 question ans answer pairs, 162 from *ask ubuntu* and 89 from *Web Applications* (see Table 2.1 and Table 2.2 for the distribution of intents and entity types). Example entries from all three corpora are shown in Appendix B.

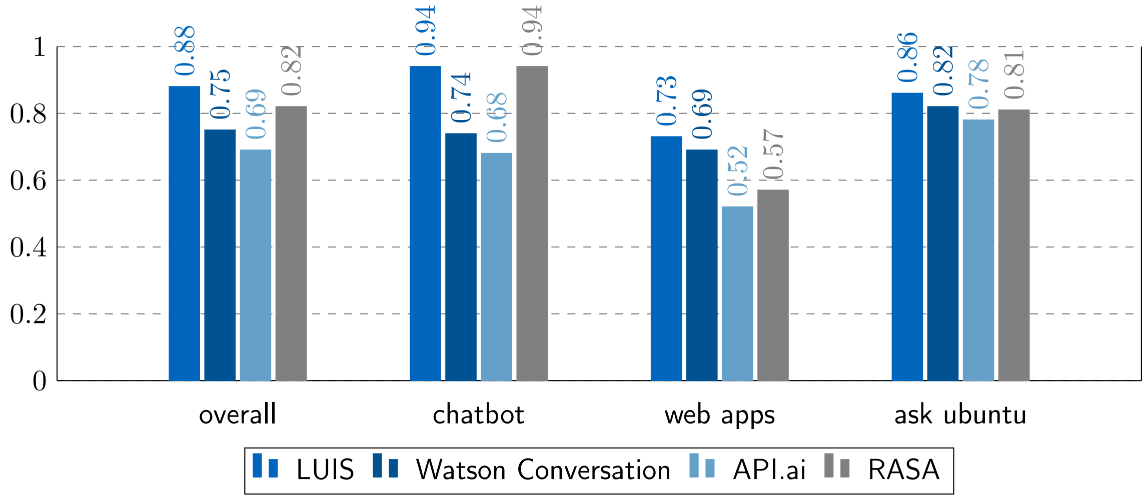


Figure 2.3.: F1-scores for the different NLU services, grouped by corpus

2.2.1.2. Evaluation

We used these three corpora to compare the performance of Microsoft LUIS, IBM Watson Conversation, Dialog Flow, and Rasa Core[8] (Bocklisch et al., 2017). Rasa is an open source library that can be used as a drop-in replacement for all three services by mimicking their API. It was first published in 2016 and has since become one of the most popular libraries in the field. Rasa integrates existing ML libraries and offers multiple of them from which the user can choose. At the time of our evaluation, the standard pipeline used the MIT Information Extraction (MITIE) library, which, based on the results of this evaluation, will also use for the information extraction approach we describe in Section 6.4.

We trained and tested all services with a set of automated Python scripts that we published under the open MIT license.[9] The scripts calculate true positives, false positives, and false negatives, based on exact matches. Based on this data, they compute precision and recall as well as the F1-score for intents, entity types, and corpora, as well as overall results. The results presented here were previously presented by Braun et al. (2017a) and obtained in January 2017. From a high-level perspective, in this evaluation LUIS performed best with an overall F1-score of 0.88, followed by Rasa with 0.82, Watson Conversation with 0.75 and Dialog Flow with 0.69 (see Figure 2.3).[10]

A very similar evaluation was carried out two years later by Liu et al. (2019), using a new, bigger corpus. The results from Liu et al. (2019) are shown in Table 2.3 and confirm that LUIS and Rasa are the best-performing services.

Because of the good results that Rasa showed in both evaluations, and also because it is open source, we decided to use the technology that is applied by Rasa for the semantic analysis of

[8] `https://rasa.com/docs/rasa/core/about/`

[9] `https://github.com/sebischair/NLU-Evaluation-Scripts`

[10] For more detailed results see Braun et al. (2017a).

	Intent			Entity			Overall		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Rasa	0.863	0.863	0.863	0.859	0.694	0.768	0.862	0.787	0.822
Dialogflow	0.870	0.859	0.864	0.782	0.709	0.743	0.832	0.791	0.811
LUIS	0.855	0.855	0.855	0.837	0.725	0.777	0.848	0.796	0.821
Watson	0.884	0.881	0.882	0.354	0.787	0.488	0.540	0.838	0.657

Table 2.3.: NLU evaluation results from Liu et al. (2019)

standard form contracts. Since the intent classification has the aforementioned limitation that it does not support multi-label classification, we only used the entity extraction technology, which, in comparison to the other services, performs even better than the intent classification (see Table 2.3).

2.2.2. Constituency and Dependency Trees

A common way to parse and represent the structure of sentences are so-called phrase structure grammars (Chomsky, 1985, p. 26 ff.), also known as constituency grammars. At their core, they are based on constituency relations like: a noun phrase (NP) can be constituted of a determiner (DT) and a noun (N): $NP \rightarrow DT{+}N$ (see Appendix A.1 for an overview of the PoS and sentence constituent tags). The common way to represent the result of parsing a sentence with such a grammar is a constituency tree. Figure 2.4 shows the constituency tree for the sentence “This does not affect your statutory rights.”, as produced by the Stanford CoreNLP library (Manning et al., 2014).

In this thesis, we will use a different, more sophisticated approach to analyze the structure of sentences, so-called dependency grammars. Dependency grammars take the (finite) verb of a sentence as its root and represent the structure by one-to-one relationships between the words. In the sentence “This does not affect your statutory rights.”, one of these relationships would, e.g., be a negation modifier (neg) between “affect” and “not” (`neg(affect, not)`) and an adjectival modifier (amod) relation between “rights” and “statutory” (`amod(rights, statutory)`). We show in Chapter 6 that this representation of the sentence structure is better suited for the extraction of information and subsequent legal assessment of clauses in standard form contracts than constituency trees. A list of the dependency types used in this thesis can be found in Appendix A.2. The common way to represent the results of dependency parsing are so-called dependency trees. Figure 2.5 shows the dependency graph for the sentence “This does not affect your statutory rights.”, as produced by the Stanford CoreNLP library.

2.3. Legal Foundations

This work is by no means a legal treatise and does not aspire to be. The research it reports was conducted in an interdisciplinary team, involving legal scholars and practitioners who work on consumer protection matters on a daily basis.

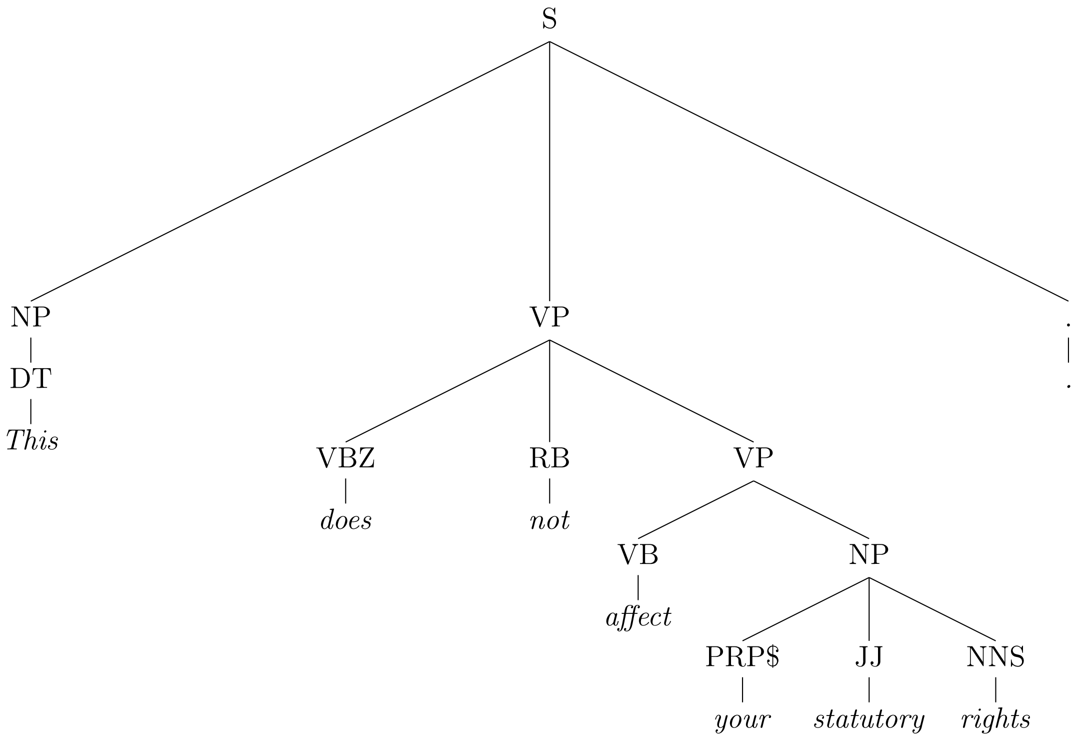


Figure 2.4.: Constituency tree with PoS tags for the sentence "This does not affect your statutory rights."

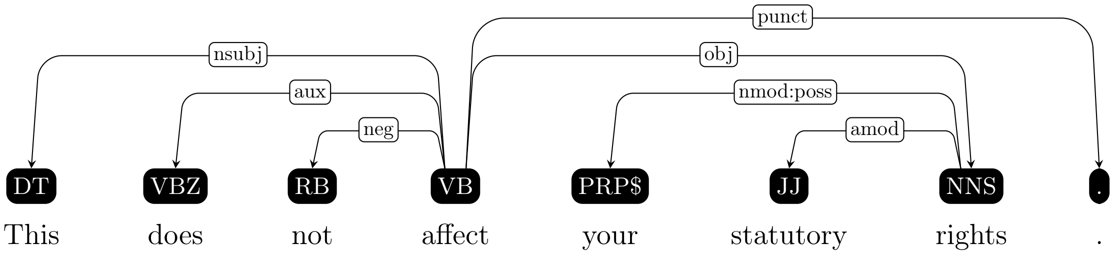


Figure 2.5.: Dependency tree with PoS tags for the sentence "This does not affect your statutory rights."

In order to do justice to the legal aspects of this work and provide the foundations that are necessary to understand the issue we are working on, we are going to outline the legal foundations of consumer protection with a specific emphasis on Germany (see Section 2.3.1) and more broadly within the single market of the EU (see Section 2.3.2). The remarks in this section are not targeted at an audience of legal experts and will be a simplified representation of the legal situation.

The folk wisdom that being right does not automatically lead to getting justice is specifically true for the area of consumer protection, where there is regularly a strong imbalance of power between the involved parties, a single consumer on one side and a potentially large corporation on the other side. In acknowledgment of this fact, the legislators have given NGOs in the area of consumer protection special and extensive rights to assist and represent consumers and their interests. In addition to the laws that grant consumers their rights, we will therefore also briefly discuss the legislative basis on which consumer protection agencies operate.

2.3.1. Germany

Consumer protection has historically always played an important role in Germany. The first federal NGO for consumer protection, the "Arbeitsgemeinschaft der Verbraucherverbände e.V. (AgV)" (working group of the consumer associations) was founded in 1953. In 1961, all (at the time) eleven federal states of Germany had founded local "Verbraucherzentralen" (consumer protection agencies). After the German reunification in 1990, the five new federal states successively also founded "Verbraucherzentralen". Since 2000, the now 16 "Verbraucherzentralen" and 26 other organizations are represented by the vzbv on the federal level. (Verbraucherzentrale Bundesverband e.V., 2020a)

From 2013, the ever-increasing importance of consumer protection was also reflected politically, when the former Federal Ministry of Food, Agriculture and Forests ("Bundesministerium für Ernährung, Landwirtschaft und Forsten") was renamed to Federal Ministry of Consumer Protection, Food and Agriculture ("Bundesministerium für Verbraucherschutz, Ernährung und Landwirtschaft"). In 2013, the competency for consumer protection moved from the Federal Ministry of Food and Agriculture to the Federal Ministry of Justice, which was subsequently renamed to Federal Ministry of Justice and Consumer Protection (Bundesministerium für Justiz und Verbraucherschutz (BMJV)).

2.3.1.1. Governing Law

While many countries have dedicated consumer protection laws, e.g., the Konsumentenschutzgesetz (KSchG) in Austria (BGBl, 2018) or the "Consumer Protection Act 1987" in the United Kingdom (c. 43, 1987), in Germany, matters of consumer protection can be found in different pieces of legislation. As is the case in many areas of the law nowadays, a lot of the national legislation is based on directives and regulations from the EU (see Section 2.3.2).

The most important piece of legislation in Germany with regard to standard form contracts is the German civil code BGB. In §§ 305-310, the BGB outlines the legal framework for drafting standard form contracts in Germany, with direct reference to the Council Directive 93/13/EEC of 5 April 1993 on unfair terms in consumer contracts. Since the focus of this thesis is on T&C from online shops, another important applicable German law was the distance selling act (Fernabsatzgesetz (FernAbsG)), which in 2002 was incorporated into the BGB (§§ 312b ff).

The BGB contains both very general fundamental principles for the drafting of standard form contracts, as well as very specific bans of individual clauses. Many of the regulations contain so-called vague legal terms.

Definition 2. *A vague legal term ("unbestimmter Rechtsbegriff") is a term in a law or another legislative document that is deliberately vague or unexhaustive and therefore has to be interpreted individually on every application of the law. Examples of such terms are "appropriate" or "sufficient".*

A fundamental principle regulating the drafting of standard form contracts that uses a vague legal term can be found in §305c Abs. 1 BGB: *"Provisions in standard business terms which in the circumstances, in particular with regard to the outward appearance of the contract, are so unusual that the other party to the contract with the user need not expect to encounter them, do not form part of the contract."* [11]

Another is §307 Abs. 1 BGB: *"Provisions in standard business terms are ineffective if, contrary to the requirement of good faith, they unreasonably disadvantage the other party to the contract with the user. An unreasonable disadvantage may also arise from the provision not being clear and comprehensible."* [11]

The interpretation and application of such vague legal terms to assess whether a clause in a standard form contract is void or not is a problem that is regularly also difficult for humans to solve. The main challenge is the substantiation of the vague term. As soon as it is defined what an *"unreasonably high remuneration for enjoyment or use of a thing"*[11] is (§308 Abs. 7a BGB), it becomes much easier to assess whether a clause adheres to this principle or not. The legal assessment that we perform is not directly based on legislative texts but on a knowledge-base that is created and maintained by legal experts (see Section 7). We do not concern ourselves with the automatic extraction of regulations from legislative texts. It is therefore within the responsibility of the legal experts to specify vague legal terms, based on their expertise or existing specifications from judgments or commentaries, where it is necessary. In cases where the legal assessment is based on data-driven approaches, the specification of a vague term is implicitly given by the annotations that are made in the training examples.

The bans of specific clauses contained in the BGB are sorted by whether they contain a vague legal term (§308) or not (§309). Clauses containing one of the following elements are deemed to be void: [11,12]

- unreasonably long or insufficiently specific periods of time for the acceptance or rejection of an offer (§308 Abs. 1)

[11] All translation of the BGB from `https://www.gesetze-im-internet.de/englisch_bgb/`

[12] Please note: The purpose of this list is to give the reader an idea about what areas of standard form contracts are regulated by the law. It is not exhaustive and simplified.

- unreasonably long payment deadline (§308 Abs. 1a)
- unreasonably long period for the review and acceptance of compensatory measures (§308 Abs. 1b)
- unreasonably long or insufficiently specific grace periods (§308 Abs. 2)
- right to free oneself from obligations without any objectively justified reason indicated in the contract (§308 Abs. 3)
- right to modify the performance promised or deviate from it in an unexpected way (§308 Abs. 4)
- provisions by which omitting a specific act is treated as if a declaration was made (§308 Abs. 5)
- provisions by which a declaration of special importance is deemed to have been received, even if not proven (§308 Abs. 6)
- unreasonably high remuneration for the use of a product or unreasonably high reimbursement of expenses in case the contract is revoked (§308 Abs. 7)
- possibility to revoke a contract without the obligation to inform or reimburse the other party without undue delay (§308 Abs. 8)
- price increase for goods or services that are to be delivered within four months (§309 Abs. 1)
- exclusion or restriction of retention rights (§309 Abs. 2)
- restrict of the right to charge up against claims that are uncontested or have been finally and non-appealably established (§309 Abs. 3)
- exemption from the requirement of giving warning notices for unfulfilled obligations (§309 Abs. 4)
- agreement of a flat-fees for compensations, if the sum exceeds the damage expected under normal circumstances or the other party is not expressly permitted to prove that the damage has either not occurred or is substantially less than the flat-fee (§309 Abs. 5)
- agreement of a contractual penalty in the event of non-acceptance, payment default or withdrawal from the contract (§309 Abs. 6)
- exclusion or limitation of liability for damage from injury to life, body or health (§309 Abs. 7a)
- exclusion or limitation of liability for damage arising from a grossly negligent breach of duty (§309 Abs. 7b)
- exclude or restrict the right to withdraw from the contract in case of a breach of duty for which the drafting party is responsible and which does not consist in a defect of the sold good or work (§309 Abs. 8a)

- exclude or limit claims against defects for entire goods or individual parts or setting a cut-off period for giving notices about of non-obvious defects (§309 Abs. 8b)
- a contract term of more than two years, an automatic extension of the contractual relationship by more than one year or a notice period longer than three months prior to the expiry of the duration of the contract (§309 Abs. 9)
- allowing a third party to take the rights and duties under the contract (§309 Abs. 10)
- modifications of the burden of proof (§309 Abs. 12)
- notices or declarations tied to a more stringent form than the written form or tied to special receipt requirements (§309 Abs. 13)
- mandatory extrajudicial dispute resolution (§309 Abs. 14)
- unusually high progress payments or unusually low provision of security in contracts for work and labor (§309 Abs. 15)

All of these regulations are designed to protect the party that did not draft the standard form contract, which, in our context, is the consumer. A company can *not* grant itself an "unreasonably long payment deadline", but it can grant one to the consumer.

The paragraphs §§ 312b ff. BGB (former FernAbsG) contain among other things the regulations regarding the withdrawal right (§ 312 d BGB) and the for the legal assessment of T&C important addition that no agreements deviating from the provisions of these paragraphs may be made, that are to the disadvantage of consumers (§ 312f BGB).

2.3.1.2. Consumer Advocates

In Germany, the Legal Services Act (Rechtsdienstleistungsgesetz (RDG)) regulates which entities are allowed to offer individual out-of-court legal services. This law grants consumer advocates a special position in the German legal system since it allows "Verbraucherzentralen" (consumer protection agencies) and "other publicly funded consumer protection organizations" to offer out-of-court legal services in consumer protection matters (§8 Abs. 1 Nr. 4 RDG).

The German Injunctions in the Case of Breaches of Consumer Protection and Other Laws act (Injunctions Act, Unterlassungsklagengesetz (UKlaG)) gives consumer protection organizations that fulfill certain requirements (§4 UKlaG) even more power by giving them the right to file injunctions if companies break consumer protection or other laws. As of February 2020, the Federal Office of Justice (Bundesamt für Justiz) lists 78 consumer protection organizations with this right, including the 16 "Verbraucherzentralen" and their federal organization.

The "Verbraucherzentralen" take, in many aspects, a special position among consumer protection organizations in Germany. As only organizations, they are explicitly named in §4 UKlaG and §8 RDG and they are the only organizations that receive permanent governmental funding in all 16 states and on the federal level and are endorsed by the federal ministry (Presse- und Informationsamt der Bundesregierung, 2019). As such, they have a certain right of representation for consumers and their interests, and we are very happy that, for this research, we were able to cooperate with two of the “Verbraucherzentralen”.

2.3.2. European Union

In its effort to harmonize the European single market, the EU also pursues high standards for consumer protection within the single market (Article 114 Treaty on the Functioning of the European Union (TFEU)). Most of the aforementioned German laws and regulations are based on EU directives and regulations and in similar form also applicable in other countries that participate in the single market (including non-EU members).

2.3.2.1. Governing Law

The Council Directive 93/13/EEC of 5 April 1993 on unfair terms in consumer contracts laid the foundations for many of the regulations for standard form contracts that are still in place today and mainly outlines the same regulations discussed in Section 2.3.1.1. One of its last big updates, the Directive 2011/83/EU of the European Parliament and of the Council of 25 October 2011 on consumer rights, which introduced especially now regulations regarding how to inform consumers about their withdrawal rights.

2.3.2.2. Consumer Advocates

The European Commission lists four federations of national consumer protection bodies from across the EU with which it works together[13]:

- European Association for Coordinating Consumer Representation in Standardisation (ANEC),
- European Consumer’s Organisation (BEUC),
- Confederation of Family Organisations in the EU (COFACE), and
- European Community of Consumer Cooperatives (Euro Coop).

Two of which, ANEC and BEUC, are also subsidized by the commission. The German vzbv is a member of BEUC.

In 2020, the EU and its member states agreed on the Directive of the European Parliament and of the Council on Representative Actions for the Protection of the Collective Interests of Consumers, which, in the future, will give consumer protection organizations from member states the right to file legal proceedings on behalf of consumers in all member states and on EU level. However, before the regulation comes into effect, it has to be implemented in national law by all member states within two years.

[13] `https://ec.europa.eu/info/policies/consumers/consumer-protection/our-partners-consumer-issues/european-and-international-consumer-organisations_en`, last accessed 2020-10-14

CHAPTER 3

Related Work

In this chapter, we will give an overview of related research that was conducted in the area of semantic analysis of legal texts (Section 3.1), knowledge representation in the legal domain (Section 3.2), and text summarization (Section 3.3). We will show how the existing work differs from the research described in this thesis for each of the areas. Towards the end of the chapter, we will also briefly discuss the commercial market in the area by looking at commercial tools for contract analysis (Section 3.4) and (semi-)automated T&C drafting services (Section 3.5).

3.1. Semantic Analysis of Legal Texts

The semantic analysis of legal texts is a relatively broad field, based on the number of tasks that exist, but also based on the variety of available document types. The three, arguably, most important document types are laws, contracts, and court decisions. Since the introduction of the General Data Protection Regulation (GDPR), privacy policies are an ever more important document type, not just with regard to automated processing. Other documents, which generally receive less attention, include legal commentaries and scholarly legal texts. This section starts with a detailed description of two closely related larger research projects, CLAUDETTE and LEXIA. Subsequently, we will describe smaller projects, based on the type of document they are working with (privacy policies, contracts, laws, and court decisions).

3.1.1. CLAUDETTE

The most closely related work was conducted by Lippi et al. as part of the CLAUDETTE project at the European University Institute (Micklitz et al., 2017; Lippi et al., 2017; Contissa et al., 2018b,a; Lippi et al., 2019b,a,c; Liepina et al., 2019). CLAUDETTE focuses on the detection of unfair clauses in terms of the legislation of the EU. The project started in 2017 with uTerms. uTerms identifies five classes of unfair clauses in English (unilateral change, unilateral termination, liability, choice of law and jurisdiction), based on regular expressions. One expression used to identify a potentially unfair clause on the unilateral termination looks like this: `([(may|can)] [stop] [providing] []* [you])`. uTerms can take the URL of a T&C page as input; however, no content extraction (see Section 6.2.1) is performed, i.e., all website elements, like menus, headers, and footers, are part of the result. On a not further specified set of 20 contracts consisting of 109.000 words, the authors report that their tool has a recall of 1.0 and a precision of 0.82. (Micklitz et al., 2017)

Afterwards, Lippi et al. (2017) started to use ML instead of regular expressions to detect unfair clauses. They first annotated Terms of Services from 20 English online platforms (from 9gag.com, Academia.edu, Amazon, eBay, Dropbox, Facebook, Google, Linden Lab, Microsoft, Netflix, Rovio, Snapchat, Spotify, Supercell, Twitter, Vimeo, World of Warcraft, Yahoo, YouTube, and Zynga) with eight classes of potentially unfair clauses: jurisdiction, choice of law, limitation of liability, unilateral change, unilateral termination, arbitration, contract by using, and content removal. The corpus was generated manually, i.e., the content was extracted manually from the websites, and sentence boundaries were also manually annotated. For each instance of these clauses the authors found, they assigned whether it was "clearly fair", "potentially unfair" or "clearly unfair". The authors used this corpus to train Support Vector Machines (SVMs) and tree kernels and evaluated the classification performance in a leave-one-document-out evaluation. The best performance was achieved with an SVM and an F1-score of 0.648. The tree kernel method achieved an F1-score of 0.603. (Lippi et al., 2017)

Building upon this work, Lippi et al. (2019b) increased the size of their corpus from 20 to 50 documents, in the same process as described above. Additionally, this time, they compared eight different classifiers, including convolutional neural networks, long short-term memory networks, and an ensemble method, combining the outputs of multiple classifiers. Table 3.1 shows the different classifiers that were evaluated and how they performed in a leave-one-document-out evaluation. Given the larger corpus, all previously tested classifiers improved their classification performance. Overall, the ensemble method (C8) performed best, with an F1-score of 0.806. (Lippi et al., 2019b)

Recently, in light of the GDPR, the focus of the CLAUDETTE project has shifted from ToS to privacy policies (Contissa et al., 2018a; Liepina et al., 2019). First, Contissa et al. (2018a) built a corpus consisting of 14 privacy policies, including those from Apple, Facebook, Google, and Microsoft. The corpus consists of 80,398 words or 3,658 sentences. The corpus was annotated on a sentence level with three different tags: "unclear language", "problematic processing", and "insufficient information". Overall, 1,641 sentences (44.9%) were labeled with one of the tags. Based on this corpus, multiple classifiers were compared. For the detection of "unclear language", a simple keyword search reached a recall of 0.89 and a precision of 0.25. An SVM achieved a recall of 0.72 with a precision of 0.3, a combination of both approaches 0.81 recall and 0.32 precision. No separate evaluation was performed for clauses tagged with "problematic processing". For clauses tagged with "unclear language" and "problematic processing", the rule-based approach achieved a recall of 0.92 with a precision of 0.31. The SVM approach achieved 0.7 and 0.5

Classifier	Precision	Recall	F1
C1 SVM – Single Model	0.729	0.830	0.769
C2 SVM – Combined Model	0.806	0.779	0.784
C3 Tree Kernels	0.777	0.718	0.739
C4 Convolutional Neural Networks	0.729	0.739	0.722
C5 Long Short-Term Memory Networks	0.696	0.723	0.698
C6 SVM-HMM – Single Model	0.759	0.778	0.758
C7 SVM-HMM – Combined Model	0.848	0.720	0.772
C8 Ensemble (C1+C2+C3+C6+C7)	0.828	0.798	0.806

Table 3.1.: Evaluation results for the classification of unfair clauses in ToS from Lippi et al. (2019b)

(precision / recall) and the combination 0.75 and 0.47. Finally, the classification of "some required information clauses" was evaluated with rule-based approaches. The reported precision values were between 0.78 and 0.91, and the recall values between 0.83 and 0.91. However, given that only favorable results were reported and "other tags are much more heterogeneous and thus difficult to detect with handcrafted rules", these values are not very meaningful. (Contissa et al., 2018a)

In their most recent work, Liepina et al. (2019) built upon this work by extending their corpus to 32 privacy policies, annotated with three slightly different labels: vague language, unfair clauses with respect to advertising, and unfair clauses with regard to the purpose of data processing. In a leave-one-document-out evaluation, the authors trained a classifier that combines SVMs and Hidden Markov Models (HMMs). For the three different labels, the classifier achieved a precision between 0.408 and 0.602 and a recall between 0.565 and 0.612. (Liepina et al., 2019)

While this thesis describes a complete NLP-pipeline, the work of Lippi et al. focuses on the detection of unfair clauses (i.e., legal assessment of clauses based on EU regulations). They do not consider pre-processing and only use manually extracted and annotated corpora. Their results are limited by the fact that they only perform leave-one-document-out evaluations on relatively small data sets (50 ToS documents, 32 privacy policies), while we work with different evaluation methodologies and larger corpora. Additionally, they only look into a very limited number of clause types (eight at most), while we investigate a broader range of more than 20 clause types.

Lippi et al. work with ToS of online platforms. Naturally, such platforms do not have any clauses regarding the shipping of products, which are important to the area we focus on, T&C of online shops. Most of the applied approaches are black-box approaches that do not use an explicit description of the underlying law and provide no explanations or summarizations. Although multilingualism is shortly mentioned in one of the works, which suggests using machine translation (Liepina et al., 2019), the approaches are never applied or evaluated on other languages.

3.1.2. LEXIA

The LEXIA project from Waltl et al. (Waltl et al., 2016b, 2017b; Waltl, 2018) investigated a broad range of topics regarding the semantic analysis of German legal texts. Unlike CLAUDETTE, the project was not focused on one document type but generally investigated laws and judgments (Waltl, 2018, p. 44). Therefore, the analyses performed by Waltl et al. are on a much higher level and more abstract than what was investigated by CLAUDETTE and what we will present in this thesis. Waltl et al. (2016a), for example, tried to automatically detect references within legal documents or classify statutory legal norms with their functional type, like rights, obligations, and definitions. On a corpus of the German tenancy law, they were able to achieve an F1-score of 0.78. (Waltl, 2018, p. 167) Such high-level classifications themselves are not sufficient for further analysis regarding lawfulness and, arguably, of limited value outside of scholarly analysis. Generally speaking, by focusing on standard form contracts, we can perform more in-depth semantic analyses, which are more focused on practical application, while Waltl et al. present a broader approach, across different document types, that is rooted in legal theory.

Waltl et al. (2015) also presented an approach for automated legal assessment, however, it is based on the traditional approach of trying to formalize the law in the form of deontic logic and is more focused on providing a framework in which such logical deductions can be created manually. One of the key concepts in legislation regarding consumer protection in standard form contracts is “vague clauses”, a concept which can hardly be formalized in simple logic deductions. We will therefore not try to formalize the law but rather codify the evaluation schemes and rules of thumb that consumer advocates have developed for the legal assessment of standard form contracts, based on laws, court decisions, and years of experience.

While we will present different approaches to automatically extract (see Section 6.4) and segment (see Section 6.2) T&C from websites, Waltl (2018) presents an architecture which enables the use of adaptors and importers which have to be written individually for each source and formatting, and are not the focus of their work.

From a technical perspective, Waltl (2018) has a stronger focus on rule-based approaches (specifically, Apache UIMA) and traditional ML, while we will also look at deep learning technologies, some of which did not even exist in 2018.

3.1.3. Privacy Policies

Since the commencement of the GDPR in 2018, there has been much focus on privacy policies, both from practitioners and researchers. A project that predates the GDPR is “The Usable Privacy Policy Project” by Sadeh et al. (2013). They chose a semi-automated approach, which combines NLP with crowd-sourcing to analyze privacy policies and compare them to user preferences. However, the amount of automation they use in order to analyze privacy policies is very limited (Ammar et al., 2012).

Harkous et al. (2018) presented a technologically more sophisticated approach that uses deep learning to categorize clauses in privacy policies into nine topics (1st party collection of data, 3rd party sharing of data, user choices, data security, specific audiences, access, edit, and delete, policy changes, data retention, and do not track) with an F1-score of 0.84. We will show in Section 6.3 that, despite having a much more fine-grained taxonomy with 23 topics, we are able to achieve better classification results on T&C with our approaches.

Torre et al. (2020) suggested to automatically analyze privacy policies in order to check them for completeness according to the criteria of the GDPR. They use keywords and text similarity, based on word embeddings, to find missing information obligations, which have to be present in privacy policies. In an evaluation with 24 privacy policies, their approach found 45 out of 47 missing clauses.

3.1.4. Contracts

Lee et al. (2019) presented an approach to automatically detect what they call "poisonous clauses" in international construction contracts. By their definition, a poisonous clause is a clause that provides "a risk that may adversely affect the contractor". They apply a rule-based approach which takes the plain text of the contract as input, segments it into sentences, and then applies a dependency parser to generate dependency trees. The manually created rules are then applied to these dependency trees to detect poisonous clauses, i.e., perform a binary classification. In their evaluation, they report a precision and recall of 81.8%, however, the evaluation is only based on 13 poisonous clauses. The technical approach Lee et al. chose, applying rules to dependency trees, is very similar to one of the approaches we applied and presented in Braun et al. (2018c), however, our rules are specifically tailored to different clause types and are based on legal requirements, while Lee et al. work on a more abstract, linguistic level.

Joshi et al. (2018) presented an approach to identify and extract location information from English contracts, more specifically, the locations of the involved parties and the jurisdiction. They use a naive Bayes classifier and decision trees with Tf-idf vectors as input to classify relevant clauses. For the detection of jurisdiction clauses, they report an F1-score of 94.54. We will show later that we can achieve an F1-score of 0.98 in German and 1.00 in English T&C for that task (see Section 6.3) by using logistic regression with Tf-idf vectors as input.

Chalkidis et al. (2017) compared different approaches, including handcrafted rules and logistic regression, to extract information like the title, the start and end date, governing law, and jurisdiction from contracts. With their best performing approach, logistic regression, they achieved F1-scores between 0.62 and 0.94, depending on the type of information, and a macro-average of 0.80. For jurisdiction, they achieved an F1-score of 0.81.

In addition to these (and other) scientific works, there is also a growing market of commercial tools for contract analysis, which we will talk more about in Section 3.4.

3.1.5. Laws

Similar to the work of Waltl et al. (2016a), Bartolini et al. (2004a,b) developed a system to automatically distinguish different categories of provisions (obligations, definitions, modifications) and their sub-categories (obligations: obligations, permissions, prohibitions, and penalties; modifications: replacements, insertions, and repeals) in Italian laws. Their system called “SALEM” uses dependency trees on which it performs a rule-based classification, similar to one of the approaches we and Lee et al. (2019) use. In an evaluation with 473 paragraphs from Italian law, SALEM achieved a precision of 0.97 and a recall of 0.96. The authors do not mention from which law the paragraphs were taken or how they were selected. Biagioli et al. (2005) presented an approach using SVM for the same task. They performed an evaluation using 582 paragraphs from Italian laws and achieved an accuracy of 92.44% when performing a leave-one-one-out evaluation.

de Maat and Winkels (2007, 2010) presented an approach to classify sentences in Dutch laws according to their provision type. They used a slightly different classification, which included the classes definition, permission, obligation, delegation, publication provision, application provision, enactment date, citation title, value assignment, penalization, change, mixed type, and a default class for all other types of sentences. In an evaluation with 592 sentences from eighteen Dutch laws, they were able to classify 91% of all sentences correctly.

For French laws, Lame (2004) presented an approach using Tf-idf to automatically extract ontology elements, including concepts and five kinds of relations: “Is a legal sort of”, “Is a general sort of”, “Is a component of”, “Is related to”, and “Is another sense of”.

Although this is by no means an exhaustive list of the work in the area, there is a noticeable shortage of literature performing semantic analysis on English laws, which is also reflected by broader literature reviews, e.g., by Zhong et al. (2020). This is unusual, given that in most NLP tasks, English is the predominant language. The reason is, most likely, that in the Anglo-American case-law system, statutory law, i.e., laws passed by parliaments and other bodies of legislation play a less important role and court decisions (see Section 3.1.6) a more important role.

A lot of the work using US law rather focuses on predicting whether a proposed bill will be enacted rather than analyzing enacted laws. Examples of such work include Nay (2017), Goldblatt and O’Neil (2012), and Gerrish and Blei (2011), who used different machine learning techniques to predict votes in the US congress, based on the content of a proposed bill. Since in the period from 2001 to 2015, only 3.6% of all proposed bills were enacted (Nay, 2017), evaluating approaches on a non-balanced data set favors models that simply always predict that a proposed bill will not be enacted (i.e., predict “no”). Gerrish and Blei (2011) report that they could achieve a 90% accuracy (85% for “yes” predictions) when ignoring temporal order and 87% (overall) for sequential prediction, i.e., using only older decisions to predict newer ones. On the same corpus, Goldblatt and O’Neil (2012) also reported an accuracy of 90% when ignoring temporal order. Nay (2017) reports an AUC of 0.96 for sequential prediction. As pointed out by Nay (2017), the relationship between the text and the votes these systems utilize is only correlational and other factors, like partisanship, play an important role in the actual decision making.

3.1.6. Court Decisions

Prediction of outcomes is also one of the most active fields of research when it comes to court decisions. As early as 1963, the US lawyer Reed C. Lawlor publicly discussed the idea of using computers to predict court decisions. He thought that “even if, as some seem to believe, the use of computers in aiding the prediction of judicial decisions be futile, great benefits may flow from the effort” (Lawlor, 1963). From today’s perspective, one might argue that the opposite has happened: We got awfully good in predicting court decisions with black-box deep learning approaches, which taught us nothing about the problem.

A lot of work exists regarding predicting decisions of the US supreme court, many of which claim to have reached superhuman performance[1] (see Ruger et al. (2004) and Martin et al. (2004) for a comparison of different approaches). Another court that is popularly targeted by research on decision prediction is the European Court of Human Rights (ECHR) (Aletras et al., 2016; Medvedeva et al., 2020; Kaur and Bozic, 2019). And even on lower levels, like the Court of Appeal for England and Wales (Strickson and De La Iglesia, 2020) or German financial courts (Waltl et al., 2017a), work has been carried out.

However, there is also a corpus of work that is concerned with the more in-depth semantic analysis of court decisions. Walter and Pinkal (2006) for example, presented an approach to extracting definitions from German court decisions with a precision of more than 70% by using the grammatical structure of a sentence and a list of keywords. Moens and Angheluta (2003) investigated the automated extraction of legal concepts from US court decisions. Chieze et al. (2008) built a tool that can extract information from Canadian immigration decisions and also automatically summarize them.

3.2. Knowledge Representation in the Legal Domain

Law is often purposefully drafted vaguely. Laws have to be vague in order to be flexible enough to still be applicable in a changing world, especially when it comes to technological development. Sometimes, legislators also leave some specifics up for interpretation by the jurisdiction on purpose. However, a law that is too vague might be unconstitutional and introduce uncertainty among citizens. (Hadfield, 1994; Raban, 2009) In addition to such purposefully introduced vagueness, our natural languages that are used to draft law are inherently vague.

For decades and centuries, philosophers, mathematicians, and legal theorists (and more recently also computer scientists) have worked on formalizing the law (Jones and Sergot, 1992; Royakkers, 2013; Navarro and Rodríguez, 2014) and legal reasoning (Breuker, 1993; Hage, 1996; Jøsang and Bondi, 2000). We will refrain from such ambitious, controversial, and arguably rather philosophical endeavors and instead focus on more practice-oriented formalization of knowledge representation within the legal domain.

[1] Comparisons against the predictions of human experts are commonly used metrics in court case outcome prediction. We would like to note that, by its very nature, especially on the level of supreme courts, for every court case, there are at least two opposing views from experts, one representing the prosecution and one representing the defense.

LegalRuleML (Palmirani et al., 2011), for example, is an XML-based open-standard for formalizing legal knowledge from laws, contracts, and judgments and allows to perform legal reasoning on these sources by providing deontic operators. With this ambitious goal, it is no surprise that LegalRuleML operates on a theoretical and rather abstract level. The Legal Knowledge Interchange Format (Gordon, 2010) is a comparable standard with less powerful expressions.

MetaLex (Boer et al., 2002) is another XML-standard for the representation of legal knowledge. However, instead of formalizing legal documents and conducting reasoning on them, it is purely focused on providing a common format to structure the content of legal documents.

Valente (2005) conducted a literature review on the usage of ontologies in the field of AI and Law and proposed a framework for the classification of types and roles of legal ontologies. The five roles Valente proposed are:

- organize and structure information
- reasoning and problem solving
- semantic indexing and search
- semantic integration and interoperation
- understanding the domain

In this classification, the knowledge-base we present in Chapter 7 fulfills the roles of semantic indexing and search and reasoning and problem solving. The first, because we will use the underlying ontology to extract structured information from clauses and the latter because we will use the information stored in the knowledge-base to assess clauses legally. The type system that Valente proposes for legal ontologies is based on the format the information is stored in the system. He differentiates between less structured types, like audio files, videos, and raw texts, and more structured types, such as structured textual information (e.g., XML), databases and tuples, categorized information like taxonomies, and formal knowledge, like logic formulas.

We will combine the three most structured types in our knowledge-base: we will use tuples to store the information extracted from clauses, a taxonomy for the classification of clause topics, and formal knowledge to assess the clauses legally.

From the thirteen legal ontologies that Valente found in the literature, only six were using an existing standard to model their ontology (four of which used OWL and two the Knowledge Interchange Format). We decided not to use any of the existing formats and instead created a very lightweight JSON-based knowledge-base. Many of the existing formats introduce massive overheads, which we do not need since we are not planning to model the whole domain or all aspects of it, but have a very sharp focus on which information we want to model and how we want to use this information.

3.3. Text Summarization

In this section, we will give an overview of the state-of-the-art in text summarization. The content of the section is based on a more extensive literature review that we published earlier (Klymenko et al., 2020).

Manual text summarization is a time-consuming, difficult, and therefore expensive task. Automatic text summarization is the process of automatically shortening a text in a way that retains its most important information. The goal of summarization systems is to produce a concise and coherent summary that allows readers to understand the content of the input text without having to read it entirely. A good summary should be:

- fluent,
- consistent,
- contain all important information,
- but no unnecessary or duplicate content.

Automatic text summarization has applications in many domains. It can be used by companies for the generation of reports, by students and scientists to find relevant research, by physicians to summarize medical information, by journalists in order to find relevant sources, and, as we will show later, to summarize standard form contracts for expert and non-expert readers.

For a long time, so-called "extractive" techniques (see Section 3.3.1) have been the primary focus of research. Recently, the focus has shifted to so-called "abstractive" summarization (see Section 3.3.2), driven by advances in ML.

There is a number of scientific works, e.g., by Nenkova et al. (2011), Lloret and Palomar (2012), and Saggion and Poibeau (2013), that provide an extensive overview of different automatic summarization methods that were proposed ever since the publication of the first paper on the topic in 1958 by Luhn. Here, we, therefore, want to focus on more recent advances and approaches that are currently considered to be state-of-the-art, as well as describe the techniques that are currently used to evaluate summaries automatically.

3.3.1. Extractive Methods

Extractive summarization methods generate summaries by concatenating sentences or text units from the source text exactly as they occur. The main challenge, therefore, is to determine which units are most important and should be included in the summary. Recent approaches treat the task as a sequence labeling problem, in which each text unit is labeled with the information whether or not it should be included in the summary. In the following, we will shortly introduce some of the most recent work.

Cheng and Lapata (2016) developed a data-driven single-document summarization framework based on a hierarchical document encoder and an attention-based extractor. It enables the development of different classes of summarization models, which can be trained on large-scale data sets and learn informativeness features based on continuous representations without any access to linguistic annotation. The authors tested their approach on different corpora and reported results that are comparable to other state-of-the-art approaches.

Nallapati et al. (2017) presented “SummaRuNNer”, a RNN-based sequence model for extractive single-document summarization. Their approach tries to identify a set of text units, which collectively achieve the highest ROUGE score, a metric for the automated evaluation of summaries which we discuss in Section 3.3.3, with respect to the gold standard. They also provide an explanation component, which visualizes additional information about the predictions, such as information content, salience, and novelty. The approach was evaluated on three data sets: The Daily Mail and joint CNN/Daily Mail corpus from Hermann et al. (2015) and the DUC 2002 corpus (of Standards and Technology, 2002). The authors compared their results on the Daily-Mail corpus with Cheng and Lapata (2016) using ROUGE recall with a summary length of 75 and 275 bytes. The authors report that their model performs significantly better for a summary length of 75 bytes and similar for a summary length of 275 bytes. On the joint CNN/Daily Mail corpus, the authors compared their work against Nallapati et al. (2016), which was the only work at the time, that reported performance on this corpus, and report that their model performs significantly better. On the DUC 2002 corpus, the authors report to achieve similar results as Cheng and Lapata (2016), however, could not match the results reported by Parveen et al. (2015) and Wan (2010).

Narayan et al. (2017) used an architecture similar to those of Nallapati et al. (2017). For sentence ranking, however, they introduced a global optimization framework, which combines the maximum-likelihood cross-entropy loss with rewards from policy gradient reinforcement learning to optimize the final evaluation metric, ROUGE. The model was applied to the CNN/Daily Mail data set and automatically evaluated using ROUGE, outperforming both of the systems mentioned above, as well as some of the most prominent abstractive systems (Nallapati et al., 2016; Chen et al., 2016; See et al., 2017), which we discuss in the next section. However, their approach of directly optimizing for ROUGE score might be seen as problematic, especially given the rather weak link between the score and human evaluation, which some studies have found (see Section 3.3.3).

Mehta (2016) argued that extractive summarization methods might already have reached their peak performance. Mehta therefore suggests to either focus on ensemble models of existing extractive methods or focus on abstractive techniques.

3.3.2. Abstractive Methods

Choosing a subset of text units from the source text as summary has multiple inherent drawbacks. Among the most severe ones is that the produced summary lacks cohesion and coherence, and meaning might be lost due to the fact that sentences are taken out of context. In order to generate summaries of human quality, it is not enough to just extract sentences from the source text; we have to generate a new, coherent text as summary. Abstractive summarization methods produce summaries by generating new text, which was not present in the source document. However, this also makes abstractive summaries more challenging because we do not only have to find the most important information in the source, which in itself is a challenging problem but also

Extractive	**Abstractive**
When we tried to hook up the first one, it was broken - the motor would not eject discs or close the door.[1] The build quality feels solid, it doesn't shake or whine while playing discs, and the picture and sound is top notch (both dts and dd5.1 sound good).[2] The progressive scan option can be turned off easily by a button on the remote control which is one of the simplest and easiest remote controls i have ever seen or used.[3] It plays original dvds and cds and plays mp3s and jpegs.[4]	Customers had mixed opinions about the Apex AD2600[1,2] possibly because users were divided on the range of compatible disc formats[3,4] and there was disagreement among the users about the video output[5,6]. However, users did agree on some things. Some purchasers found the extra features[7] to be very good and some customers really liked the surround sound support[8] and thought the user interface[9] was poor.

Table 3.2.: Comparison of an extractive and abstractive summary (from Carenini and Cheung (2008); numbers indicate which review the facts are taken from)

transform it into coherent and grammatically correct texts, which, again, in itself is already a challenging problem.

Carenini and Cheung (2008) performed a user study comparing extractive and abstractive summarizers. Table 3.2 shows an example from their paper, with an extractive and an abstractive summary of multiple product reviews. The numbers in the text refer to the source review from which information was taken. The example highlights some of the core differences between the two techniques:

- In the extractive summary, each sentence originates from one specific source. In the abstractive summary, a sentence can contain information from multiple sources.
- Because the sentences in the extractive summary are taken out of their context, the reader might sometimes not be able to understand them anymore, e.g., because referring expressions cannot be resolved anymore (like the "it" in the last sentences of the extractive summary).
- While the abstractive summary is able to point out disagreement specifically ("there was disagreement"), the extractive summary simply concatenates opposing statements without any transition ("it was broken" - "the build quality feels solid"), which feels unnatural to readers.

Most recent research on abstractive summarization relies on sequence-to-sequence models (Sutskever et al., 2014), which were first introduced as Encoder-Decoder Models by Cho et al. (2014) and later extended by Bahdanau et al. (2014) with so-called "attention mechanism". Originally, sequence-to-sequence models were developed for machine translation.

A group of researchers at Facebook (Rush et al., 2015) presented a fully data-driven approach to abstractive sentence summarization. They use neural attention-based model that generates each word of the summary conditioned on the input sentence. This probabilistic model is combined with a generation algorithm that produces accurate abstractive summaries. The attention-based model provides less linguistic structure comparing to other abstractive summarization approaches but is easily scalable for training on large amounts of data. Furthermore, the lack of vocabulary constraints in the system makes it possible to train the model on diverse input-output pairs. Chopra et al. (2016) later extended that model to a RNN architecture. Although these approaches are considered to be state-of-the-art, they are far from perfect in that they sometimes inaccurately reproduce factual details, are unable to deal with out-of-vocabulary words and can only deal with very short documents.

Chen et al. (2016) explored neural summarization technologies for articles that contain thousands of words. Their model is also based on an encoder-decoder framework, however, instead of focusing on attention to get the local context like most of the recent work does, they incorporate a coverage mechanism to “distract” the model to different parts of a document to avoid focusing on only one aspect of the source text. Without engineering any features, they trained the models on two large data sets and test them on LCSTS corpus (Hu et al., 2015). The proposed approach achieved better performance than the best result reported by Hu et al. (2015).

See et al. (2017) proposed a novel architecture that enhances the standard sequence-to-sequence attentional model with a hybrid pointer-to-pointer generator network and a coverage mechanism. The resulting hybrid network is similar to the ones proposed by Gu et al. (2016) and Miao and Blunsom (2016). It uses pointing (Vinyals et al., 2015) to copy words from the input text, which provides better accuracy and is better in dealing with out-of-vocabulary words while retaining the ability to generate words. To ensure coverage of the input document and thus reduce repetitions in the summary, the authors used an adapted version of the coverage vector model by Tu et al. (2016). The model was applied to the CNN/Daily Mail data set and performed better than the system by Nallapati et al. (2016).

3.3.3. Automatic Evaluation

Evaluating the quality of text summarizations - manually or automatically - is a challenging task. Not only can different humans create very different summaries for the same text, even the same person can create significantly different summarizations for one text. The phenomenon was, e.g., shown by Rath et al. (1961). Not only did they find a low agreement rate between different subjects when it came to producing an extract of a text, they also found that a subject may produce a significantly different extract when asked to summarize the same document eight weeks later. Moreover, manual evaluations require a lot of time and effort and are, therefore, expensive. To address these problems, automatic methods for the evaluation of summaries have been developed. The three arguably most popular ones are BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). Of the three, ROUGE is the only metric that was designed specifically to evaluate text summarization. BLEU and METEOR were developed to evaluate machine translation.

3.3.3.1. BLEU

BLEU is short for bilingual evaluation understudy and was introduced by Papineni et al. (2002) to automatically evaluate machine translations. The underlying idea is to calculate the quality of a machine translation by computing the precision. In order to calculate the precision, two numbers are needed: the number of words that occur in the machine translation and in human translations of the same sentence and the total number of words in the machine translation. This naive approach has two drawbacks: It rewards (or at least not punish) over-generation and does not take into account the position of words. If the human translation would, e.g., be “The cat sat on the mat.”, the naive precision score would judge the machine translation “the the the.” with 0.66, although the “translation” is nonsense. The translation “Mat the the on cat sat” would even receive the “perfect” score of 1.0.

In order to overcome these shortcomings, the idea of BLEU is to calculate the precision for n-grams rather than unigrams. If we use bigrams in the above example ((the, cat), (cat, sat), (sat, on), (on, the), (the, mat)), the translation “the the the.” would score 0, and “Mat the the on cat sat” would score 0.2. An evaluation using bigrams is also referred to as BLEU-2, using trigrams as BLEU-3, and so on. Instead of using just one type of n-grams, they can also be combined to form a cumulative score, which takes into account, e.g., weighted unigram, bigram, and trigram precision.

3.3.3.2. ROUGE

ROUGE is short for recall-oriented understudy for gisting evaluation and was introduced by Lin (2004). Instead of being based on precision, ROUGE is based on recall. Therefore, ROUGE is less punitive towards over-generation and words that do not occur in the human reference and is considered more appropriate for the evaluation of summarization tasks.

There are several variations of ROUGE:

- ROUGE-N is based on an n-gram recall between a candidate summary and a set of reference summaries, i.e., the score is computed as the ratio between the number of shared n-grams across the candidate, and the references and the number of n-grams in the reference summary.
- ROUGE-L takes into account the longest common sub-sequences instead of just n-grams. Advantages of this metric are that it only requires in-sequence matches but not consecutive matches, which is more appropriate for shorter texts, automatically includes longest in-sequence common n-grams and takes into account sentence-level structure in a natural way.
- ROUGE-S uses skip-bigram co-occurrence, i.e., unlike for bigrams, where two words have to occur in direct sequence, an arbitrary of words can occur in between these two, as long as the first word of the skip-bigram still precedes the second.

- ROUGE-SU is an extension of ROUGE-S which takes into account unigrams in addition to skip-bigrams, thus avoiding the assignment of zero scores to a sentence which does not share a skip-bigram but has common unigrams (e.g., “This is a cat.” and “Cat a is this.”). Correlation analysis results on DUC summarization data show that ROUGE-SU correlates the best with human judgments.

The fact that ROUGE relies purely on lexical overlaps, which can significantly underrate summarization score, especially for documents that contain many synonyms, terminology variations, and paraphrasing, is one of its main shortcomings. In a detailed analysis of ROUGE’s effectiveness for evaluating scientific summaries by Cohan and Goharian (2016), ROUGE proved to be unreliable for the assessment of such summaries.

3.3.3.3. METEOR

METEOR stands for metric for evaluation of translation with explicit ordering and was introduced by Banerjee and Lavie (2005). Instead of using precision or recall as the base, METEOR calculates the harmonic mean of precision and recall, i.e., the F1-score. Unlike BLEU or ROUGE, METEOR only uses unigrams but considers not only exact word matches but also stemmed forms and synonyms.

3.3.3.4. Validity

Automatic evaluation metrics like BLEU, ROUGE, and METEOR have become omnipresent. With the increasing size of corpora and models, it becomes more and more expensive to conduct human evaluations. With their precise continuous numbers as results, automated metrics *seemingly* offer a higher degree of comparability than sometimes somewhat vague evaluation of humans. A property that seems especially desirable when parts of the research community have started to focus on improving the state-of-the-art by the hundredth or thousandth part.

The claim of these metrics to represent a qualitative analysis in their respective domains is based on the assumption that their scores correlate with human evaluation, i.e., if a text is perceived as a better translation or summary by humans, the inventors of these automated metrics claim, that they will also achieve a higher score.

Whether this is really the case, however, is debatable. Liu and Liu (2008), for example, investigated the correlation between human judgment and ROUGE scores on extractive meeting summaries and only found a low correlation. Belz and Reiter (2006) compared the correlation of BLEU, ROUGE, and NIST with human judgment on texts generated by NLG systems. They found that NIST had the highest correlation, followed by BLEU. For ROUGE, they conclude that it provides little advantage over simple metrics such as string-edit distance. They also found that whether or not the metrics correlate with human judgment depends strongly on the (perceived) quality of the gold standard. Reiter (2018) conducted a broad literature review, in which he analyzed the reported correlation between BLEU and human judgment from 34 papers. He concluded that BLEU seems to have a bias towards certain technologies and that whether or not BLEU correlates with human judgments depends on contextual factors that are unclear.

Given all these factors, it is not surprising that Reiter (2018) concludes that "BLEU should not be the primary evaluation technique in NLP papers". More generally speaking, van der Lee et al. (2019) suggest to "always conduct a human evaluation (if possible)". We will follow their suggestion and evaluate our summarization system by performing a human evaluation.

3.4. Commercial Tools for Contract Analysis

The analysis of standard form contracts and contracts more generally, is not only of scientific relevance but also commercially relevant, as we will also show in Chapter 4. Accordingly, the problem of contract analysis is not only tackled by researchers but also by corporations with commercial interests. Unfortunately, most of these companies are rather secretive about their products and even more so about the underlying technology. For researchers, it is often difficult to get in touch with these companies to get additional information on their products. Therefore, we partnered with a company from the financial sector, which was looking for a software to support its legal department in the analysis of large inter-bank loan agreements. European inter-bank loans are largely based on standardized loan agreements by the European Loan Market Association (Gadanecz, 2004) and can therefore be considered a type of standard form contracts. Through this partnership, we were able to conduct a survey of the market for contract analysis tools.

The information in this section is based on presentations given by software vendors and interviews we conducted with the vendors. While we believe, based on live demonstrations, that the information given to us is credible, we have no way to confirm it. We still believe that it is valuable not only to consider the state-of-the-art in science but also in the industry.

3.4.1. Vendors

In the first step, we conducted a broad market survey by identifying as many vendors in the market as possible and the contract analysis products they offer. The result of this search is shown in Table 3.3. While there are a few established vendors that entered the market, like iManage LLC, best known for their document management systems, most tools are developed by start-up companies. And even the tool that is offered by iManage was originally developed by a start-up that the company acquired in 2017. Another interesting pattern that is visible is that most companies that are not headquartered in the US are based in London. This is most likely an expression of the fact that London is one of the most important financial centers in the world.

3.4.1.1. Analysis of Tools

In consultation with our partner company, we chose six tools for which we contacted the vendors to request a presentation with a live demo and a subsequent interview. We only chose six tools due to time constraints and because several tools were ruled out by our partner based on the product websites. Before the presentations, we developed a catalog of criteria, based on the

Tool	Company	Origin
archii	Archii ApS	Copenhagen, Denmark
ANALYZELAW	IntraFind Software AG	Munich, Germany
ANVI	Ganot Lex Solutions Private Ltd.	Hyderabad, India
Change-Pro	Litera Corp.	McLeansville, NC, USA
Eigen Technologies	Eigen Technologies Ltd.	London, UK
H4	H4 Ltd.	London, UK
iManage RAVN	iManage LLC	Chicago, IL, USA
Kira Systems	Kira Inc.	Toronto, Ontario, Canada
Legito	Legito s.r.o.	Brno, Czech Republic
Lexical Labs	Lexical Consulting Ltd.	London, UK
LexPredict	LexPredict, LLC	Chicago, IL, USA
Luminance	Luminance Technologies Ltd.	London, UK
Nammu21	Nammu21, Inc.	New York, NY, USA
Primer	Primer Technologies Inc.	San Francisco, CA, USA
rfrnz	rfrnz GmbH	Munich, Germany
Seal	Seal Software Inc.	CA, USA / London, UK
SoftLaw	SoftLaw SAS	Paris, France
Summize	Summize Ltd.	Manchester, UK
Things Thinking	thingsTHINKING GmbH	Karlsruhe, Germany

Table 3.3.: Vendors of commercial solutions for automated contract analysis and their products

needs of our partner company, by which we evaluated the tools. The drafting of this catalog was very insightful because many of the criteria that were important for our partner are rarely considered in scientific works. While, for reasons of confidentiality, we cannot disclose all criteria, the ones we can disclose are still interesting to consider:

- **Input format**: Many of the contracts are stored in Microsoft Word-format (.doc/.docx), therefore, it is important for the company that the tool is able to process these formats, among more standard formats like PDF. This is also an interesting difference, compared to the consumer contracts we investigate in this thesis, which are virtually never received in editable formats like Word.
- **Deployment**: In the financial sector, rules about privacy protection are even more rigorous than in other industries. Therefore, only solutions that can be deployed on-premise are viable options.
- **Analytic power**: One of the central criteria is the analytic power, in terms of NLP, that the tools provide.
- **Required training samples**: Although data scarcity is also a common problem in research and lately received more attention in the form of few-shot learning (see, e.g., Yan et al. (2018)), many existing approaches in the scientific literature are still based on massive data sets, while especially smaller companies often only have small data sets of a few hundred contracts. Therefore, it is important for companies how much data is required to train stochastic elements that are offered by contract analysis tools.

	Input				Deployment		
Tool	**DOCX**	**PDF**	**Img**	**OCR**	**Cloud**	**Local**	**Training Samples**
ANVI	✓	✓	✓	✓	✓	✓	> 30
iManage RAVN	✓	✓	✓	✓	✓	✓	> 30
Lexical Labs	✓	✓	-	-	✓	✓	10-20
Primer	-	✓	-	-	✓	✓	500 - 1,000
Seal	✓	✓	✓	✓	✓	✓	10,000
Summize	✓	✓	✓	✓	✓	✓	X

Table 3.4.: Overview of the evaluated tools

Table 3.4 shows the six tools which we analyzed and how they performed in three of the four criteria. The most notable differences are visible for the number of training samples that the tools need. Summize was the only tool in the comparison that does not allow the user directly to train their own models but offers pre-trained models for similarity detection. Three providers claimed that their tool needs between 10 and 50 contracts to train a working model for clause topic detection and other tasks. It is difficult to assess whether this is actually true and also, as usual, strongly depends on the training examples.

Generally speaking, the types of analysis the different tools offer are quite similar. All of them offer clause topic labeling, either based on a set of defined keywords or based on machine learning. Most also offer a similarity search, which finds similar clauses in other documents. All of them also offer a standard full-text search, some only in single documents, others in a whole corpus of documents. The tool "ANVI" also offers document classification to differentiate between different types of contracts (e.g., Non-Disclosure Agreement (NDA), Loan Agreement, ...). Only one Tool, Lexical Labs, offers information extraction to turn the unstructured information contained in a clause into structured information. All of the tools offer a web-based User Interface (UI) and some also collaboration features like writing comments or assigning tasks to other users.

In summary, the tools we evaluated do not offer a deep semantic analysis of the contracts but usually stay on the level of topic labeling. Because they claim to work with any kind of contract, it would also be very difficult to offer a deeper analysis of the content. We, on the other hand, will focus on one specific kind of contract, consumer standard form contracts, and therefore will be able to conduct an in-depth semantic analysis and even a legal assessment, which none of the existing tools offer.

3.5. T&C Generators

Another class of commercial tools that is closely related to our research are T&C generators. They promise tailored and legally secure T&C to a price much lower than what a law firm would charge to draft such a document. Usually, they are based on a series of selection options and form fields, which are used to adapt a template text. Most services offer a free and a premium version. The premium version usually offers a liability warranty, in case the T&C turn out to be void, and an update service in case relevant laws change. Example of such generators include Trusted Shop[2], janolaw[3], Termify[4], and Termly[5]. In this section, we will show on the example of Trusted Shop, which is focused on the German market, and Termly, which is focused on the US and British market, how such generators work and what the T&C they generate look like.

3.5.1. Trusted Shops

Trusted Shops is a company located in Cologne, Germany, which specializes in auditing and certifying online shops. They also offer generators for four different types of legal texts: imprints, T&C, privacy policies, and cancellation policies. The texts are by default generated in German and based on German laws.

In order to generate T&C, the user has to fill in forms for six major categories: general information, contract, payment, shipping, restriction, and service. The following enumeration lists all questions which the users are asked during the generation.[6]:

- General
  - For which platforms are the T&C?
    - Online shop
    - eBay
    - Amazon
    - Hood.de
  - What is the URL of the shop?
  - Who are the customers?
    - Consumers (B2C)
    - Consumers and Companies
    - Companies (B2B)
- Contract
  - Company name
  - How is the contract concluded in your online shop?
    - By submission of the order by the customer
    - By declaring acceptance by the vendor

[2] `https://www.trustedshops.com`

[3] `https://www.janolaw.de`

[4] `https://termify.io/`

[5] `https://termly.io/`

[6] All questions are direct quotes from the Trusted Shops website (`https://www.trustedshops.de/`), last accessed 2020-09-15

    - Do you store the text of the contract in your system?
    - Which languages are available in your online shop?
        - German
        - English
        - ...
- Payment
    - Which forms of payment do you offer?
        - Credit card
        - Direct debit
        - PayPal
        - ...
- Shipment
    - Do you charge delivery fees?
    - Do you deliver to pick-up stations?
    - Can the customer pick-up their order at your store?
- Restrictions
    - Do you want to include a retention of title?
    - Do you want to ask the customer to help with transport damage without obligation?
    - Do you want to include regulations about transport damage for companies?
    - Do you want to restrict warranty?
    - Do you want to restrict liability?
    - Do you offer age-restricted goods (e.g., alcohol)?
    - Would you like to make a choice of law towards companies (German law)?
    - Are you a merchant within the meaning of the German commercial law?

An example for T&C generated by the Trusted Shops can be found in Appendix D.1. If the free version of the tool is used, the phrase “AGB erstellt mit rechtstexter.de.” (“T&C generated with rechtstexter.de.”) has to be put at the end of the T&C. A Google search for that exact phrase in May 2020 brought 103 results. In order to get an estimate for how many online shops use the service, including the premium version, we chose one of the headings generated by the service, which we believe to be rather distinctive and unlikely to appear in T&C that were not drafted using the Trusted Shops service (“2. Vertragspartner, Vertragsschluss, Korrekturmöglichkeiten”, “2. Contracting Party, Conclusion of Contract, Correction Options”). This search returned 293 results on Google.

3.5.2. Termly

Termly is a young start-up headquartered in the US with offices in the UK and Taiwan. Their only business is generating legal texts for websites, more specifically T&C, privacy policies, refund & return policies, disclaimer, and cookie consents. The approach for generating these documents is very similar to the approach taken by Trusted Shops: Termly asks users to fill in forms for five major categories: company, website, website content, disputes, and final details. The questions the user has to answer are[7]:

- Company
  - Fill legal name of company
  - Are you doing business under a short form or trade name (also known as a D/B/A)?
  - Address
- Website
  - URL
  - Which country is your website hosted in?
  - Does your website offer goods and services OR target users/customers in the United States?
  - Can users under the age of 18 use your website?
  - Can users under the age of 13 use your website?
  - Can users create an account on your website?
  - Can users link their accounts to their social media accounts?
  - Does your website have a mobile application?
- Website content
  - Does your website have a privacy policy?
  - Please provide the link to your privacy policy.
  - Please select the activities you would like to prohibit your users from engaging in on your website:
    * Systematically retrieving data or content from your site to create a collection or database without written permission from you.
    * Trick, defraud, or mislead you or other users, especially in any attempt to learn sensitive account information such as user passwords.
    * ...

[7] All questions are direct quotes from the Termly website (`https://termly.io/`), last accessed 2020-09-15

  - Can users upload or post content on your website? (e.g., comments, articles, photos, audio, videos, etc.)
  - Can users submit reviews of the products on your website?
  - Does your website link to any other websites that you do not own?
  - Do you have third-party advertisements on your website?
- Disputes
  - Do you want to include a clause regarding how users can notify you regarding copyright infringements?
  - How will disputes between you and your users be resolved?
    - Litigation
    - Arbitration
    - Informal negotiations, then arbitration
  - In which country will it take place?
  - Do you want to limit the amount of your liability for any claims against your website?
  - Do you want to limit the length of time that you can be liable for a user claim?
- Final details
  - Do you want to add your own clause to these terms?
  - Which country's laws will govern this Terms of Use?
  - What is the effective date for this Terms of Use?

An example of T&C generated by Termly can be found in Appendix D.2. Although the service is also specifically marketed for online shops, the questions show that the focus is less on the buying process and more on the usage of the website. In the free version, the sentence "These terms of use were created using Termly's Terms and Conditions Generator." has to be written at the end of the T&C. A Google search in May 2020 found 280 pages. A search for a phrase that we believe to be distinctive ("We will alert you about any changes by updating the "Last updated" date of these Terms of Use, and you waive any right to receive specific notice of each such change.") did not bring additional results.

The two examples of Trusted Shops and Termly show that T&C generators are fairly simplistic template selection machines that either include or exclude certain text blocks based on the user choices and fill in some blanks within these text blocks. Still, the questions they ask and the types of clauses they generate help us to understand what some of the most commonly used building blocks for T&C are.

CHAPTER 4

Empirical Relevance of Standard Form Contracts

One of the key aspects of action research is that it addresses real-world problems. While most people will have a good intuition about whether unread standard form contracts and possibly void clauses are a real-world problem, not least from personal experience, we will underpin this intuition with data in this chapter. First, we will shortly recap the relevance of standard form contracts in our modern economy, which we already touched in the introduction (Chapter 1). Subsequently, we will show why the automated semantic analysis, legal assessment, and summarization of standard form contracts will be relevant to both consumers (Section 4.2) and consumer advocates (Section 4.3).

For practical reasons, we will focus our analysis on a subset of clauses that can typically be found in standard form contracts and especially in T&C from online shops. In Section 4.4, we explain how we chose this subset and why it supports our goals of empirical relevance and consumer protection.

Whenever we conduct research, but especially when we try to tackle real-world problems, we have to think about the ethical implications our work could have. We are driven by the goals of tackling the existing imbalance of power, democratizing the access to legal advice, and supporting consumers. Ethical considerations should neither be an afterthought, nor can they be dealt with exhaustively in one section, but have to be taken into account during each step. Nevertheless, we want to discuss some of the more general ethical implications that arise from the goals we pursue at this point in Section 4.5.

4.1. Relevance for the Economy

Our modern economy is unimaginable without the use of standard form contracts, or as Marotta-Wurgler and Taylor (2013) put it, “standard-form contracting is the engine of the mass-market economy”. Businesses like Google and Amazon, but also more traditional businesses like banks, car rental, and utility providers, could not operate at scale anymore if they were to negotiate individual contracts with each of their customers.

Standard form contracts are not only crucial for businesses, but also from a consumer perspective. Some of the biggest decisions we make in our lives are connected to standard form contracts, from buying a house to getting insurance; even prenuptial agreements can be found in boilerplate versions on the internet. And with the increasing importance of e-commerce, they will only get more important.

While there is no sound estimate on how often per day we are confronted with standard form contracts in general, McDonald and Cranor (2008) estimated that, based on the average amount of new websites an average US citizens visits per year, they would need to read 1,462 privacy policies per year. Given an average reading time of ten minutes per policy, McDonald and Cranor calculated that the average citizen would spend 244 hours (a bit more than ten days) per year reading privacy policies, which would lead to a national reading time of 53.8 billion hours and costs of 781$ billion per year for reading privacy policies, not including all the other types of standard form contracts we face daily. Since McDonald and Cranor estimate that more than 30% of that time would have to be spend at work, it is fair to say that, to a certain degree, our economy depends on the fact that we do not read all the standard form contracts we agree to. These numbers show that the automated assessment of standard form contracts could be of great relevance and value for both consumers and the economy as a whole.

Standard form contracts and their clauses are also regularly the subject of legal proceedings. The EUR-Lex database[1] of the EU lists 260 legal proceedings at the Court of Justice of the European Union concerning the Council Directive 93/13/EEC of 5 April 1993 on unfair clauses alone between 2008 and June of 2020. In the same time frame, the German legal database Juris[2] lists 6,520 court rulings related to the control of individual T&C clauses at all levels of the German legal system. In addition, there is an unknown number of cease-and-desist orders send each year, which is likely to be in the thousands. The “Verbraucherzentralen” alone send out 634 cease-and-desist orders in 2019 (Verbraucherzentrale Bundesverband e.V., 2020b), a record high, after 266 in 2018 (Verbraucherzentrale Bundesverband e.V., 2019b) and 281 in 2017 (Verbraucherzentrale Bundesverband e.V., 2018b).

These numbers show that standard form contracts do not just have a big economic value but also are regularly the subject of legal proceedings. Therefore, their analysis and legal assessment are also of great practical value.

[1] `https://eur-lex.europa.eu/`

[2] `https://www.juris.de/`

4.2. Relevance for Consumers

As we have outlined in Section 1.1.1, there are plenty of studies showing that consumers, most of the time, do not read T&C before they order online or register for a service (Hillman, 2005; Plaut and Bartlett III, 2012; Bakos et al., 2014; Obar and Oeldorf-Hirsch, 2020). We wanted to get a better understanding of why consumers do not read T&C, if they care about void clauses, and other questions surrounding the behavior of consumers towards T&C.

In order to answer these questions, we conducted an online survey. Since we focus on German standard form contracts, we constructed the survey in German. The initial draft of the survey consisted of ten questions. We conducted a pilot study with five PhD candidates to gather feedback on the clarity of the questions and possible missing aspects. Based on the feedback, we changed the text of multiple questions and added two additional ones. The answers given by the participants of the pilot study were not used for further evaluation.

For the conduction of the main study, we used the service Prolific, "a subject pool for online experiments" (Palan and Schitter, 2018), comparable to Amazon's Mechanical Turk, however, specifically designed for research purposes and with higher standards regarding both workers and their payment. We conducted a second pilot with ten participants recruited through the platform to check for technical problems and get an estimate for the time needed to complete the survey. Since there were no technical issues, no further changes were necessary, and we could include the results in our final analysis. The average completion time was around 3:30 minutes. We paid each participant 0.57 GBP, which, assuming a completion time of four minutes, corresponds to an hourly rate of 8.55 GBP or 9.47 EUR, which is slightly above the German national minimum wage of 9.35 EUR. Based on the actual time spent by the participants, the average pay per hour, in the end, was even slightly higher at 9.04 GBP. We only invited participants whose first language is German and who are currently living in Germany to participate. We decided not to collect any further demographic data because of privacy and data handling reasons, but also because they were simply not relevant to us. In the following, we will describe the insights we gained from the survey and how they support the empirical relevance of our research. In total, we got responses from 100 participants.

The first question we asked our participants was, how often they buy something online, because one of our hypotheses was that people who buy online more often are less likely to read the T&C before they order. As Figure 4.1 shows, 34% of the participants are buying online at least once a week, and only 11% order less than once a month.

The next two questions asked how often participants do read T&C before they order something at a new online shop (Figure 4.2) and how often before registering for a new service. At large, the responses to these questions confirm the finding of previous studies. Only a marginal amount of participants regularly reads T&C. 32% - 34% never reads any T&C at all, and the majority of participants claimed that they read them in some cases, however in less than 25% of all cases. Participants seem to be a little bit more likely to read T&C when registering for an online service, compared to online shopping; however, the difference is only marginal. In any case, we have to remember that these figures are based on self-assessment and assume that the real numbers might be even lower.

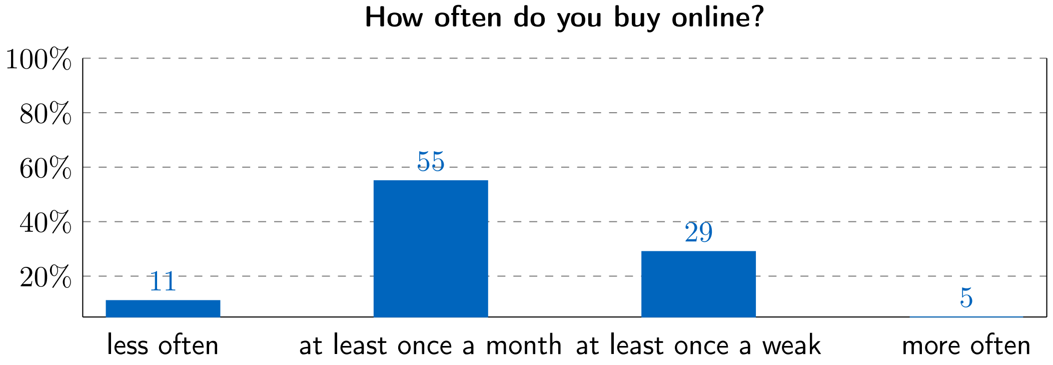


Figure 4.1.: Responses in percent for the question “How often do you buy online?”

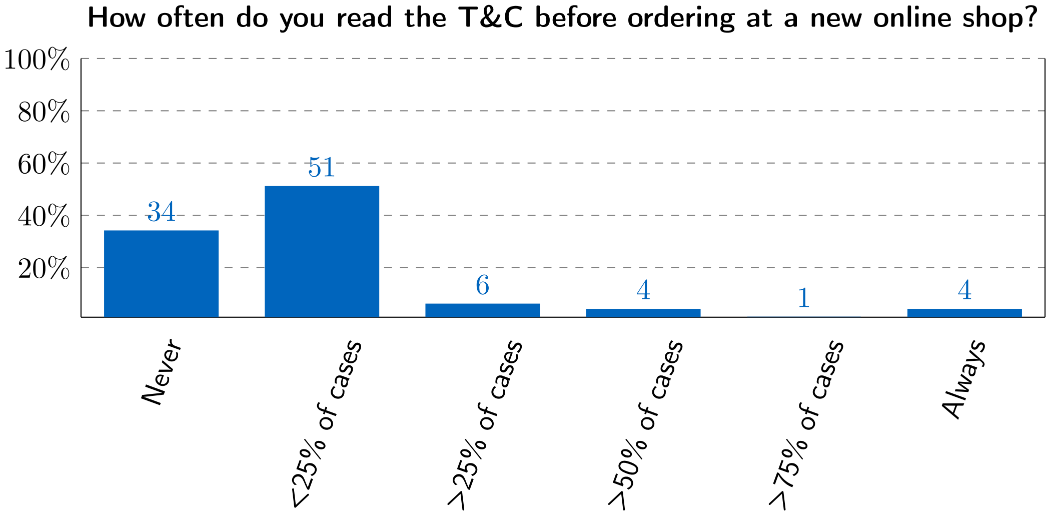


Figure 4.2.: Responses in percent for the question “How often do you read the T&C before ordering at a new online shop?”

In both cases, the participants who order multiple times a week responded all that they read the T&C in less than 50% of the cases, however, a Chi-squared test of independence showed that there is no statistically significant influence of the order frequency on how often somebody reads the T&C ($p = 0.9781$).

We asked participants who responded that they sometimes read T&C, but not always, to fill in a free text field to explain how they decide when they read T&C. The answers we received can be categorized into eight groups. Most participants said that they would read the T&C if the vendor makes a dubious first impression, e.g., because of the look of the website (21 participants). The second most common answer was that participants would read the T&C if they are looking for a specific point (19). For most participants, such points would be either data protection or cancellation policies. For 14 participants, the amount of the transaction they are about to make is decisive when deciding whether or not to read the T&C. Ten participants responded they read T&C that are short. Another common theme was that participants read

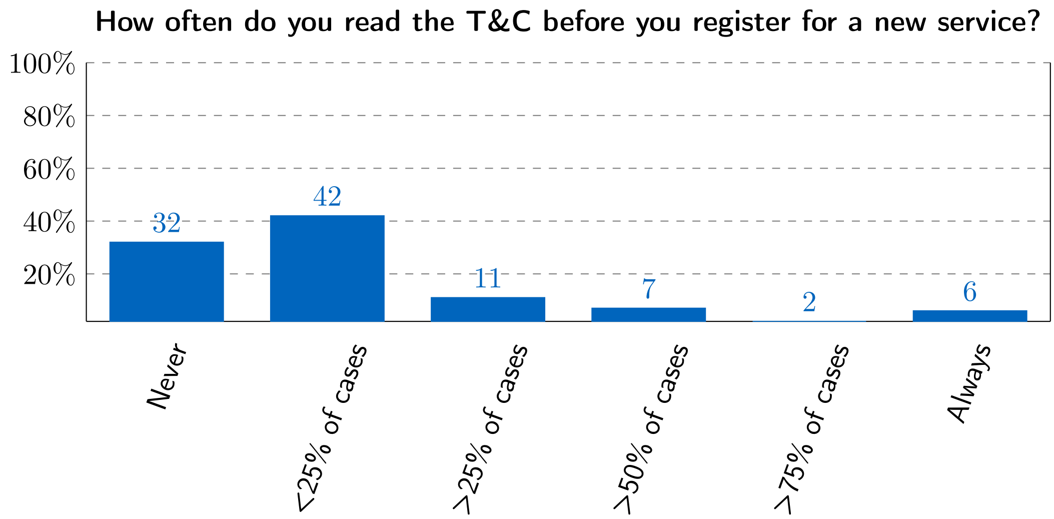


Figure 4.3.: Responses in percent for the question “How often do you read the T&C before you register for a new service (e.g., social media)?”

T&C if the shop is either in general not well-known (5), not known by any of their friends (4), or has bad reviews (4).

In the next question, we turned the tables and asked all participants what are reasons for them to *not* read T&C. We gave them five options to chose from (multi-selection was allowed) and a free text field for additional reasons. The five given reasons were: “Text too long”, “Difficult to understand”, “Not interested”, “Trust in legal regulations”, “Have to accept them anyway”. To avoid inducing biases between the five predefined options, they were shown in random order. The distribution of the responses is shown in Figure 4.4.

The most frequent response was that the text length stops participants from reading T&C (83%). This is a strong indication that our goal of summarizing T&C could indeed be of value for consumers. However, the results also show that summarizing alone will probably not be enough for many consumers. The summarized text should not just be shortened but also be simplified in comparison to the original text because 46% of participants refrain from reading T&C because they are too difficult to understand. The least often chosen response (22%) was “not interested”. Although it was chosen least often, still, more than 20% of consumers simply are not interested in T&C and technology is unlikely to change that. 38% trust that legal regulations will protect them, and a staggering number of 77% said that they do not read T&C because they have to accept them anyway in the end. Eight participants gave additional reasons. Three said that they simply do not have the time, two said that they trust in the vendors to have fair T&C and one participant said that he does not think that T&C are in general difficult to understand, but they are “purposefully written in a way that is exhausting to read”. These results underline that focusing on consumers alone will not be sufficient to address the existing imbalance of power and we instead also need to focus on consumer advocates and how we can support their work.

When we asked participants directly whether they would keep away from buying in a shop if

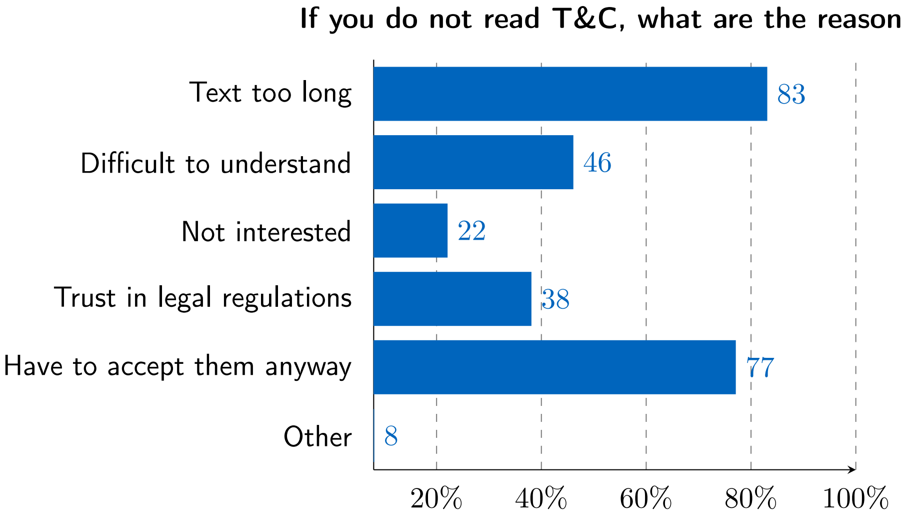


Figure 4.4.: Responses in percent for the question "If you do not read T&C, what are the reasons?"

void clauses would be pointed out to them, a majority of 81% answered yes, 16% answered that it would depend on the severity of the clause, and only 3% said they would order anyway, which is encouraging, but also inconsistent with the 77% who responded that they have to accept T&C anyway.

We also asked participants whether they ever had problems after they bought something online (e.g., not delivered goods or a refusal to take back goods). 52% reported that they already had such negative experiences. Participants who always read T&C were just as likely to have negative experiences as people who never read them. We also asked the participants to specify in a free text field what kind of problems they were facing. The categorized responses are shown in Figure 4.5.

The most common problem that 28% of the participants already faced was that ordered goods have not been delivered, followed by products that arrived broken (14%). The third most common problem were products that were inferior to the description given during the buying process. 6% faced problems receiving refunds for products they send back. Exceeded delivery times and deliveries to wrong addresses were faced by 4%. In what can only be called a very blunt case, one participant reported that his computer got infected with a virus, and when they complained, the shop owner referred to the T&C which included a clause stating that customers are aware of the fact that they will install a virus and agree with it.

While 69% of the participants that had problems were able to resolve them to their own satisfaction, 31% were not able to do so. We also asked participants if they had ever taken on legal advice from a lawyer or a consumer protection agency regarding an online transaction, and out of all participants, only two involved a lawyer; the 98 others have never taken on any legal advice for such a matter. This is a strong sign that the access barriers for legal advice in this area are still too high. At least 29 participants had problems that they could not resolve and never even

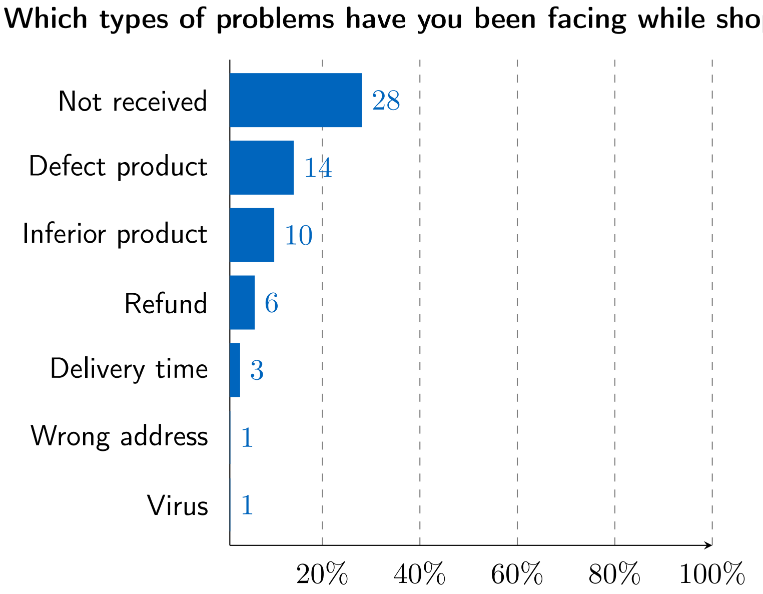


Figure 4.5.: Categorized responses in percent for the question “Which types of problems have you been facing while shopping online?”

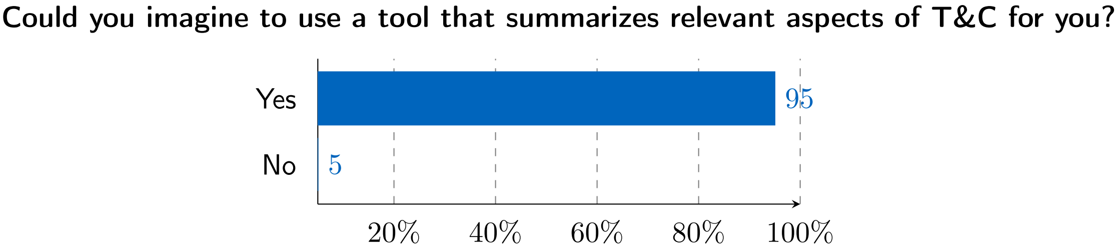


Figure 4.6.: Responses in percent for the question “Could you imagine to use a tool that summarises relevant aspects of T&C for you?”

tried to get legal advice. We believe that this shows that there is a need for additional support for consumers in such situations.

Since starting this project was, we were repeatedly asked by legal scholars and professionals why a consumer should care about void clauses at all; after all, they are void. However, we were never asked this question from consumers. This seems to be aligned with the fact that consumers hesitate to involve legal professionals and is a reminiscence to the German proverb that being right does not automatically lead to getting justice.

As the final question, we asked participants whether they could imagine using a tool that summarizes the relevant aspects of T&C for them before they order. An overwhelming majority of 95% answered with yes. However, a mere declaration of intent is easy to give and should not be overrated.

4.3. Relevance for Consumer Advocates

In 2019, the German consumer protection agencies “Verbraucherzentralen” received 119,611 complaints from consumers. While there is no data available about how much of these complaints were related to standard form contracts, we know that 4,378 of the complaints we about electricity contracts, 6,131 about insurance contracts, and 5,797 about bank saving plans, to name just three examples, which are most likely all based on standard form contracts. 4,770 complaints that were received were related to online shopping. (Verbraucherzentrale Bundesverband e.V., 2020b) These numbers show that standard form contracts are an important part of the counseling the Verbraucherzentralen do, which is not surprising given their economic importance. Hence, support in analyzing and assessing these types of contracts by technology could potentially support the work of the Verbraucherzentralen significantly.

The Verbraucherzentralen do not only act upon individual complaints, but they also monitor the online market actively, i.e., detect void clauses in T&C before they are reported by consumers (see Chapter 5 for more information). From 2015 to 2019, they received federal funding for this specific task under the “Marktwächter” (market guard) program. Since then, however, there has been no follow-up funding, and the organizations have, therefore, mostly stopped to monitor the markets actively or at least reduced their involvement considerably. Technological support could help to pick up this important task again.

Standard form contracts are also relevant for the Verbraucherzentralen from a financial perspective. Verbraucherzentralen charge consumers for individual legal advice. The amount that is charged varies between the organizations in the different states. The Verbraucherzentrale Bavaria charges, e.g., 14,50 Euro per 30 minutes, while the Verbraucherzentrale Baden-Wuerttemberg charges 21,45 Euro for 20 minutes. The Verbraucherzentrale Bayern earned in total 576.143,72 EUR from these fees in 2019.

Additionally, cease-and-desist orders are usually linked with a contractual penalty and a fee that is charged from the receiving company. We analyzed the latest available annual reports from all 16 state Verbraucherzentralen and the federal association to find out how much revenue these positions generate. Unfortunately, the level of detail in which revenues are reported vary widely between the different organizations, and it is not possible for each of them to identify the financial contributions from legal proceedings against companies. Table 4.1 shows for each of the organizations how much income they declared in the latest available annual report in the category to which contractual penalties and fees from cease-and-desist orders belong. Where these numbers are reported separately, their contribution to the income is usually in the mid-five-figures. E.g., 19,650 EUR in Berlin, 53,851 EUR in Hessen, or 153,841 EUR in Hamburg, most of which was generated through litigation. Most organizations do not report this income stream separately, but rather as part of their “self-funding”, in comparison to federal and state funding, which is the biggest income stream for all 16 Verbraucherzentralen and the federal association. In Berlin, the proceedings from cease-and-desist fees and contractual penalties contribute about 6.2% of this self-funding. If this share is anywhere near this figure for the bigger organizations, like North-Rhine Westphalia with reported self-funding of 3,433,600 EUR, this revenue stream generates hundreds of thousands of Euros per year for the Verbraucherzentralen.

Organisation	Year	Income Category	Value
Federal Association	2019	Reimbursement of legal costs, contractual penelties, and fees for cease-and-desist orders	225,446.62
Berlin	2018	Cease-and-desist orders and contractual penelties	10.596,48 + 9.053,65
Brandenburg	2019	Other monetary income	321,310.95
Saxony Anhalt	2018	Self-funding, membership fees, balance from 2017, and others	290,750.99
Saxony	2018	Fines	20,498.56
Thuringia	2019	Income from talks, counceling, donations, cease-and-desist orders etc.	79,066.29
Mecklenburg-West Pomerania	2018	Various incomes	48,739.32
Schleswig-Holstein	2018	Others	52,163.32
Lower Saxony	2018	Self-funding	578,512.77
Hamburg	2019	Class action suits, donations and fines	131,658.00 + 22,183.00
Bremen	2018	Revenue from counceling and other revenue	257,292.69
Hessen	2019	Reimbursement of legal costs, donations, and fines	28,838.54 + 25,012.72
Rhineland Palatinate	2019	Self-funding	149,616.81
Baden Wurttemberg	2019	Other income (income from contractual penelties, donations, and others)	383,679.91
Saarland	2016	Self-funding	90,569.94
Bavaria	2019	Other income	83,942.68
North Rhine Westphalia	2019	Self-funding	3,433,600.00

Table 4.1.: Revenue report from Verbraucherzentralen including contractual penalties and fees for cease-and-desist orders

4.4. Relevant Clauses

Standard form contracts can, especially if they include void clauses, regulate pretty much everything one can possibly think of. And even if we restrict ourselves to T&C from online shops, there are still hundreds if not thousands of subjects that can possibly be regulated. In order to be able to conduct a meaningful analysis and evaluation, we have to restrict ourselves to a subset of clauses. Two primary goals guided us in the selection of this subset. First, we want to cover as much of the most frequently appearing clauses as possible. Automation can unfold its full potential in situations that occur frequently. There is comparatively little value in being able to thoroughly analyze and legally assess an exotic clause that appears in only one T&C document, compared to a clause that appears in almost every contract. It makes more sense to focus the in-depth analysis on frequently occurring clauses, while it would already be helpful to be able to just point out "exotic" clauses and leave the analysis to human experts. Second, we want to focus on clauses that are relevant from a consumer and consumer advocate perspective. Severability clauses, for example, can be found in many T&C from German online shops, however, from a consumer protection perspective, they are mostly irrelevant because the issue is regulated in § 306 Abs. 1 BGB anyway and such clauses usually do not add any tangible information.

While these are the two main criteria for the selection of relevant clauses, there is a third point we took into consideration. From the perspective of social impact, it does make sense to prioritize clauses that are known to be void often or at least have been in the past. However, it is not our goal to foster as many cease-and-desist orders as possible, but to foster consumer protection (see Section 4.5 for a further elaboration of this consideration). Therefore this point is only relevant in conjunction and balance with the criterion of relevance from a consumer protection perspective.

To build a list of relevant clauses that satisfies all of these goals, we started with building a taxonomy to classify T&C clauses by their topic (Section 4.4.1). Based on this taxonomy, we analyzed the position of consumers and consumer advocates to find out which types of clauses are more important or relevant from their individual perspective (Section 4.4.2).

4.4.1. Taxonomy of T&C Clause Topics

We combined multiple approaches and sources to create our taxonomy for T&C clauses:

- Legal templates: We searched legal literature, material from industry associations and online T&C generators (see Section 3.5) for templates and analyzed which types of clauses are present in these templates, assuming that they are likely to appear in most of the T&C found online.
- Bottom-up approach: In addition, we annotated more than 5,000 clauses from T&C of online shops to find frequently occurring types of clauses that are not covered by the templates.

4.4.1.1. Literature

Personalized legal advice can be costly, and standard form contracts are a highly regulated class of documents. Many businesses, therefore, use templates for their T&C. Templates promise legally watertight texts, which can be easily adapted to specific needs at very low costs. Historically, such templates have been mainly provided by legal publishing houses and industry associations, but more recently also have been digitized and are now offered as online generators.

So-called "Formularbücher" (form books) are books that contain standardized legal templates that can be adapted or extended by the reader. In addition to the template itself, they often contain explanations and pointers to relevant judgments and commentaries. German publishing houses offer such books for a broad range of topics, from employment law (Klemm et al., 2014) to sports contracts (Partikel, 2015).

We analyzed two such publications that contain templates for T&C for online shops. The "Beck'sches Formularbuch IT-Recht" (Beck's form book for IT-Law) (Weitnauer and Mueller-Stöfen, 2017) and the "Vetrags- und Formularbuch" (contract and form book) (Fingerhut, 2009b). In the following, we will offer a taxonomy for classifying T&C clauses regarding their content, based on the templates contained in these two publications.

We started the analysis with the template from the "Beck'sches Formularbuch IT-Recht", which was written by Sommer and von Stumm (2017). While the original template is German, the book also contains an English translation. Together with two consumer advocates, we went through every section and analyzed their topics and created a class in our taxonomy for each new topic. One section in the template can contain multiple topics. Classes are ordered hierarchically; subclasses of "withdrawal" are, for example, denoted with "withdrawal:". Table 4.2 shows the result of the classification process. In addition to the 13 numbered section, the chapter also contains a separate section for the "information on the right of withdrawal" (Sommer and von Stumm, 2017, p. 723), which is listed in Table 4.2 under "WITH".

Section	Topic	Class
Sec. 1.1	applicability	applicability
Sec. 1.1	contracting party	party
Sec. 1.2	contract language	language
Sec. 2	applicable law	applicableLaw
Sec. 3.1	conclusion of contract - binding of offers	conlusionOfContract:binding
Sec. 3.2	conclusion of contract - submission of order	conclusionOfContract:steps
Sec. 3.3	conclusion of contract - necessary steps to order	conclusionOfContract:steps
Sec. 4	adjustments to orders	conclusionOfContract:changeOfOrder
Sec. 5	storage of the contract text	textStorage
Sec. 6.1	personal data - reason for storage of personal data	personalData:reason

Table 4.2.: Taxonomy for the classification of the topic of clauses from T&C, derived from Sommer and von Stumm (2017)

Section	Topic	Class
Sec. 6.2	personal data - information needed to order	personalData:information
Sec. 6.3	personal data - information needed to create an account	personalData:information
Sec. 6.4	personal data - usage of the stored information	personalData:usage
Sec. 6.5	personal data - duration of storage	personalData:duration
Sec. 6.6	personal data - update of personal data	personalData:update
Sec. 7	payment - methods	payment:methods
Sec. 8	retention of title	retentionOfTitle
Sec. 9	delivery conditions	delivery
Sec. 10	right of withdrawal	withdrawal
Sec. 11.1	warranty - options (e.g., withdraw or reduce price)	warranty:options
Sec. 11.2	warranty - period of limitation	warranty:period
Sec. 11.3	warranty - rights in case of defect	warranty:options
Sec. 12.1	limitation of liability - cases in which the company is liable	liability
Sec. 12.1	limitation of liability - cases in which the company is not liable	liability
Sec. 12.2	limitation of liability - liability for shop system	liability
Sec. 13.1	final remarks - additions and changes	changes
Sec. 13.2	final remarks - place of jurisdiction	placeOfJurisdiction
Sec. 13.3	final remarks - arbitration	arbitration
Sec. 13.4	final remarks - severability clause	severability
WITH	withdrawal - period	withdrawal:period
WITH	withdrawal - form of withdrawal	withdrawal:form
WITH	withdrawal - effects of withdrawal	withdrawal:effects
WITH	withdrawal - model withdrawal form	withdrawal:model

Table 4.2.: Taxonomy for the classification of the topic of clauses from T&C, derived from Sommer and von Stumm (2017)

After we analyzed the template from Sommer and von Stumm (2017), we performed the same analysis on the template from Fingerhut (2009a). The result is shown in Table 4.3. The template from Fingerhut (2009a) is significantly shorter, in both text length and the number of covered clauses. Nevertheless, it introduces new classes, which are printed italic in Table 4.3. Another interesting point is that some of the clauses we already found in Sommer and von Stumm (2017) are also contained in Fingerhut (2009a), however, organized in a completely different structure. While Sommer and von Stumm (2017), e.g., have a dedicated clause for limitations of liability, Fingerhut (2009a) puts the liability clause in the warranty section, which makes it potentially more difficult to find them. In total, based on both templates, our taxonomy so far consists of 43 different clause topics.

Section	**Topic**	**Class**
I	applicability	applicability
II	product description	*description*
III.1	conclusion of contract - submission of order	conclusionOfContract:submission
III.2	conclusion of contract - withdrawal of the company from the contract	*conclusionOfContract:withdrawal*
IV.1	delivery - partial delivery	*delivery:partial*
IV.2	delivery - time	*delivery:time*
V.1	delivery - costs	*delivery:costs*
V.2	delivery - customs	*delivery:customs*
VI.1	payment - methods	payment:methods
VI.2	payment - restraint	*payment:restraint*
VII.1	withdrawal - period	withdrawal:period
VII.1	withdrawal - form of withdrawal	withdrawal:form
VII.2	withdrawal - shipping method	*withdrawal:shippingMethod*
VII.3	withdrawal - shipping costs for returning goods	*withdrawal:shippingCosts*
VII.4	withdrawal - shipping method	*withdrawal:shippingMethod*
VII.5	withdrawal - compensation	*withdrawal:compensation*
VII.6	withdrawal - exclusion of products	*withdrawal:exclusion*
VIII	retention of title	retentionOfTitle
IX.1	recycling - battery	*disposal*
IX.2	recycling - battery	*disposal*
X.1	warranty - broken packaging	*delivery:brokenPackaging*
X.2	warranty - options (e.g., withdraw or reduce price)	warranty:options
X.3	warranty - cases in which the company is liable	liability
X.3	warranty - cases in which the company is not liable	liability
XI	liability for links	liablity
XII	intellectual property	*intellectualProperty*
XIII	personal data - reason for processing of personal data	personalData:reason
XIII	personal data - usage of stored data	personalData:usage
XIV.1	applicable law	applicableLaw
XIV.2	place of jurisdiction	placeOfJurisdiction
XV	severability clause	severability

Table 4.3.: Taxonomy for the classification of the topic of clauses from T&C, derived from Fingerhut (2009a)

4.4.1.2. Industry associations

Another source for T&C templates we identified are industry associations, who often provide such documents for their members. In Germany, trade companies have to be a member of one of 79 existing Chamber of Commerce and Industry (Industrie- und Handelskammer (IHK)). While some local branches actively discourage the use of T&C templates (IHK Frankfurt am Main, 2020), others provide such templates for their members. We analyzed exemplarily the template provided by the IHK Munich and Upper Bavaria (2020). The template is based on a template from a book written by Föhlisch and Groß (2018), which was published by the federal union of all chambers of commerce and industry in Germany. The result of the analysis is shown in Table 4.4.

We found that some of the clauses were very similar to clauses from the template by Sommer and von Stumm (2017) and are probably based on those. However, there were also three new clauses that were not present in the other two templates.

Section	Topic	Class
1	applicability	applicability
2	contracting party	party
3.1	conclusion of contract - binding of offers	conclusionOfContract:binding
3.2	conclusion of contract - submission of order	conclusionOfContract:submission
4.1	withdrawal	withdrawal
4.2	withdrawal - shipping costs for returning goods	withdrawal:shippingCosts
4.3	withdrawal	withdrawal
WITH	withdrawal - period	withdrawal:period
WITH	withdrawal - form of withdrawal	withdrawal:form
WITH	withdrawal - effects of withdrawal	withdrawal:effects
WITH	withdrawal - shipping costs for returning goods	withdrawal:shippingCosts
WITH	withdrawal - compensation	withdrawal:compensation
5.1	prices - VAT	*prices:vat*
5.2	delivery - costs	delivery:costs
5.3	delivery - costs	delivery:costs
6.1	delivery - destinations	*delivery:destination*
6.1	delivery - delivery company	*delivery:methods*
6.2	delivery - time	delivery:time
7.1	payment - methods	payment:methods
7.2	payment - methods	payment:methods
8	retention of title	retentionOfTitle
9	arbitration	arbitration

Table 4.4.: Taxonomy for the classification of the topic of clauses from T&C, derived from IHK Munich and Upper Bavaria (2020)

Another industry association that provides a template for T&C to their members is the "Bundesverband E-Commerce und Versandhandel Deutschland" (Federal Association of E-Commerce and Mail Order Germany). The template they provide was drafted by Schirmbacher (2018). The analysis is shown in Table 4.5. It only adds two new classes to our taxonomy, one for fees for different payment methods and one for regulations regarding late payments.

Section	Topic	Class
§ 1.1	applicability	applicability
§ 1.1	contracting party	party
§ 1.2	minimum age	*age*
§ 1.3	applicability	applicability
§ 1.4	contract language	language
§ 1.5	storage of the contract text	textStorage
§ 2.1	conclusion of contract - binding of offers	conclusionOfContract:binding
§ 2.2	conclusion of contract - submission of order	conclusionOfContract:steps
§ 2.3	conclusion of contract - binding of offers	conclusionOfContract:binding
§ 2.4	conclusion of contract - binding of offers	conclusionOfContract:binding
§ 3	price - vat	prices:vat
§ 3	price - delivery costs	delivery:costs
§ 4.1	payment - methods	payment:methods
§ 4.2	payment - methods	payment:methods
§ 4.3	payment - methods	payment:methods
§ 4.4	payment - fees	*payment:fee*
§ 4.5	payment - methods	payment:methods
§ 4.6	payment - methods	payment:methods
§ 4.7	payment - fees	*payment:fee*
§ 4.8	payment - late payment	*payment:late*
§ 5.1	payment - restraint	payment:restraint
§ 5.2	payment - restraint	payment:restraint

Table 4.5.: Taxonomy for the classification of the topic of clauses from T&C, derived from Schirmbacher (2018)

4.4.1.3. T&C Generators

We also analyzed T&C generated by T&C generators, like the one described in Section 3.5. We only used to text generated by Trusted Shops, which can be found in Appendix D.1, because, as previously mentioned, the text generated by Termly is not focused on online shops.

Section	Topic	Class
1	applicability	applicability
2	contracting party	party
2	conclusion of contract - binding of offers	conclusionOfContract:binding

Table 4.6.: Taxonomy for the classification of the topic of clauses from T&C, derived from Appendix D.1

Section	Topic	Class
2	conclusion of contract - submission of order	conclusionOfContract:setps
3	contract language	language
3	storage of the contract text	textStorage
4	delivery - costs	delivery:costs
4	delivery - method	delivery:method
5	payment - methods	payment:methods
6	retention of title	retentionOfTitle
7	delivery - broken packaging	delivery:brokenPackaging
8	warranty - period	warranty:period
8	liability - inclusion	liability
8	liability - exclusion	liability
9	liablity - cases in which the company is liable	liability
9	liablity - cases in which the company is not liable	liability:exclusion

Table 4.6.: Taxonomy for the classification of the topic of clauses from T&C, derived from Appendix D.1

4.4.1.4. Bottom-up identification

By the combination of the above described three sources, the taxonomy we derived consisted of 22 classes with 36 sub-classes. In order to get an estimate of how much of "real" T&C from online shops we can cover with this taxonomy, we manually annotated the T&C from 151 online shops (see Section 6.1.1 for more information on how these T&C were gathered and annotated), which left us with a list of 5,020 annotated clauses. Of these clauses, the taxonomy covered 4,684 clauses or 90.08%. The remaining 336 clauses fell into only two classes: 285 clauses contained information regarding vouchers and gift cards, and 51 clauses contained information about codes of conduct. We added both classes to the taxonomy as payment:vouchers and codeOfConduct. The final taxonomy, therefore, consists of 23 classes with 37 sub-classes. The complete taxonomy can be found in Appendix E.

Due to the broad range of sources we used for the creation of the taxonomy, one could even argue that clauses that are not covered by the taxonomy are likely to fall under §305c BGB which bans the use of clauses that are so unusual that the contracting party could not have anticipated them. Although this is certainly not applicable for all clauses which are not covered by the taxonomy, it could be used as a first indication to highlight such clauses as potentially important for the legal analysis.

4.4.1.5. Transferability

Naturally, the taxonomy is one part of this thesis which does not generalize to different kind of standard form contracts but is mostly limited to the concrete domain of T&C from online shops. Topics like delivery and disposal are simply not relevant for many other forms of standard form contracts, e.g., in banking or property rental. Nevertheless, some types of clauses can be found in other types of standard form contracts. Especially “technical” clause types, which do not directly touch the subject matter of a contract, but rather govern the contract itself, like the applicable law, the place of jurisdiction, or the language of the contract.

We exemplarily tried to classify the T&C of Commerzbank, one of Germany’s largest banks, with the taxonomy and found that only 17 out of 71 clauses (about 24%) could be assigned to one of the topics from the taxonomy, including the applicability of the contract, liability, changes of contract, applicable law, and place of jurisdiction. For a boilerplate rental agreement from the German “Mieterschutzbund” (association for tenants protection), we were even only able to classify one clause (contracting party) with the taxonomy.

4.4.2. Importance of Clause Topics

Based on the taxonomy, we wanted to find out which of these clauses are especially important from a consumer and consumer advocate perspective. We used two different approaches to answer this question:

- Consumer survey: We used the result from the previous two steps to ask consumers what aspects of T&C they care most about.
- Experience and data from consumer advocates: Additionally, we asked consumer advocates from the Verbraucherzentralen to provide data and insights into which types of clauses are relevant for consumers and consumer protection and which types of clauses they have seen to be void frequently in the past.

4.4.2.1. Consumer survey

Based on the clause topics we identified from the legal templates and the bottom-up approach, we asked consumers which clauses are most relevant for them. The question was part of the survey described in Section 4.2. In order not to overwhelm the participants, we only used the identified super-categories and omitted some rather technical clause types, like severability clauses. We ended up with the following list of clause topics:

- Withdrawal
- Payment
- Data protection
- Delivery

- Warranty
- Liability
- Place of jurisdiction
- Retention of title
- Contracting party
- Applicable law
- Disposal
- Conclusion of contract
- Copyright

Participants were asked to rank these topics according to how relevant they are for themselves, resulting in an ordered list from 1 (most important) to 13 (least important). The order in which the options were shown to the participants was randomized in order to avoid inducing any biases. Figure 4.7 shows a graphical representation of the rankings given by the participants. The topic that was ranked as most important by most participants is withdrawal, which was ranked as the first priority by 25 participants, followed by data protection (20), and payment and warranty (15 each). The least important topics for participants are the place of jurisdiction (26), disposal (22), and copyright (20).

Overall the participants ranked withdrawal as the most important clause topic for them (mean 2.91, median 2, see Table 4.7), followed by warranty and delivery. The three least important topics are copyright, place of jurisdiction, and disposal. Although data protection is ranked relatively high (sixth place), it will not be one of our priorities. First, data protection regulations are regularly not part of the T&C, but rather a separate document or page. Second, there is already a broad range of existing work that deals specifically and only with privacy policies (see Section 3.1.3).

4.4.2.2. Consumer advocate perspective

According to a report from Föhlisch (2019), 45% of all cease-and-desist orders in German e-commerce are send by competitors. The different consumer protection organizations account for only about 15% of all cease-and-desist orders. 15% of all cease-and-desist orders that were sent in 2019 were targeted at void or missing clauses about withdrawal rights according to Föhlisch. 10% were targeted at missing or void clauses about arbitration, 3% about data protection, 2% at missing clauses about warranty rights, and 1% at missing clauses about the storage of the contract text.

For consumer advocates, such clauses, with the exception of withdrawal rights, are rarely interesting. In interviews, consumer advocates told us that most cases in which consumers seek their counseling are related to clauses that are more or less directly connected to monetary obligations. Examples of such clauses include fees for different payment methods or default charges, but also withdrawal clauses, especially in the context of contracts with long-term obligations, like mobile

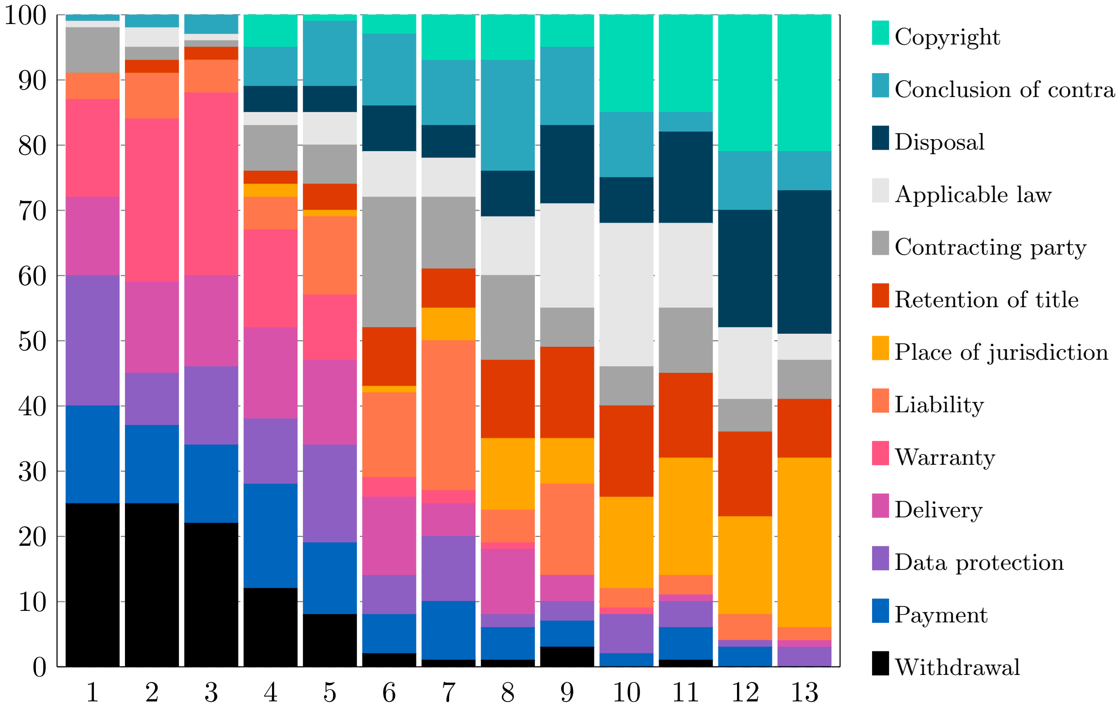


Figure 4.7.: Consumer ranking of personal relevance of different clause topics from most important (1) to least important (13)

Clause topic	Mean	Median	SD
Withdrawal	2.91	2	2.00
Warranty	3.04	3	1.59
Delivery	4.47	4	2.56
Payment	4.80	4	3.07
Data protection	4.91	4.5	3.33
Liability	6.53	7	2.83
Contracting party	7.35	7	3.19
Conclusion of contract	7.87	8	2.86
Applicable law	8.90	9.5	2.68
Retention of title	9.10	9	2.72
Copyright	10.30	11	2.51
Place of jurisdiction	10.62	11	2.22
Disposal	10.01	11	2.71

Table 4.7.: Consumer ranking of personal relevance of different clause topics from most important (1) to least important (13), ordered by mean relevance

phone contracts. Given that consumers are charged for the counseling, it is not unexpected that they would mostly seek counseling in cases where there is money at stake.

While clauses concerned with data protection and privacy play an important role for consumers and are also often the target of cease-and-desist orders, they are, relatively speaking, less important for the consumer advocates we interviewed. This is mostly because each German state has a designated state commissioner for data protection ("Landesbeauftragte für den Datenschutz") with which consumers can file complaints (free of charge) in cases of data protection and privacy violations.

4.5. Ethical Considerations

As mentioned before, our goal is to support consumers and consumer advocates in order to further consumer protection and address the imbalance of power between corporations and consumers. While these are, by most standards, worthy and ethical goals, just because something is well-intended does not mean it can not have critical or at least ambivalent consequences. In this section, we want to highlight some of the issues that can arise from the research presented in this thesis and the goals it pursues.

The laws governing T&C are changing comparably fast. § 309 BGB, to give just one example, was changed three times in 2016 and six times in total since 2008 (Liebig, 2020). For small companies, without in-house legal counseling, it can therefore be expensive and challenging to keep up with the changing legislation and keep T&C always up to date. In such cases, honest mistakes might be made in drafting and maintaining T&C which do not intend to harm consumers. Nevertheless, such mistakes can make companies vulnerable to cease-and-desist orders from competitors and organizations which specialize in sending out cease-and-desist orders, not in order to protect consumer interests but for personal financial benefit. Although the German legislator took action in 2013 to make such models less attractive and lowering the costs for cease-and-desist orders in such cases (Gesetz gegen unseriöse Geschäftspraktiken - Law against dubious business practices), they still pose an economic threat to smaller companies.

We respond to this problem in two ways. First, as set out in Section 4.4, we want to focus on clauses that are relevant for consumers and consumer protection. Competitors and so-called "Abmahnvereine" (organizations which specialize in sending cease-and-desist orders for financial benefits) often focus on rather technical aspects of T&C, changes to which are usually also less well known. Second, with our collaboration partners being Vebraucherzentralen, we choose organizations that are dedicated to consumer protection and bound to that aim by their statue and their state given mission. However, given that we publish a lot of our results under open licenses, it can not be prevented that our research can also be used by less well-intended actors. While this poses a potential threat, it allows companies on the other side to use our results in the same way on their own T&C and hence make sure they match the rule of law.

A second, arguably more philosophical issue that arises, not just from our research, but from the perspective of consumer-focused LegalTech in general, is whether our legal system is prepared for lowering the bar for accessing the system. The legal and moral standpoint on this issue is quite clear. The charter of fundamental rights of the EU guarantees in article 47 that "everyone whose rights and freedoms guaranteed by the law of the Union are violated has the right to an effective remedy before a tribunal". While the legal situation is clear, it is also clear that there are, in fact, barriers in place which make access to justice harder, whether they are of financial or procedural nature. And while it could be denied that they purposefully do so, it is difficult to deny that these barriers help to keep up the in many countries already stretched legal systems. If we would be able to denounce our neighbors by the click of a button every time they disturb the nighttime, this could not just have implications for the viability of our legal systems but also for the kind of society we live in and how we interact with each other. With apps like "Wegeheld"[3] that allows user to report parking violations through an app and so-called "online police stations" established in all German states, where reports can be filed online (Bundeskriminalamt, 2020), we might be about to find out what the consequences will be.

Concerning our work, we would argue that, if it has any influence on the legal system at all, it is designed to reduce its load. The behavior we would expect to see from consumers when they find out about a void clause is that they do not enter the contract in the first place and hence prevent a possible following lawsuit. And while the number of cease-and-desist orders send out by consumer advocates might rise, we would hope that subsequently, this would lead to fewer cases brought on by consumers about void standard form contracts.

An issue that is faced by every computer system that makes or supports decision is that, eventually, just like humans, the system will make mistakes. If users trust a system too much and follow its decisions or suggestions without questioning and checking them, they might act upon wrong outputs from the system with potentially negative consequences for them or others. In the domain we operate, if a system fails to identify a void clause, which is disadvantageous for the user, this might cause problems for the user and, in the worst case, might lead to a loss of money. On the other hand, if the system falsely claims a clause is void, this could have a negative impact on the company whose contracts are falsely claimed to be void because they might lose customers. In addition to ethical questions, liability questions arise in both cases. Not to mention legal questions arising from the German "Rechtsdienstleistungsgesetz" (Legal Services Act) under which it is questionable whether a consumer software would even be allowed to deliver such judgments.

For the first issue, we would argue that since by all we know, consumers hardly ever read T&C, even a system that only catches 50% of void clauses would still be a massive improvement compared to the status quo. One way of addressing both issues is user communication. The system should openly communicate that it is not perfect and that its output should be seen as a first estimate rather than a final judgment. Users should be encouraged to examine the clauses that were highlighted by the system further. However, given that one of the very reasons why we conducted this research is the fact that people do not read T&C (see Section 4.2), it would be more than naive to think this would be a definitive solution.

This, in combination with the liability questions and the results of the consumer survey, are the main reasons why we decided to put our focus on supporting consumer advocates rather than supporting consumers directly. As domain experts, we believe they will be less prone to blindly trusting a system and ultimately are more likely to improve the protection of consumers.

[3] https://www.wegeheld.org/

If a system that automatically analysis standard form contracts for their validity would be successful and widely adopted, one of the implications would very likely be that companies could start trying to "gamble" the system. This is a phenomenon that can be observed very well in the area of search engine optimization (Malaga, 2008) and security (Mansfield-Devine, 2018). This could potentially lead to a situation where such a system would mostly fail to detect clauses that were purposefully drafted in a consumer-aversive way and would potentially be left detecting mostly clauses that are unintentionally void, e.g., by honest mistake, and were never intended to harm consumers. If we can learn anything from search engine optimization and security, then that there is no easy or permanent fix to such problems. We, therefore, try to build our system in a way that it can be easily adapted, so that consumer advocates can change the system in a way that it will be able to detect such clauses, once they became aware of it, entering an "arms race" with malicious companies. And while "security through obscurity" is generally discouraged, search engine providers have shown that obfuscating the exact criteria helps to stay ahead of attempts to manipulate the ranking of websites. Therefore, our decision to focus on consumer advocates as users, rather than consumers themselves, can also help to mitigate the problem since companies will not know the exact rules used for the analysis and will not be able to directly test different versions of their clauses.

CHAPTER 5

Requirements Identification

In this chapter, we describe how we approached the requirement elicitation and which requirements we identified through that process. This is an essential part of the action research approach we are taking because it addresses two of the four key elements defined by O'Leary (2017) (see also Section 1.3). It helps us to ensure that we *address real-world problems* and is an important element of *participation*, in which the voice of stakeholders is being heard. The process will also help us in answering our third and fourth research question. Based on the premises described in the previous chapter, we focused on consumer advocates as stakeholders. We collaborated with seven experts from two state consumer protection agencies and one expert from the federal association of German consumer protection agencies over a period of three years, from 2017 to 2020.

Within the consumer protection agencies, we identified four main relevant stakeholder groups:

- *consumer counselors* who directly interact with consumers and provide counseling,
- *in-house legal counselors* who are responsible for taking legal actions like sending cease-and-desist orders or filing lawsuits,
- *board members* who make budget and political decisions, like settings focuses on specific legal aspects or industries, and
- *IT experts* who are responsible for administrating tools that support the other stakeholders in their work.

While we did implement a prototypical user application to evaluate the techniques we developed (see Chapter 9), the focus of this thesis is on the core algorithms for the semantic analysis, legal assessment, and summarization of standard form contracts, and not on the user application and its UI. Therefore, the requirement elicitation process was also focused on these core algorithms, and other aspects, like deployment and integration, were not considered. The stakeholder group of IT experts, which was presented among our experts by one person from the federal association, is mainly interested in operational aspects, and hence implementations and applications, and less interest in the algorithmic aspects we are focusing on.

From the remaining seven experts, one belongs to the group of board members. Six have experience as consumer counselors, and five of them have also been actively working as consumer counselors during our research. Four experts have experience as in-house legal counselors, and three of them have also been actively working in this role during our research. This already shows that, while the two stakeholder groups have distinct responsibilities, a personal overlap exists. Out of the seven experts, six are fully qualified lawyers under German law, i.e., they passed the second part of the German bar exam and are allowed to practice law in any court of the country. All future references to collaboration partners from the consumer protection agencies will refer to these seven experts.

The consumer protection agencies in Germany are organized as independent non-profit associations ("gemeinnützige Vereine" in the German law). As such, it is mandatory for them to have an elected board, which is another group of inter-organizational stakeholders. However, they do not play an active role in the daily business and have, at most, a control function. Therefore, they do not play a role in our analysis.

In addition to the inter-organizational stakeholders, there are also external stakeholders:

- *consumers* who seek counseling and are the customers of the consumer protection agencies,
- *companies* that, most of the time, prefer not the be involved with consumer advocates, and
- the *state* provides most of the funding and sometimes gives directions and sets priorities through programs and campaigns.

As mentioned before, we will focus on the consumer protection agencies and their internal stakeholder groups here, but also analyze the requirements of consumers. We did not gather requirements from companies.

The state, as the main source of funding for the consumer protection agencies, has an interest in maximizing their efficiency and hence the value of its own funding. The fact that the Federal Ministry of Justice and Consumer Protection as representative of the stakeholder group "state" funded large parts of the research presented in this thesis shows that there is an interest and willingness to invest in the research and development of AI technology.

5.1. Methodology

There are different methodologies and approaches to identify requirements. Typically those include interviews, surveys, and gathering data from existing systems or documentation (Tiwari et al., 2012; Dick et al., 2017b). Our analysis is mainly based on interviews, ranging from one-to-one interviews to group interviews with up to five experts from the consumer protection agencies and up to four researchers. The interviews were conducted from 2017 to 2020, first in larger intervals of one to two months, and later (starting in 2020) in biweekly intervals. In total, more than 30 interviews have been conducted. While there were no existing systems in place to process contracts, we did find out that two different databases are used to organize processes, one stored on a SharePoint server and one stored in an Excel file. We analyzed how they are used (see Section 5.2 for more details) in our first step, in which we analyzed existing processes related to standard form contracts and formalized them (see Section 5.2).

We used the process formalization to analyze the stakeholder requirements (see Section 5.3). The identification of the stakeholder requirements of the consumers is based on the survey presented in Section 4.2 and one-on-one interviews with ten students in their role as consumers. The students are all from computer science related fields and were recruited as volunteers through university courses. Based on the stakeholder requirements from both groups, we derived the system requirements (see Section 5.4).

5.2. Process Analysis

As part of our interview series, we conducted multiple interviews dedicated to analyze the existing processes at the consumer protection agencies. In these interviews, we asked the experts from the consumer protection agencies in which situations and how they currently work with standard form contracts. We consolidated the responses from interview notes and formalized them in a process model using Business Process Model and Notation (BPMN) (Chinosi and Trombetta, 2012). In a follow-up interview, we presented the resulting business process models to the experts and asked them whether they capture the underlying processes correctly. We adapted our models, where necessary, according to the feedback. We found two main processes in which the experts work with standard form contracts, which we will introduce in the following two sections.

5.2.1. Consumer Counseling

Currently, the most important process for the processing of standard form contracts is consumer counseling. The formalization of the process is shown in Figure 5.1. The process is initiated by an incoming complaint or questions from a consumer about a standard form contract. This can happen through different channels. Traditionally, most of the counseling is done in-person; however, an increasing amount is also done through telephonic consultancy. More recently, some consumer protection agencies have also started to offer counseling through email. A typical inquiry would be, e.g., a consumer that complains about high fees for overdue fines or a subscription that can not be canceled. As mentioned in Section 4.3, consumers usually have to pay between 10 EUR and 30 EUR for 20 - 30 minutes of counseling.

The process, independent of the channel, is always very similar. With the consumer's inquiry, the counselor receives the contract document, tries to find the clause or clauses relevant to the inquiry, and analyses them to determine whether a complaint is justified or not, or to gather the information necessary to answer questions. If a relevant clause is found to be void by the counselor, they pass the clause on to the in-house legal team for additional review. Depending

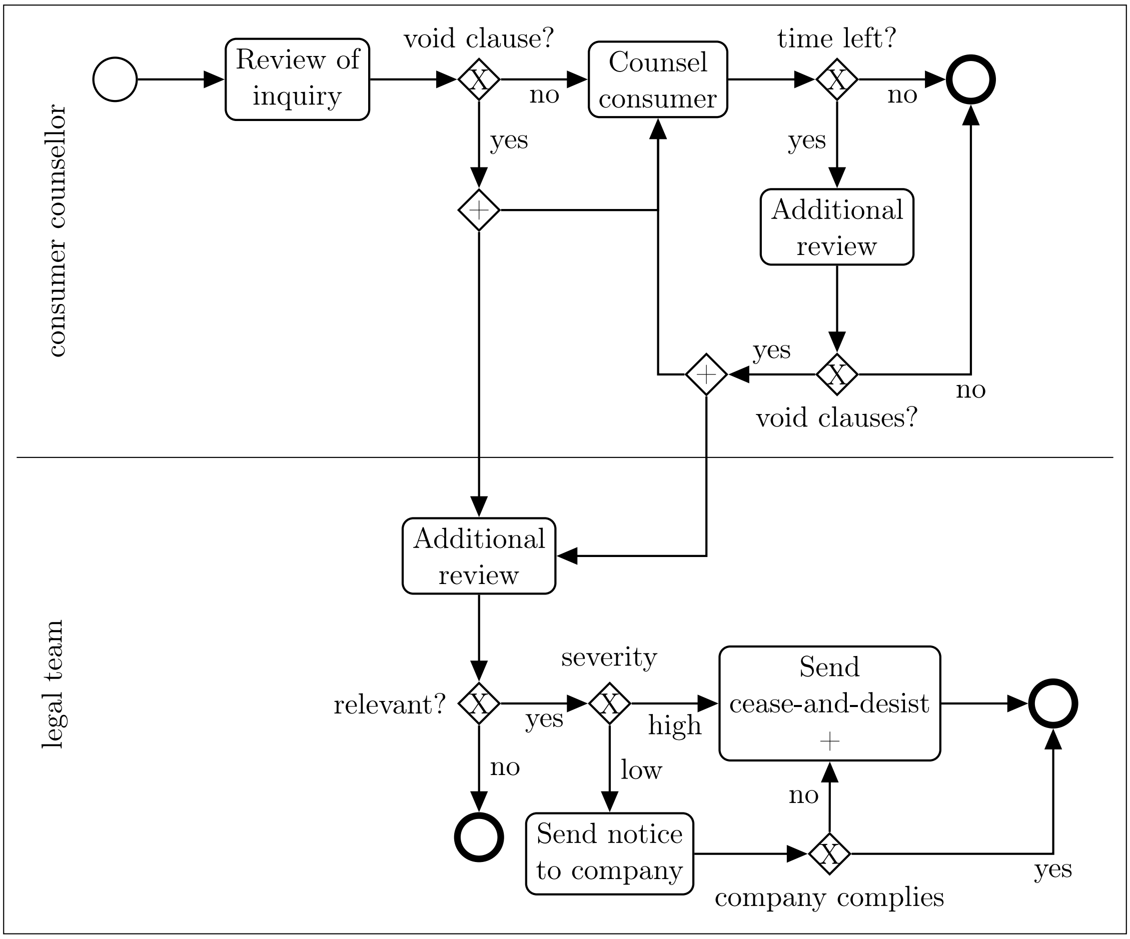


Figure 5.1.: Worfklow for the handling of consumer inquiries

on the channel and the case, the format in which the contract is received can differ from a printed contract to an electronic document, link, or fax.

Based on the analysis of the relevant parts of the contract, the counselor provides the necessary information to the consumer and discusses the case. If the counselor has additional time available, they also review related parts of the contract and pass other possibly void clauses on to the legal team as well. If no clause is found to be void, the process ends. During the interviews, we found that counselors rarely have the time to review more clauses of the documents, and even if they have the time, they usually stick to clauses closely related to the inquiry and do not typically check the whole contract.

Each clause that is handed over to the in-house legal team will be reviewed again. Generally speaking, the consumer counseling and the legal team are separate, however, especially in smaller consumer protection agencies, there might be an overlap of staff between the two teams. Based on the relevance, the process either ends or continues. While there is no fixed catalog of relevance criteria, we were able to gather a (non-exhaustive) list of criteria during our interviews. First and foremost, it is assessed whether the clause is indeed void. It is also assessed whether the clause has a potential negative influence on consumers and how high the chances are to successfully challenge the clause in court. Additionally, it is also factored in whether there is a public interest in a certain aspect or not. If a clause is deemed to be relevant, there are two possible actions. Either a cease-and-desist order is sent to the company, or the company is notified informally about the void clause. The decision on which action to take is based on the severity of the violation, but also on the company that uses the clause. If the company is large or has a history of consumer protection violations, a cease-and-desist order will be sent. However, if the company is small, and the clause might be an honest mistake, the company might just be notified. Sending a cease-and-desist order is a sub-task which is shown in Figure 5.2.

In case a notice is sent to a company, the company can either comply with the notice, i.e., change or remove the void clause, or not comply. In case the company complies, the process is finished. In case the company does not comply, a cease-and-desist order will be sent, leading to the same sub-task, which is shown in Figure 5.2

Before a cease-and-desist order can be sent, the legal team first has to check a central database called “JURA”, which is operated by the federal association of consumer protection agencies and accessible for all state-level organizations. This database contains all present and past cease-and-desist orders sent by any of the state organizations. One of the purposes of the database is the documentation and sharing of knowledge. Another, at least as important purpose, is to avoid that the same company receives cease-and-desist orders from multiple state-organizations. Historically, each separate state-organization had a clearly separated sphere of influence, based on the separation by states. In times of e-commerce, however, this is not applicable anymore, which sometimes leads to an overlap and disputes about competencies.

The database can be searched by industries, companies, and keywords. There is no common taxonomy in place to specify topics. The level of detail in which the cases are documented varies greatly between organizations and individual users. If the clause is already in the database because a cease-and-desist order was already sent or another organization is currently working on it, the process ends. If the clause is not yet in the database, it has to be “reserved” before a cease-and-desist order can be sent. This is done by creating an entry in the database. Other organizations have then one week time to react. Only after this time has passed, a cease-and-desist order can be sent.

There are three possible outcomes when a cease-and-desist order is sent to a company: the company can either sign the order, not sign the order or sign a changed version of the order. If the company signs the order, this is documented in the “JURA” database, and then both the subtask and the main process end. The fact that the processes end here is one of the problems our process analysis revealed. When signing a cease-and-desist order, a company does not only agree to not use a certain clause in their contracts anymore, they also agree to a contractual penalty, which they have to pay if they use the clause again. In order to ensure a long-term improvement for consumers and claim such penalties, the consumer protection agencies would have to check periodically whether the agreed on changes are implemented by the company and stay in place. However, this is currently not happening due to a lack of resources. Therefore, contractual penalties are usually only claimed if a new consumer inquiry is received or if the company is checked again in a different context, but there are no processes in place to regularly check compliance with cease-and-desist orders.

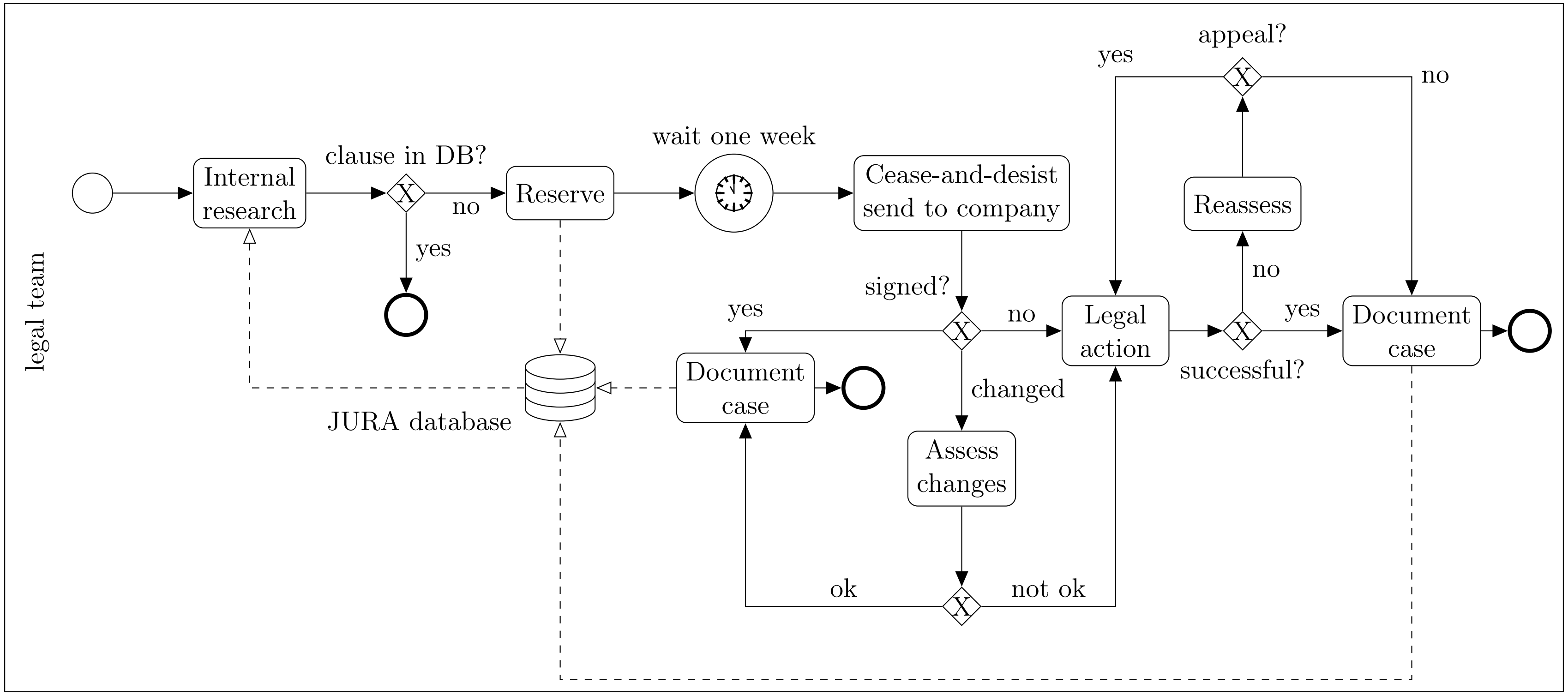


Figure 5.2.: Subtask of sending cease-and-desist orders

In case a company does not sign a cease-and-desist order, further legal actions are taken, which usually means filing a lawsuit. It was important for our interview partners to point out that they always take further legal actions if a cease-and-desist order is not signed by the company because they only send such orders for cases in which they are willing to sue and are confident enough that they can win.

If the legal actions are successful, the outcome is documented, and the process ends. If the legal actions are not successful, the case can either be dropped, or an appeal can be filed. This decision is based on a reassessment of the relevance of the case.

If the company signs a changed version of the order, the changes are assessed. If the changes are acceptable, the result is documented, and the process ends. If the changes are not acceptable, further legal actions are taken.

5.2.2. Campaigns

The above-described process is the most important one, according to our interviewees, when it comes to standard form contracts. One drawback that was identified by the interviewees is that most of the time, clauses are only challenged after consumers have already faced problems because of them. Another possible problem is the selection bias that is introduced. Because there is a fee for the counseling, only cases with a relevant monetary value will be brought to attention. Consumers might also be hesitant to bring up more "delicate" matters, e.g., connected to gambling or adult entertainment websites.

Until 2019, the federal government of Germany funded a program called "Marktwächter" (market watchman), which provided the consumer protection agencies with money to actively monitor markets and detect possible negative developments, like void clauses, before consumers bring them to their attention. Each state organization was responsible for a different market segment, like energy and electricity, pension schemes, or e-commerce. However, since the program ended, consumer protection agencies do not have the resources anymore to do that at scale. In our interviews, all experts, from the board member to the counselors, expressed that they see this as a great loss and would like to be able to do that again.

On a smaller scale, the consumer protection agencies still occasionally run similar actions, e.g., if they find a new "trend" in the usage of a specific clause or a new malpractice in a specific industry. Due to their similarity, we consolidated both in a process which we entitled "campaign" because it consists of a series of actions targeted towards a specific market area, industry, area of the law, or type of clause.

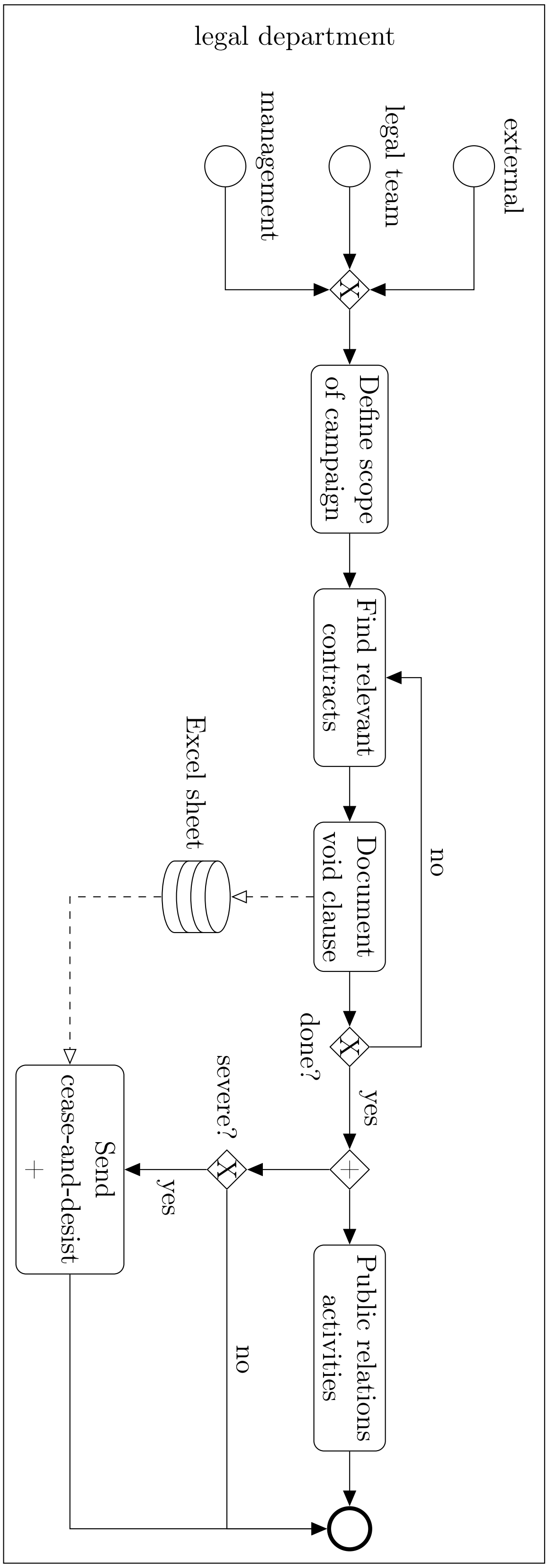


Figure 5.3.: Worfklow for “campaigns”

In the interviews, we identified three possible sources that can trigger such a campaign:

1. A campaign can be triggered by an external source, which is most of the time, either the federal or the state government, which can define priority areas or provide additional funding for specific areas.
2. If the legal team notices an increase of (justified) complaints in a certain area, this might trigger a broader analysis of this area.
3. Sometimes campaigns might be triggered by the management because they identified a certain area as specifically relevant with respect to public interest.

The first step in each campaign is the definition of the scope, i.e., by which criteria it will be decided whether a certain contract is relevant. In the past, such criteria included the industry (e.g., gyms) or a specific clause type (e.g., overdue fines). Once the scope is defined, the legal team starts to gather relevant contracts. Depending on the industry, they apply different techniques. In e-commerce, for example, T&C are openly available and can directly be gathered by the legal team. In other industries (e.g., insurance services), that is not necessarily the case. In such instances, the legal team either requests the contracts directly from the companies or asks consumers to share their existing contracts with them.

These documents are then searched for relevant void clauses, which are documented using excel sheets. As long as the campaign is still running, either because there is still time left or the desired amount of clauses was not yet reached, the process is repeated. Once this process is finished, the findings are usually used for public relations activities to publicly highlight the identified problems. Our interview partners emphasized that it is usually not the goal of such campaigns to send cease-and-desist orders or contact individual companies but to raise awareness for broader issues among consumers and political decision-makers. Only in very severe exceptional cases, the findings of a campaign might also lead to a cease-and-desist order being sent.

5.3. Stakeholder Requirements

Stakeholder requirements are part of the problem domain, i.e., “the domain in which a system is going to be used” (Dick et al., 2017b, p. 113). They should describe the goals that stakeholders want to achieve by using a system, without already referring to specific solutions (Dick et al., 2017a, p. 22). The main question that should be answered is what stakeholders want to be able to do with a system.

One way to identify stakeholder requirements is the usage of user scenarios. A user scenario is a high-level description of a time-series of events that has to happen in order to achieve a certain task or goal. By using user scenarios, the interviewed stakeholders are invited to reflect on how they want to fulfill a certain task, which provides a framework in which their needs can be explored. (Dick et al., 2017b) We presented experts (i.e., stakeholders from consumer protection agencies) and consumers with different scenarios and gathered their respective stakeholder requirements.

5.3.1. Experts

Based on the process analysis we performed, we derived three user scenarios for the group of expert stakeholders:

1. consumer counseling
2. sending cease-and-desist orders
3. campaigns

We went through all three scenarios with the experts and asked them what they would want a support system to do for them during these processes. Below we present their condensed responses in the form of stakeholder requirements. For each requirement, we also note by which stakeholder group it was brought up.

> **Requirement 1: Find clauses within a contract**
> Consumer counselors want to find clauses that are related to consumer inquiries within large contract documents.

Depending on the contract type, consumer counselors report that they often spend a lot of time searching for a specific clause in the documents provided by the consumer. Therefore, the first step during consumer counseling in which they would like a system to support them is finding specific clauses.

They did not see much value in trying to automate the assessment of clauses that are related directly to the inquiry because they will have to analyze the clauses anyway to explain them to the consumer. As pointed out before, this individual counseling is also an important funding source for the consumer protection agencies.

> **Requirement 2: Check complete documents**
> Consumer counselors want to check a whole document and want to be notified about potentially void clauses.

At the end of the counseling process, the consumer counselors would like to have the possibility to "quick-check" a whole contract for any void clauses, independently from the topic of the inquiry because they currently do not have the time to do that manually.

> **Requirement 3: Assessment based on law and court decisions**
> For consumer counselors, it is important that the assessment of the validity of clauses is not only based on the law but also on the latest court decisions.

It is important for the consumer counselors that the assessment of whether a clause is void or not is not only based on the law but also on relevant court decisions, which often interpret the vague consumer protection laws and introduce more concrete checkpoints.

Requirement 4: Full-text access
The consumer counselors at all points want to be able to access the full text of the contract, mostly because they do not want to have to rely on the assessments made by the system.

While our interviewees, in general, were very open and even enthusiastic about the idea of a support system, understandably, they do not want to have to rely blindly on the judgment of the system. Therefore, they want to be able to make their own judgments and need access to the full text of contracts at all times.

Requirement 5: Automatic documentation and checking
The legal team would like to have a way to automatically document clauses they identified as being void, and they would like to automatically check whether a certain clause was documented before.

The legal team generally did not saw much potential in applying a support tool during the process of sending cease-and-desist orders because the clauses are already pre-filtered by the counseling team, and every clause will be manually checked anyway because the result of the assessment is potentially an expensive legal process. However, they did see the possibility of automatizing or at least supporting the documentation of void clauses in a central database, like the "JURA" database, and also the process of checking in the database whether a specific clause was already documented by another organization.

Requirement 6: Automatic adherence check
The board member and the legal team would like to be able to automatically check the adherence to declarations to cease and desist periodically.

The board member among our interview partners and also the legal team want a tool to automatically check the adherence to declarations to cease and desist periodically because they do not have the resources to do that manually. It is not only important for them from a consumer protection perspective. Enforcing contractual penalties also is a possible new stream of revenues.

Requirement 7: Find relevant contracts for campaigns
The board member and the legal team would like to be able to automatically identify contracts that are relevant to a campaign from the internet or a corpus of documents.

Finally, the legal team saw the biggest opportunity in applying software support in campaigns in order to search for relevant documents, i.e., contracts that contain certain void clauses, on the internet or in large corpora. This is also desirable from the board member's perspective because such campaigns are important marketing tools, and they currently lack funding to conduct larger campaigns.

5.3.2. Consumers

For consumers, we provided one user scenario in which they visit the website of an online shop in order to buy a product and asked them what they would want a system to do for them during the process.

Requirement 8: Being informed about T&C at a glance
Consumers want to be informed about the T&C at a glance, either directly when entering the online shop or during checkout. They want an overall assessment of whether the T&C are "okay", as well as a short summary of the content.

The main point we heard from consumers is that they want to be informed at a glance. They do not want to have to use a separate website or application. The desire for short summaries is also in line with the results of the survey, where the length of the texts was one of the main reasons why consumers do not read T&C in the first place. There was a disagreement among consumers when they want to receive the information. Some interview partners would prefer to see the information as soon as they enter the online shop so that they do not spend time searching for items if they do not agree with the T&C. Others would prefer to receive the information during the checkout process when they have to indicate that they agree to the T&C.

Another interesting aspect was that consumers do not think about T&C in categories like lawful or void, but rather in categories like "okay" or "reasonable". This is also emphasized by the next requirement.

Requirement 9: Define own standards and priorities
Consumers want to be able to define their own standards against which the T&C should be evaluated and also prioritize different topics.

In some aspects, like privacy or withdrawal rights, some consumers find regulations desirable that go beyond the legal minimum, and they want to be able to tell a software that it should perform a check based on their personal requirements rather than just the legal regulations. In some other cases, they would still be willing to order, even if an online shop uses void clauses; in such cases, consumers would like to be able to de-priorities certain aspects.

5.4. System Requirements

Based on the stakeholder requirements we gathered, we derived system requirements and discussed their priorities with the expert stakeholders. System requirements are part of the solution domain, i.e., the domain in which we use (software) engineering to solve problems (Dick et al., 2017c, p. 135). Table 5.1 shows the list of system requirements we derived and which stakeholder requirement they were derived from. As mentioned at the beginning of this chapter, we want to focus on the core algorithms rather than a user application. Therefore, we did not derive any requirements regarding user interfaces or system integration. For each system requirement, the table also contains two prioritization, which can be either low, medium, or high. The *user priority* entry reflects the priority from a user perspective, i.e., from the perspective of the consumer counselors and the legal team at the consumer protection agencies. The *research priority* reflects the priority from the perspective of our research agenda.

While our action research approach is generally driven by user needs and real-world problems, we have to acknowledge that sometimes, things that are important to users do not fit a certain research agenda or have little scientific relevance. In our concrete example, this applies, e.g., to Optical Character Recognition (OCR) (SysRq 1). Especially during in-person counseling, most consumers provide their contracts in print. From the counselors perspective, it is therefore important that the system is also able to perform OCR in order to digitize printed contracts. For us, however, OCR is not part of our research agenda, and there are many commercial solutions available, which address the problem (see, e.g., Rahman et al. (2019) or Vithlani and Kumbharana (2015) for a comparison of existing tools). In order to make these opposing interests transparent, we provide both priority assessments, which we tried to balance during the conduction of our research.

ID	**requirement**	**related Req.**	**user prio**	**research prio**
SysRq 1	The system should be able to handle input from different sources, like websites, PDF documents, and printed documents.	Req. 1	high	low
SysRq 2	The system should be able to automatically separate a contract into its clauses.	Req. 1	high	high
SysRq 3	The system should be able to automatically detect the topic of individual clauses.	Req. 1	high	high
SysRq 4	The system should be able to automatically assess whether a clause is potentially void or not.	Req. 2 + Req. 8	high	high
SysRq 5	Users should be able to define their own evaluation criteria within the system and train new ML models.	Req. 3 + Req. 9	high	medium
SysRq 6	The system should be able to justify its decisions by providing the text passage which led to a certain decision.	Req. 4	high	medium
SysRq 7	The system should be able to automatically create reports which contain the text of the clause, its topic, its source, and why it was classified void.	Req. 5	medium	medium
SysRq 8	The user should be able to define a list of websites that should be checked periodically and which specific clauses should be checked.	Req. 6	high	medium
SysRq 9	The system should be able to notify a user, e.g., via email, in case a periodical check finds a void clause.	Req. 6	low	low

Table 5.1.: List of system requirements and their priority

ID	**requirement**	**related Req.**	**user prio**	**research prio**
SysRq 10	The system should be able to automatically identify documents and websites which represent T&C.	Req. 7	high	high
SysRq 11	The system should be able to automatically identify T&C that contain a certain (void) clause.	Req. 7	high	high
SysRq 12	The system should be able to assess T&C as a whole, as to whether or not they contain any potentially void clauses or undesired clauses.	Req. 8	high	medium
SysRq 13	The system should be able to shortly summarize the content of the T&C in natural language.	Req. 8	medium	high

Table 5.1.: List of system requirements and their priority

CHAPTER 6

Semantic Analysis of Standard Form Contracts

This chapter describes the core part of this thesis, which is the NLP-pipeline for the semantic analysis of standard form contracts that we developed. The pipeline consists of multiple consecutive modules that start with a standard form contract document as input and end with a machine-readable semantic representation of this document. In the following chapters, this semantic representation will be used to legally assess (Chapter 7) and summarize (Chapter 8) the standard form contracts. This chapter focuses on a conceptual description of the modules of the pipeline, a more technical description is given in Chapter 9, where the pipeline is implemented as part of a prototype user application.

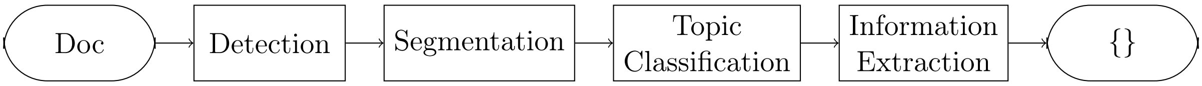


Figure 6.1.: NLP-pipeline for the semantic analysis of standard form contracts

Figure 6.1 shows the four modules of the pipeline, which align with the four sections of this chapter:

- **Detection (Section 6.1):** The first module is concerned with the automatic detection of documents that contain standard form contracts by classifying each document with regard to whether it represents a standard form contract or not.
- **Segmentation (Section 6.2):** Documents that were identified as standard form contracts by the first module are segmented in the next module. First, the main content is segmented from possible distractors like footers, then, each sentence is segmented, and finally, sentences are segmented into paragraphs, which are then organized according to the document hierarchy.

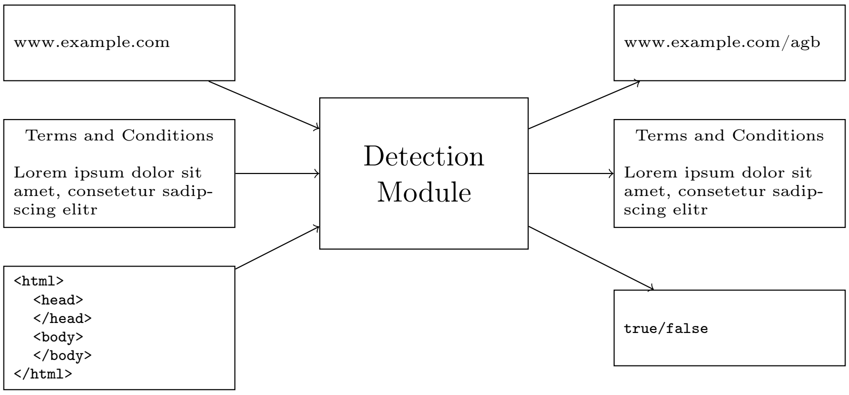


Figure 6.2.: Detection module with possible inputs and outputs

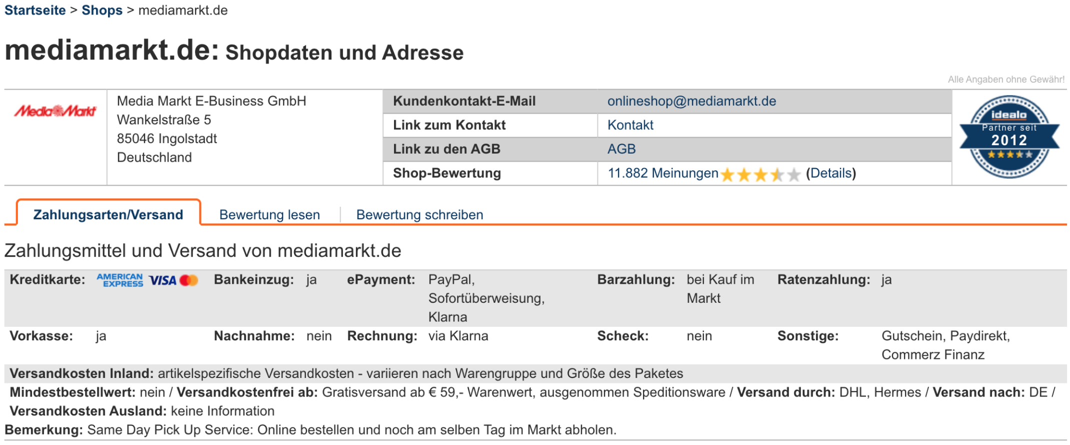
- **Topic Classification (Section 6.3):** The third module labels each clause with the topics it contains, based on the taxonomy introduced in Chapter 4.
- **Extraction (Section 6.4):** In the final step of the pipeline, relevant information is extracted based on the type identified clause topics. For a withdrawal clause, for example, the withdrawal period and the form of withdrawal are extracted, while for a warranty clause, the warranty period is extracted.

Combined, the four modules generate a machine-readable semantic representation of the input document. The results of this chapter are summarized in Section 6.5.

6.1. Automatic Detection

In order to automate or at least support the analysis and assessment of large amounts of standard form contracts, especially from the internet, not just their analysis and assessment but also their detection has to be automatized. While the task might sound simple, it is important to get this first step right because, in pipeline architectures, errors that are made at the beginning of the pipeline propagate through the whole process. Only if a document can be identified as a standard form contract it can be correctly processed subsequently. From a practical perspective, we want a detection module to be able to handle different input and output formats, depending on the situation. In the basic case, the input is a document, whether it is represented by text, HTML, or a URL, and the expected output is a binary classification whether the document contains T&C or not. However, we also want the module to be able to take a URL or HTML document as input and search through the outgoing links for T&C and return either the URL or content of the page. From an NLP perspective, both tasks, conceptually, can be treated as a binary document classification problem. A document should either be classified as standard form contract or not. In this section, we focus on this classification problem.

We conducted a pilot experiment in 2017 (Braun et al., 2017b) and a more elaborate version in 2020. This section first shortly describes the pilot experiment (Section 6.1.2) and then the final experiment in more detail (Section 6.1.3 and 6.1.4). In both experiments, we used T&C



mediamarkt.de: **Shopdaten und Adresse**

Media Markt E-Business GmbH
Wankelstraße 5
85046 Ingolstadt
Deutschland

Kundenkontakt-E-Mail	onlineshop@mediamarkt.de
Link zum Kontakt	Kontakt
Link zu den AGB	AGB
Shop-Bewertung	11.882 Meinungen ★★★★☆ (Details)

Zahlungsmittel und Versand von mediamarkt.de

Kreditkarte:	AMERICAN EXPRESS VISA	**Bankeinzug:**	ja	**ePayment:**	PayPal, Sofortüberweisung, Klarna	**Barzahlung:**	bei Kauf im Markt	**Ratenzahlung:**	ja
Vorkasse:	ja	**Nachnahme:**	nein	**Rechnung:**	via Klarna	**Scheck:**	nein	**Sonstige:**	Gutschein, Paydirekt, Commerz Finanz

Versandkosten Inland: artikelspezifische Versandkosten - variieren nach Warengruppe und Größe des Paketes
Mindestbestellwert: nein / **Versandkostenfrei ab:** Gratisversand ab € 59,- Warenwert, ausgenommen Speditionsware / **Versand durch:** DHL, Hermes / **Versand nach:** DE / **Versandkosten Ausland:** keine Information
Bemerkung: Same Day Pick Up Service: Online bestellen und noch am selben Tag im Markt abholen.

Figure 6.3.: Detailed shop listing on idealo.de, including link to the T&C[1]

from online shops, which we retrieved from price comparison website (see Section 6.1.1 for more details). Section 6.1.5 shows that the described approach can also be transferred to different kinds of standard form contracts. While the pilot experiment was only conducted with German T&C, the main experiment also included T&C in English.

6.1.1. Corpus

To test and train the different classification approaches, we needed a sufficiently large corpus of labeled T&C from German and English online shops. Since no such corpus was available, we built a new corpus by automatically parsing the list of merchants from two German price comparison websites ("Idealo"[2] and "Geizhals"[3]) that also offer a version of their website in English, targeted to the British market[4]. On both websites, shop operators manually report the URLs to their T&C pages, which we extracted from their semi-structured representation, shown in Figure 6.3, with a web-crawler we implemented using Python and the Beautiful Soup library[5].

In this way, we could crawl 4,459 manually annotated links to German T&C from Idealo and 1,335 from Geizhals. After removing duplicates, our corpus consisted of 4,875 distinct links to German T&C. We were able to download 4,869 pages from these 4,875 links. Six could not be downloaded because the websites were (permanently or temporarily) not available. In order to also have negative examples in our corpus (i.e., pages that do not contain T&C), we also downloaded the landing page of each shop (4,852) and a random other non-landing and non-T&C page (4,687). We chose this page by randomly selecting an outgoing internal link from the landing page and checking that it does not link to the T&C page or the landing page itself. We

[1] Source: `https://www.idealo.de/preisvergleich/Shop/285519.html`, last accessed: 13.05.2020
[2] `www.idealo.de`
[3] `www.geizhals.de`
[4] `www.idealo.co.uk` and `www.skinflint.co.uk`
[5] `https://www.crummy.com/software/BeautifulSoup/`

Language	**Page Type**	**Number of Pages**
German	T&C	4,869
	Landing	4,852
	Other	4,687
	Σ	*14,408*
English	T&C	543
	Landing	549
	Other	486
	Σ	*1,578*
Total	Σ	*15,986*

Table 6.1.: Corpus of pages from German and English online shops

Technique	**Precision**	**Recall**	**F-score**	**⌀t in s**
Naive Bayes	0.91	0.82	0.86	1.44
Rule-based URL analysis	0.80	0.54	0.64	0.001

Table 6.2.: Evaluation of different techniques for the automatic detection of T&C in German online shops (pilot)

performed the same process for the English versions of both websites and were able to download 543 T&C pages, 549 landing pages, and 486 random other pages. Table 6.1 shows an overview of the corpus. In total, our corpus consists of 15,986 HTML pages and their URLs.

We used the Article Extractor from BoilerPy3[6] to extract the plain text from the HTML pages by removing HTML tags, but also noise like navigation menus and footers. Since a surprisingly high number of pages consisted of invalid HTML, which can not be processed by BoilerPy, we first ran all pages through Beautiful Soup to fix the invalid HTML markup.

6.1.2. Pilot Experiment

When we conducted the pilot experiment in 2017, we did not yet have the full corpus described in Section 6.1.1. Instead, the corpus consisted of just 2,592 German T&C pages extracted from Idealo and 832 other, randomly extracted, pages from German online shops. We tested two approaches on this data set. The first approach was a rule-based URL classifier that checks the path and query parameters of a URL for an occurrence of one of the following strings: “agb”, “allgemeine_geschäftsbedingungen”, “allgemeinegeschäftsbedingungen”, “allgemeine-geschäftsbedingungen”, “allgemeine_geschaeftsbedingungen”, “allgemeinegeschaefts-bedingungen”, and “allgemeine-geschaeftsbedingungen”. As second approach, we implemented a naive Bayes classifier and trained it with a bag-of-words model from 400 pages (200 T&C and 200 others). We evaluated both approaches on the rest of the corpus. The results of the evaluation are shown in Table 6.2. In addition to precision, recall, and F1-score, we also evaluated how much time it took each of the techniques on average to classify a single document.

[6] `https://github.com/jmriebold/BoilerPy3`

We learned from this pilot that even a (literally) naive classification approach can achieve good results for this task. The results of the rule-based URL analysis were surprisingly bad, especially with regard to precision. On closer inspection of the results, we identified a problem with our corpus. When selecting the random pages for our negative examples, we did not properly ensure that this random page is not the T&C page of the online shop in the way we did for the final corpus previously described. Therefore, some of the pages that were labeled as non-T&C pages were actual T&C pages which were correctly classified by the rule-based URL analysis, which lead to "false false positives". Therefore, another lesson we learned from the pilot experiment was to select the negative examples in the corpus more carefully.

6.1.3. Approaches

Based on the experiences we made during the pilot experiment, we conducted the main experiment and compared four different techniques (two rule-based and two stochastic) on the larger corpus, including German and English pages.

Rule-based URL Analysis

The first rule-based approach we evaluated was a refined version of the rule-based URL analysis that we used in the pilot. It performs a simple keyword matching by analyzing whether the path or parameters passed in the URL contain any of the following keywords: "agb", "geschaefts-bedingungen", "geschaftsbedingungen", "terms", "conditions", "gtc", "tcs", "tac". In addition to simplifying the list of keywords, we also added additional keywords for English. Since this simplistic approach does not need any training data, we tested it on the complete corpus, i.e., 14,408 URLs from German online shops and 1,578 URLs from English online shops.

Rule-based Link-text Analysis

The second rule-based approach we investigated does not take the URL of a link into account, but its text, i.e., the text that is written between the `<a></a>` tags. This approach can not be evaluated in the same fashion as the other approaches since it does not take a link or its content as input but analyzes all outgoing links from a page, in our case, the landing page. Therefore we analyzed this approach by parsing all landing pages and checked whether the T&C link that is in the corpus could be retrieved. We again performed a simple matching of keywords. The list of phrases we searched for in the link texts are: "agb", "allgemeine geschäftsbedingungen", "geschaftsbedingungen", "terms", "conditions", and "t&c".

Logistic Regression

As a stochastic baseline, we trained a logistic regression classifier on a bag-of-words model of the documents using Scikit-learn (Pedregosa et al., 2011), i.e., we converted each document in a sparse count vector representation of the words it contains, after removing stopwords based

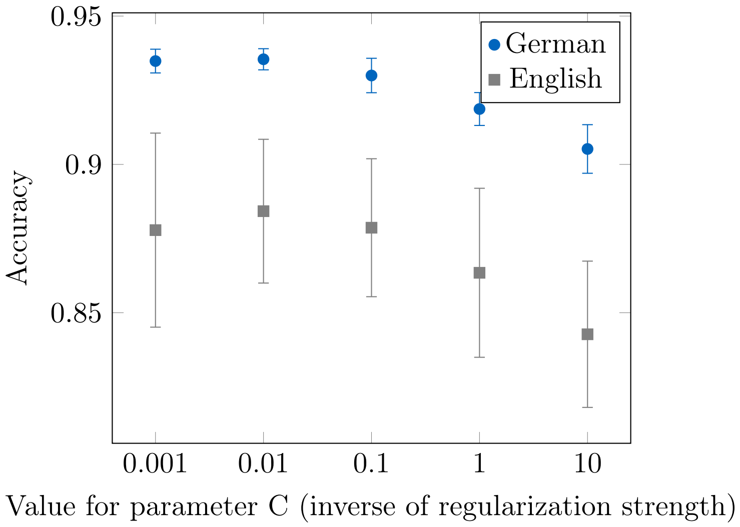


Figure 6.4.: Results of the grid search with 10-fold cross-validation to find the parameter for the regularization strength of the logistic regression

on the "Stopwords ISO"[7] lists for German and English. For the German corpus, the document vectors have almost 190,000 dimensions, and for the much smaller English corpus, 26,000 dimensions.

We split the corpus into a training (80%) and a test set (20%) while maintaining the original distribution of both classes (roughly 2:1), resulting in a German training data set of roughly 11,500 documents and an English training data set with roughly 1,200 documents. We used the training data set to perform a grid search with a stratified 10-fold cross-validation to find the parameter for the regularization strength. The results of the grid search are shown in Figure 6.4. For both languages, the best accuracy was achieved at $C = 0.01$. Finally, we evaluated the approach on the left-out test-set.

Transformer Model

Lastly, we used a transformer model that we fine-tuned for the task using Keras[8]. We used the Multilingual Universal Sentence Encoder transformer model[9] (Yang et al., 2019), which is pre-trained on 16 languages (including German and English) and produces vectors of 512 dimensions from the input text. Figure 6.6 shows a visualization of the document vectors the transformer model generated from the corpus. We used t-SNE (Maaten and Hinton, 2008) to reduce the 512-dimensional vectors to a three-dimensional space. Despite the low dimensionality, the clusters formed by the T&C and the other documents are still clearly recognizable.

[7] https://github.com/stopwords-iso/stopwords-iso

[8] https://keras.io/

[9] https://tfhub.dev/google/universal-sentence-encoder-multilingual-large/3

Language	Technique	Accuracy	Precision	Recall	F1-score
German	Rule-based URL analysis	0.97	0.99	0.91	0.95
	Logistic regression	0.93	0.98	0.82	0.89
	Transformer model (German)	0.93	0.93	0.93	0.93
	Transformer model (multiling.)	0.94	0.94	0.94	0.94
English	Rule-based URL analysis	0.93	0.99	0.81	0.89
	Logistic regression	0.89	0.96	0.72	0.81
	Transformer model (English)	0.91	0.92	0.90	0.90
	Transformer model (multiling.)	0.88	0.89	0.87	0.87
	Transformer model (German)	0.88	0.89	0.87	0.87

Table 6.3.: Evaluation of different techniques for the automatic detection of T&C in German and English online shops

For the classification, we added a hidden layer that uses the ReLU activation function, a dropout layer to prevent overfitting, and an output layer with one neuron that uses the sigmoid activation function to return the final binary classification to the transformer model. We used the same training/test split as before and performed a grid search with a stratified 10-fold cross-validation again to tune the hyper-parameters batch size, number of epochs, neurons in the hidden layer, and the dropout-rate for German and English. We performed several iterations of it in which we refined the search space. In the final iteration, we explored the following search space for the hyper-parameters:

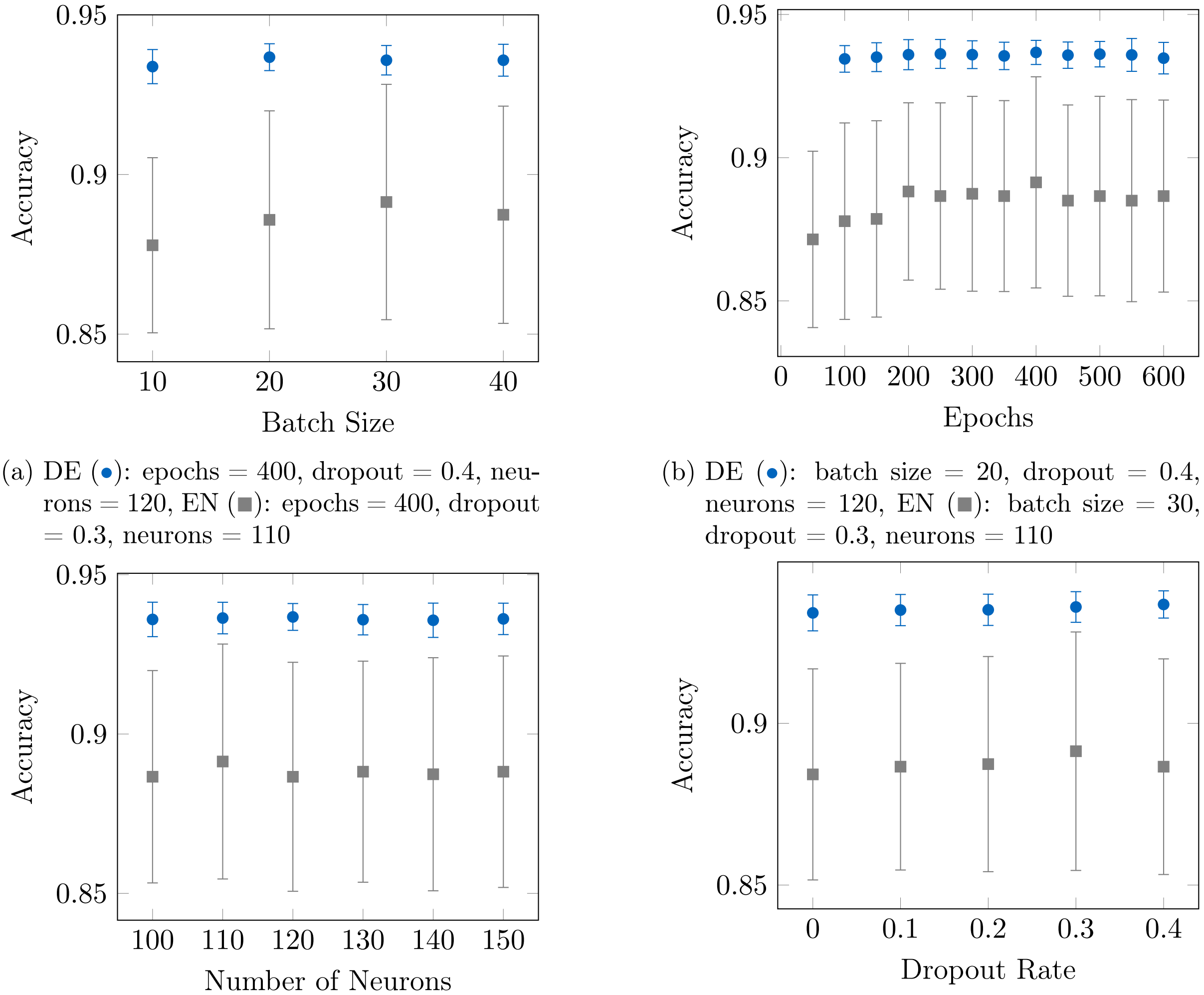
- batch size: 10, 20, 30, 40
- epochs: 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600
- number of neurons: 100, 110, 120, 130, 140, 150
- dropout rate: 0.0, 0.1, 0.2, 0.3, 0.4

This generates 1,440 different settings per language for each of which we performed a 10-fold cross-validation, which leads to 28,800 training and test runs, which took several days of computation. For German, we found the best-performing combination of parameter to be batch size = 20, epochs = 400, neurons = 120, and dropout rate = 0.4, for English it was batch size = 30, epochs = 400, neurons = 110, and dropout rate = 0.3 (see Figure 6.5).

For each language, we trained the network on the training set with the above parameters. First, we trained and tested only on the same language. In a second step, we also trained the classifier jointly on both languages and then tested it separately on German and English.

6.1.4. Evaluation

The results of the evaluation of the different approaches are shown in Table 6.3. The rule-based URL analysis performed very well on both languages, with a precision of 0.99 for German and English. The very high recall also shows that we could eliminate the problem we had with the corpus of the pilot experiment. Most false negatives (i.e., not detected T&C pages) originated



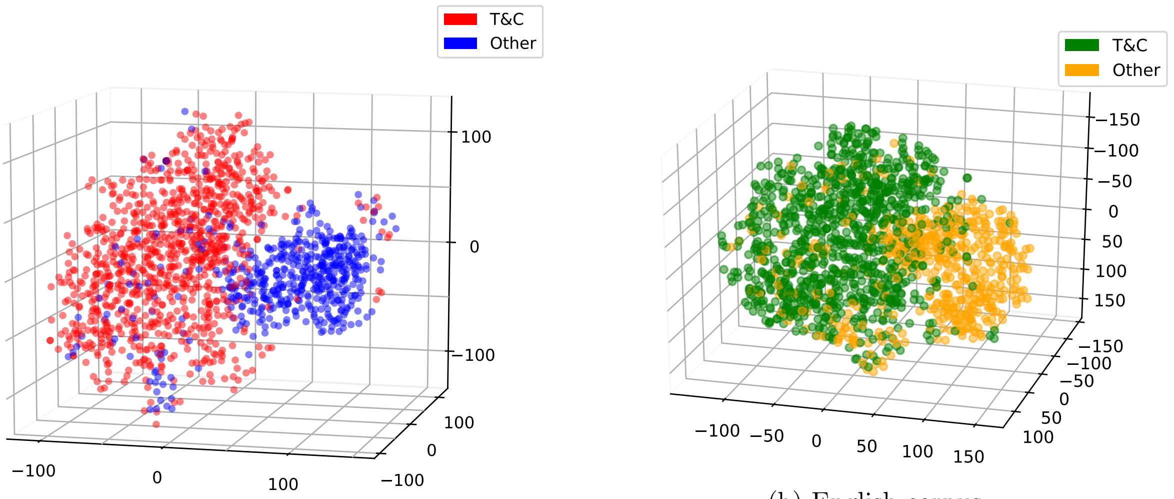
(a) DE (●): epochs = 400, dropout = 0.4, neurons = 120, EN (■): epochs = 400, dropout = 0.3, neurons = 110

(b) DE (●): batch size = 20, dropout = 0.4, neurons = 120, EN (■): batch size = 30, dropout = 0.3, neurons = 110

(c) DE (●): batch size = 20, epochs = 400, dropout = 0.4, EN (■): batch size = 30, epochs = 400, dropout = 0.3

(d) DE (●): batch size = 20, epochs = 400, neurons = 120, EN (■): batch size = 30, epochs = 400, neurons = 110

Figure 6.5.: Results of the grid search with 10-fold cross-validation to find the parameter for batch size, epochs, number of neurons, and dropout rate



(a) German corpus

(b) English corpus

Figure 6.6.: Visualization of the document vectors generated by the transformer model projected to three-dimensional space using t-SNE

from URLs that do not contain any human-readable identifiers and mostly consisted of numerical IDs. In the German corpus, some URLs did contain a misspelled version of "geschaeftsbedingungen" and were therefore not detected. The significantly lower recall in English is partially caused by the greater variety in naming. In German, the term "Allgemeine Geschäftsbedingungen" directly originates from the legislation. Since this is not the case for "Terms and Conditions" in English, a wider variety of names is used.

With the rule-based link-text analysis, we were able to find the correct T&C page on 4,383 out of 4,837 German landing pages. In 454 cases, the T&C page was not found. In no case, a wrong page was identified as T&C. Hence, we were able to extract the T&C page correctly in 90.6% of all cases. In English, we were able to extract the T&C page from 425 of 539 landing pages, which is 78.8% of all cases. It has to be noted that in cases where we were not able to find the T&C page using this approach, the reason is not necessarily a technical limitation of this approach. It is not mandatory for online shops to link to their T&C from their landing page. Therefore, there might be cases (and we do know that there are such cases in our data) in which the approach correctly did not extract any link. The percentage value of correctly extracted links can, therefore, alone, not be used as a measurement of correctness. However, the value does give an idea of the usefulness of the technique in practice.

Using bag-of-words to train a logistic regression classifier worked well in both languages (F1-score 0.87 in German and 0.82 in English) and achieved better recall values than the rule-based URL analysis (see Table 6.3), despite the relatively sparse training data in English. However, with regard to precision and F1-score, it performed considerably worse than the rule-based URL analysis.

For both languages, the transformer model performed significantly better than the logistic regression (F1-score 0.93 in German and 0.90 in English). The rule-based URL analysis approach, however, still performed better in both cases. While for German, the results could be slightly improved by training jointly on both languages, the results in English got slightly worse, possibly because English is significantly underrepresented in the joint dataset. Training the network only on the German data and testing it on the English data produced the same results as training it on the multilingual dataset.

With a precision close to 1.0 for both languages, the rule-based URL analysis performed extremely well. In addition to having a high precision, the approach is also computationally very cheap and can analyze thousands of links within a part of a second. The deep learning approach could only outperform the URL analysis with respect to recall. For application, we, therefore, suggest combining both approaches: First, use the faster and more precise URL analysis and if it does not provide any results, then use the deep learning approach with a higher recall.

An additional advantage of the deep learning approach compared to the URL analyzer is the fact that the deep learning approach uses the actual text for the classification and is, therefore, better transferable to other applications and could, e.g., also be used to classify emails or PDF documents. Moreover, since we use a multilingual transformer model, we are also able to transfer our results between different languages and still receive useful results.

6.1.5. Transferability

To test whether the models we trained for the deep learning and the logistic regression classifier could also be transferred to other types of standard form contracts, we evaluated them on a set of German general business conditions ("AGB") from ten of the biggest banks in Germany:

- Commerzbank,
- Deutsche Bank,
- Deutsche Kreditbank,
- HypoVereinsbank,
- ING,
- Postbank,
- PSD Bank,
- Sparkassen,
- Volksbanken, and
- Volkswagen Bank.

For each bank, we used the version that was valid in April 2020. Since all banks provided their conditions in PDF format, we used Apache Tika[10] to extract the textual content from the PDF files. We classified the ten documents with the logistic regression model and the deep learning model we previously trained on the German T&C data set. With the logistic regression classifier, we were able to correctly classify all but one document (see Table 6.4). In the one instance where

[10] https://tika.apache.org/

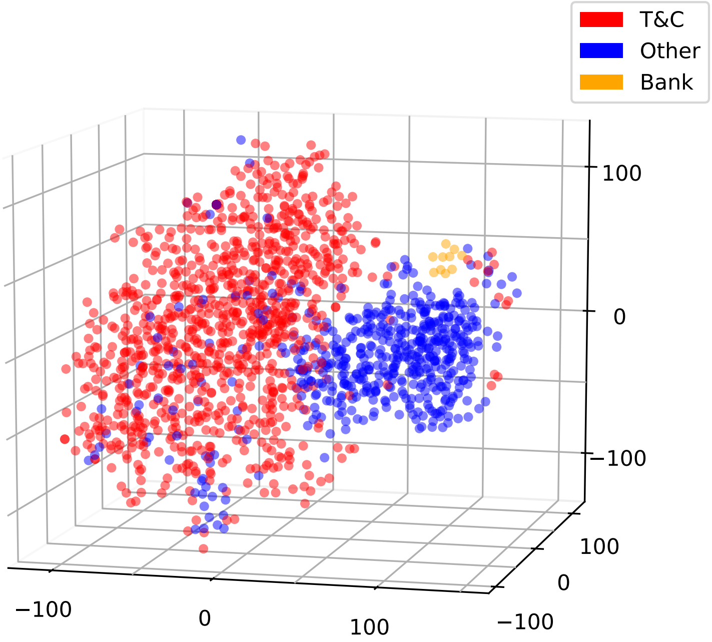


Figure 6.7.: Visualization of the vectors produced by the transformer model for the German corpus and the bank documents

Technique	Precision	Recall	F-score
Logistic regression	1.0	0.90	0.95
Neural network	1.0	0.90	0.95

Table 6.4.: Evaluation of the transferability of the the models to the banking domain

a document was classified incorrectly (Volksbanken), the most likely reason is that during the text extraction process for this document, most words were for some reason split by the parser in many parts (e.g., *"Die Bank ist be fugt, über ju ris ti sche Per so nen und im Han dels register [...]"*). For a bag-of-words approach, such an error during the extraction will almost certainly always lead to wrong classification results.

Despite the extraction problem, the vectors generated by the transformer model for the general business conditions were tightly clustered (see Figure 6.7). Nevertheless, when we used the neural network with the model trained on the German corpus, the same document was wrongly classified, and the nine other documents were classified correctly.

6.2. Segmentation

The segmentation of the content of documents, their tokens, sentences, and sections, are among the first common preprocessing steps in every NLP pipeline. Although these tasks often seem trivial to humans, they are far from trivial to machines, as the evaluations in this section show. While there is often little glory to tasks like that, the quality in which they are conducted is crucial for the overall performance of a pipeline. If sentence boundaries are not detected correctly, dependency graphs can not be built. If paragraphs are not separated correctly, clauses can not be identified.

While tasks like sentence segmentation (also called sentence splitting) might seem like domain-independent problems that can be solved with a one size fits all solution, legal language does not only have its own rules when it comes to semantics but also when it comes to syntax. The character sequence *"Art. 73 Abs. 1 Nr. 10 lit. a GG"*, for example, would probably seem odd in any text but a legal one and might prove to be challenging for sentence segmenters that were not designed with legal texts in mind. In addition, there are also language specificities that can make sentence segmenter incompatible, even between our, in many aspects so similar, Germanic languages. An example of such a small difference with a potentially large impact are the interchanged decimal and thousands separators in German and English (1,234.56 in English, but 1.234,56 in German).

This section compares multiple approaches for content extraction (Section 6.2.1), sentence segmentation (Section 6.2.2), and paragraph segmentation (Section 6.2.3) on German and English standard form contracts. It will describe what specificities of this document type lead to problems with existing solutions and how these problems can be addressed.

6.2.1. Content Extraction

Content extraction is the process of extracting the main text of a document by removing template and boilerplate elements, like page footers, headers, or navigation elements. While, in general, all document types can contain such elements, whether it is an email, a PDF, or an HTML file, this process is particularly challenging and important for HTML documents, not only do HTML files usually contain more such elements as PDFs for example, in an HTML file, the position of an element within the file itself does not necessarily correspond to its rendered position on the screen. Pop-ups, boxes, tables, and the like can be positioned almost freely, using CSS, independent from their position in the code.

The goal of the process is to remove as many of the unnecessary clutter as possible while preserving all the important content. If too much is removed, important information might get lost; if too little is removed, the possibly out-of-domain content might negatively influence the performance of the later steps in the NLP-pipeline and distract users while working with the text. While we want to avoid both, in the trade-off one usually has to make, removing too little is usually the less severe problem, it is at least for our use-case.

For the experiment in Section 6.1, we used BoilerPy3, the Python 3 port of Boilerpipe[11], which is arguably one of the most widely used tools for the task. For the mere detection of standard form contracts, the precise content extraction is not that important. For the legal assessment, however, it is. Therefore, we conducted a comparison of different content extractors that are described in the literature (Kovačič, 2012; Lejeune and Zhu, 2018; Barbaresi, 2020) as best performing and evaluated how they perform on T&C from German and English online shops.

[11] `https://code.google.com/archive/p/boilerpipe/`

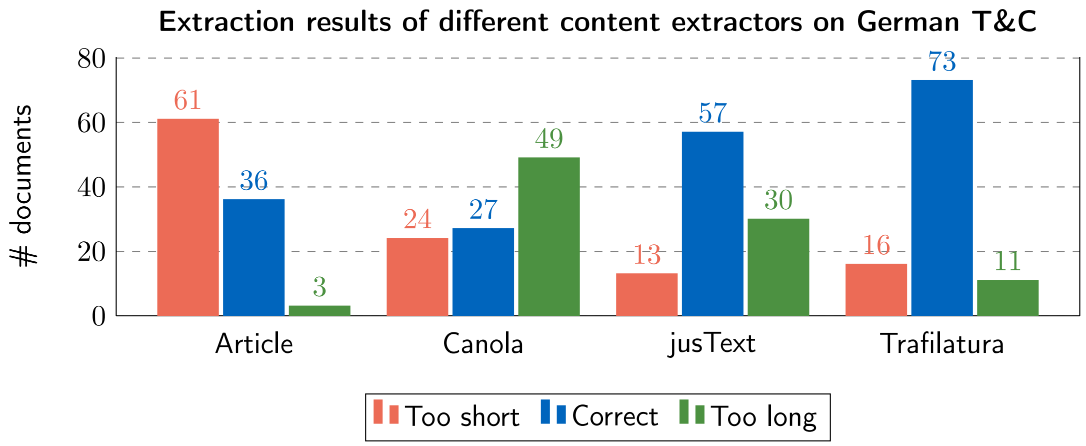


Figure 6.8.: Extraction results of different content extractors on German T&C

For each language we took the first 100 T&C from our corpus and run four different content extractors on them: BoilerPy3 with the Article extractor, BoilerPy3 with the Canola extractor, jusText[12], and Trafilatura[13]. We then manually checked and annotated each of the extraction, whether the content was extracted correctly or whether there was too much or too little extracted. In cases where both error classes could be found (e.g., too much extracted at the beginning and too little at the end), we annotated the extraction with the more severe error, i.e., too little extracted.

Boilerpipe is based on work by Kohlschütter et al. (2010). It is based on shallow text features like the number of tokens in an HTML-block, the number of links contained in a block, the average sentence length, the number of full stops, and so on. The Article extractor is optimized for the extraction of newspaper articles (the L3S-GN1[14] data set which was extracted from Google News was used for the creation), while the Canola extractor is based on the CANOLA corpus (Steger and Stemle, 2009), which consists of 216 web pages that were extracted through the Yahoo search.

jusText was originally developed by Pomikálek (2011) and works with similar shallow text features than boilerpipe, like the length of a block and the density of links within it. It was developed and evaluated using both of the aforementioned data sets and, in addition, the data set from the Cleaneval competition (Baroni et al., 2008).

Unlike the other two libraries, Trafilatura, which was developed by Barbaresi (2019), preserves some of the structure of the original document when extracting the textual content. Preserving parts of the structure of the original can make the later segmentation of paragraphs (see Section 6.2.3) easier. Trafilatura also uses the structural information, in addition to shallow text features, for the extraction algorithm itself.

The result for the evaluation on German texts is shown in Figure 6.8 and the result for the evaluation on English texts is shown in Figure 6.9. Two facts are visible at first glance: all three

[12] https://github.com/miso-belica/jusText
[13] https://github.com/adbar/trafilatura
[14] http://www.l3s.de/~kohlschuetter/boilerplate/

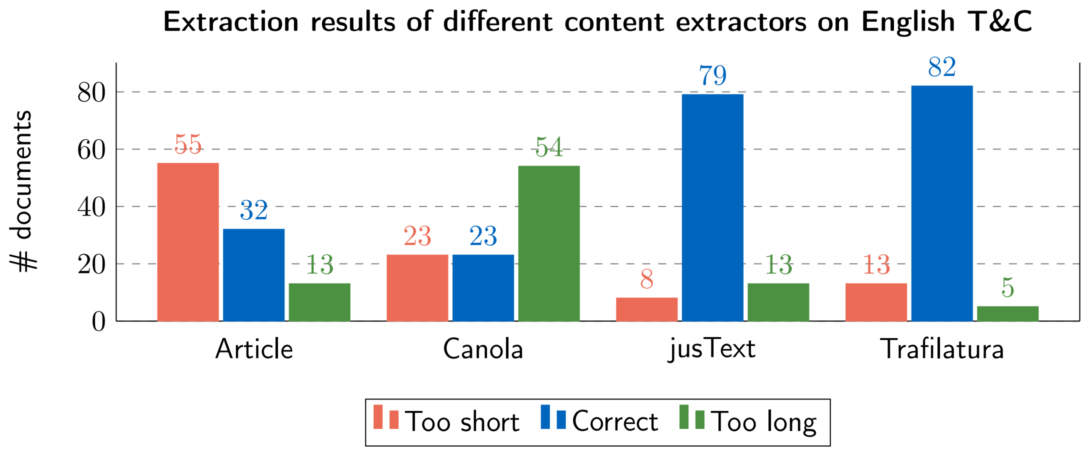


Figure 6.9.: Extraction results of different content extractors on English T&C

Language	Article	Canola	jusText	Trafilatura
German	-0.25	0.03	0.44	0.62
English	-0.23	0.00	0.71	0.69
Overall	-0.24	0.02	0.58	0.66

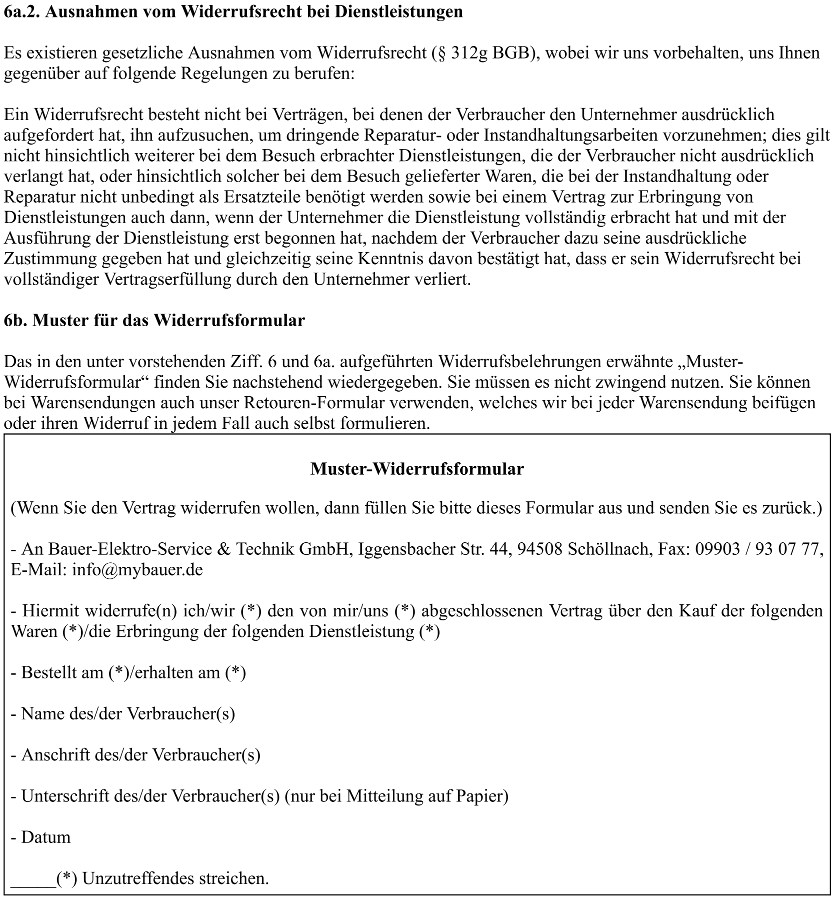
Table 6.5.: Average extraction performance of different content extractors

extractors performed better on English texts than on Germans, and Trafilatura extracted most texts correctly in both languages. By attaching a value to each of the classes (+1 for a correctly extracted document, -1 for too short, and 0 for too long) and average the values over the number of documents, we can also calculate a score for each extractor, in order to have a more direct comparison. The resulting score is between -1.00 (every extraction was too short) and 1.00 (every extraction was correct). The results in Table 6.5 show the scores for the different extractors in the different languages. Although Trafilatura extracted most texts correctly in English, jusText still received a higher score because it made fewer errors where it extracted too little text, a failure which is penalized by the score. In total, Trafilatura still performed best.

A closer inspection of the occurring errors, especially for the Article extractor, revealed that most of them were caused by domain-specific characteristics of T&C. The two main sources for too short extraction were the usage of a complicated multi-level paragraph enumeration and the inclusion of withdrawal form templates in the T&C.

Figure 6.10 shows a typical representation of a withdrawal form template within the T&C that leads to extraction problems. The form is separated from the rest of the content with a box (`div`-element), which is interpreted by the Article extractor as the end of the main content; therefore, the box itself and everything that follows is not extracted.

Some T&C use a combination of Latin (I, II, III, ...) and Arabic (1, 1.1, 1.2, 2., ...) numbers to enumerate their sections and paragraphs. Usually, the Latin numbers denote the larger sections of which only two or three exist. In such cases, the Article extractor almost always only extracted the first of these sections and did not extract the rest of the content.



6a.2. Ausnahmen vom Widerrufsrecht bei Dienstleistungen

Es existieren gesetzliche Ausnahmen vom Widerrufsrecht (§ 312g BGB), wobei wir uns vorbehalten, uns Ihnen gegenüber auf folgende Regelungen zu berufen:

Ein Widerrufsrecht besteht nicht bei Verträgen, bei denen der Verbraucher den Unternehmer ausdrücklich aufgefordert hat, ihn aufzusuchen, um dringende Reparatur- oder Instandhaltungsarbeiten vorzunehmen; dies gilt nicht hinsichtlich weiterer bei dem Besuch erbrachter Dienstleistungen, die der Verbraucher nicht ausdrücklich verlangt hat, oder hinsichtlich solcher bei dem Besuch gelieferter Waren, die bei der Instandhaltung oder Reparatur nicht unbedingt als Ersatzteile benötigt werden sowie bei einem Vertrag zur Erbringung von Dienstleistungen auch dann, wenn der Unternehmer die Dienstleistung vollständig erbracht hat und mit der Ausführung der Dienstleistung erst begonnen hat, nachdem der Verbraucher dazu seine ausdrückliche Zustimmung gegeben hat und gleichzeitig seine Kenntnis davon bestätigt hat, dass er sein Widerrufsrecht bei vollständiger Vertragserfüllung durch den Unternehmer verliert.

6b. Muster für das Widerrufsformular

Das in den unter vorstehenden Ziff. 6 und 6a. aufgeführten Widerrufsbelehrungen erwähnte „Muster-Widerrufsformular“ finden Sie nachstehend wiedergegeben. Sie müssen es nicht zwingend nutzen. Sie können bei Warensendungen auch unser Retouren-Formular verwenden, welches wir bei jeder Warensendung beifügen oder ihren Widerruf in jedem Fall auch selbst formulieren.

Muster-Widerrufsformular

(Wenn Sie den Vertrag widerrufen wollen, dann füllen Sie bitte dieses Formular aus und senden Sie es zurück.)

- An Bauer-Elektro-Service & Technik GmbH, Iggensbacher Str. 44, 94508 Schöllnach, Fax: 09903 / 93 07 77, E-Mail: info@mybauer.de

- Hiermit widerrufe(n) ich/wir (*) den von mir/uns (*) abgeschlossenen Vertrag über den Kauf der folgenden Waren (*)/die Erbringung der folgenden Dienstleistung (*)

- Bestellt am (*)/erhalten am (*)

- Name des/der Verbraucher(s)

- Anschrift des/der Verbraucher(s)

- Unterschrift des/der Verbraucher(s) (nur bei Mitteilung auf Papier)

- Datum

_____(*) Unzutreffendes streichen.

Figure 6.10.: Withdrawal form template within the T&C of a German online shop[15]

In cases where too much text was extracted, the reason was most of the time either a banner with information about the usage of cookies on the website or that the footer of the shop was not correctly identified and extracted as content. Especially in English, Trafilatura seemed to have problems extracting paragraph headings and instead removed them from the content at least five times.

Despite the fact that jusText performed better in English, we decided to use Trafilatura for both languages. The unique feature of preserving some parts of the document structure that only Trafilatura offers, makes the paragraph segmentation and the sentence segmentation easier. Handling two different output formats from Trafilatura and jusText would also make the pipeline more complex and difficult to maintain. The XML-output produced by Trafilatura contains mainly four types of elements: text blocks (`<p>`), spans (`<span>`), headings (`<head>`), and lists (`<list>`), which consist of items (`<item>`).

[15] `https://www.mybauer.de/inhalte/14183941_AGB.html;jsessionid=aaaKjpkAiyJJnhLaYwwix`, last accessed 2020-06-09

6.2.2. Sentence Segmentation

Although the task of splitting a text into its sentences might seem easy, if not trivial, machines still not always perform the task perfectly, especially for legal texts, as shown by Savelka and Ashley (2017) and Sanchez (2019).

For the task of sentence segmentation (or splitting), we decided to compare three different libraries: The sentence splitter from the NLTK library, the de-facto standard NLP library for Python, spaCy, an NLP-library which has demonstrated very good results in different evaluations for different tasks (Al Omran and Treude, 2017; Schmitt et al., 2019), and finally SoMaJo library[16] from Proisl and Uhrig (2016), which did not only win the EmpiriST 2015 shared task on the automatic linguistic annotation of computer-mediated communication / social media (Beißwenger et al., 2016) but was also already successfully applied to legal texts by Leitner et al. (2020). All three libraries do not just split the sentences but also tokenize the content of the sentences. We will show in the evaluation that many of the errors that occur during the sentence segmentation are based on tokens that are falsely split into two parts.

The NLTK sentence splitter is an implementation of the unsupervised multilingual sentence boundary detection from Kiss and Strunk (2006). It is based on the assumption that the main source of failure in sentence splitting are abbreviations. The approach, therefore, is to first learn the abbreviations based on features like length, internal periods, and occurrence without final periods and then use this knowledge for the sentence splitting. Kiss and Strunk mainly evaluate their work on corpora of news articles.

spaCy offers two different approaches to sentence segmentation. The default approach, which is evaluated here, uses the dependency parser to detect sentence boundaries. The German model is trained on the TIGER corpus (Brants et al., 2004) and the WikiNER corpus (Nothman et al., 2013). The English model is trained on the OntoNotes corpus (Hovy et al., 2006), which consists of a number of different text genres, including news articles, conversational telephone speech, weblogs, usenet newsgroups, broadcast, and talk shows. As an alternative, computationally cheaper method, spaCy also offers a simple rule-based module for sentence splitting.

SoMaJo, in contrast, works purely rule-based. It uses regular expressions and a list of abbreviations from the crowd-sourced online lexicon Wiktionary (see Section 8.6.2.5 for more information on Wiktionary). SoMaJo has a list of more than 1,000 abbreviations for German and English each. The authors evaluated their approach on the EmpiriST 2015 (Beißwenger et al., 2016) shared task social media data. (Proisl and Uhrig, 2016)

To compare the different approaches, we randomly selected ten German and ten English T&C from the 73, respectively 82, T&C where the content was extracted correctly by Trafilatura and annotated the sentence boundaries manually. On average, the documents consisted of 225 sentences, which results, in total, in 4,505 annotated sentences, 2,507 in German, and 1,998 in English. During the segmentation of sentences, two errors can occur in principle: Either the end of a sentence is not detected, resulting in one element consisting of two or more sentences, or a sentence is split before its end, resulting in two or more elements for the same sentence. In the

[16] `https://github.com/tsproisl/SoMaJo`

Document ID	Language	# Sentences	NLTK	spaCy	SoMaJo
1	de	565	580	636	565
2	de	147	195	168	147
3	de	101	107	108	101
4	de	224	236	714	224
5	de	610	690	689	610
6	de	224	360	371	224
7	de	136	144	152	136
8	de	84	89	120	84
9	de	302	355	412	302
10	de	114	133	249	114
Σ	de	2,507	2,889	3,619	2,507
1	en	145	145	192	145
2	en	265	270	286	265
3	en	115	116	155	115
4	en	24	24	25	24
5	en	283	295	326	283
6	en	96	97	100	96
7	en	450	452	462	450
8	en	218	219	265	218
9	en	317	320	397	317
10	en	85	85	87	85
Σ	en	1,998	2,023	2,295	1,998

Table 6.6.: Comparison of the number of sentences extracted by different sentence splitters

evaluation, only the later error, i.e., sentences split before their end, was present. The results are shown in Table 6.6.

In our evaluation, SoMaJo and its rule-based approach split every sentence in exactly the same way as it was done in the manual annotation, in both languages, English and German. The second-best performance was shown by NLTK, which, in German, split 382 times too often (+15%) and 25 times too often (+1%) in English. spaCy was off by 44% or 1,112 sentences in German and 15% or 297 sentences in English. The staggering performance differences in English and German by the data-drive approaches might quite well be explainable by the larger and, especially in the case of spaCy, more diverse English corpora they were trained on.

As it is the case for the content extraction, it seems that the specific enumeration of sections and paragraphs that is used by some T&C poses a challenge to the algorithms. Table 6.7 shows an example where NLTK and spaCy both failed to correctly handle the two levels of enumeration used in a German T&C page. Interestingly, they failed on different levels. While NLTK did not correctly process the first-level enumeration, spaCy failed on the second level. SoMaJo handled both levels correctly.

Enumerations were the main source of failure for spaCy and even more so for NLTK. This also explains while NLTK performed best on short documents, as visible in Table 6.6, which often use no or little enumeration.

For spaCy, we also found instances where references to laws were falsely detected as sentence

#	NLTK	spaCy	SoMaJo
1	1.	1. Allgemeines	1. Allgemeines
2	Allgemeines	(1)	(1) Die nachstehenden allgemeinen Geschäftsbedingungen [...]
3	(1) Die nachstehenden allgemeinen Geschäftsbedingungen [...]	Die nachstehenden allgemeinen Geschäftsbedingungen [...]	

Table 6.7.: Splitting results for the lines “1. Allgemeines”, “(1) Die nachstehenden allgemeinen Geschäftsbedingungen [...]”

#	NLTK	spaCy	SoMaJo
1	Fax.	Fax	Fax.: (+49) (0)7023/6210-200
2	: (+49) (0)7023/6210-200	.	
3		:	
4		(+49)	
5		(0)7023/6210-200	

Table 6.8.: Splitting results for the line “Fax.: (+49) (0)7023/6210-200”

boundaries. The line “(3) Informationspflicht nach § 36 VSBG” (Information obligation according to § 36 VSGB) was, for example, split to “(3)”, “Informationspflicht nach § 36”, and “VSGB”.

We also found some more domain-independent problems with the sentence segmentation. NLTK and spaCy, for example, both had problems to correctly identify the line “Fax.: (+49) (0)7023/6210-200”[17] as one sentence, as shown in Table 6.8. spaCy also failed to recognize the German abbreviation “vgl.” (cf.) and identified it wrongly as the end of a sentence.

Given the results of this evaluation, it was an easy decision to choose SoMaJo for our NLP-pipeline.

6.2.3. Paragraph Segmentation

From a syntactic perspective, sentences constitute a self-contained and closed structure, which can be analyzed isolated. From a semantic perspective, that is obviously not the case. Linguistic but also non-linguistic contexts influence the meaning of language. That also applies to legal language and contracts, especially when it comes to assessing their lawfulness. For this task, it is, for example, important whether the drafter of the contract is a company, a government agency, or a private person and where they are based. It is also important to take into account when a contract was closed because the applicable law might vary based on the time. While these non-linguistic contexts can not or not always be derived from the text, we can sometimes derive them from meta-data.

[17] For anonymization reasons, the digits were changed from the original number.

At this stage of the pipeline, we are interested in the linguistic context, i.e., the sentences that surround a given sentence. According to Merriam-Webster (2020d), a *paragraph* is "a subdivision of a written composition that consists of one or more sentences, deals with one point or gives the words of one speaker, and begins on a new usually indented line". In German, the word "Paragraph" additionally has a specific meaning in the context of legal texts, where it is the highest level in which laws, contracts, and other legal documents are divided, usually denoted with the paragraph symbol (§). In international and especially European law, articles take on this role.

Here, we use the word in the non-legal denotation, i.e., to describe a set of sentences that deal with the same aspect. While this definition usually applies to "legal paragraphs", it also applies to subsets of them. T&C very commonly contain, for example, a paragraph about the right of withdrawal, which contains multiple sub-paragraphs, e.g., about the withdrawal period or exemptions from the right of withdrawal (e.g., for tailored products). Our goal is not only to segment these sub-units but also to map their hierarchical relationship. This additional information can be used to improve further processing steps like the topic classification (see Section 6.3): if two sub-sections of the same section are related to withdrawal, it is very likely that a third sub-section will also be related to the same topic.

There are two general ways to approach the task of paragraph segmentation. One can either try to identify paragraphs based on an analysis of the text itself or by analyzing the existing structure of a document. The first approach is often applied to spoken language (or more accurately to its transcription) and other sources that do not have an explicit structure in their text and is also interesting for NLG applications, like automatic text summarization, which usually creates unstructured texts (Sporleder and Lapata, 2006). Hearst (1997), for example, counted how many words given sentences have in common and how many new words a sentence introduce to detect paragraph boundaries. Sporleder and Lapata (2006) used textual, syntactic, and discourse information to segment paragraphs in English, German, and Greek. Bolshakov and Gelbukh (2001) used a text cohesion metric to segment paragraphs. They proposed to use databases that contain semantic links and information of collocations, like CrossLexica for Russian (Bolshakov, 2013), to calculate the cohesion between two sentences. Sentences with high cohesion will form a paragraph. For spoken texts, prosodic information can be used as an additional source, as shown by Lai et al. (2016).

Especially in HTML documents, but also in PDF, Word, and other markup document formats, there is already a lot of structured information given by the author of a text. The boundaries of paragraphs are usually given by line-breaks, or, in the case of HTML, even explicit structure like `<p>`-Tags. The challenge that remains is the mapping of the hierarchy of the different paragraphs. Although HTML is a very structured and hierarchical format, the text itself is usually, with the exception of lists (`<ul>` and `<ol>`), not organized hierarchically, as shown in Listing 6.1. Although the section "Anwendungsbereich" is a subsection of "Kaufbedingungen", which is a subsection of "Geschäftsbedingungen", they all belong to the same parent element and are therefore on the same level in the document hierarchy.

```
<div>
    <h1>Geschäftsbedingungen</h1>
    <p>...</p>

    <h2>1. PRÄAMBEL</h2>
    <p>...</p>
    <p>...</p>

    <h2>2. KAUFBEDINGUNGEN</h2>
    <p>...</p>

    <h3>2.1 Anwendungsbereich</h3>
    <p>...</p>

    ...
</div>
```

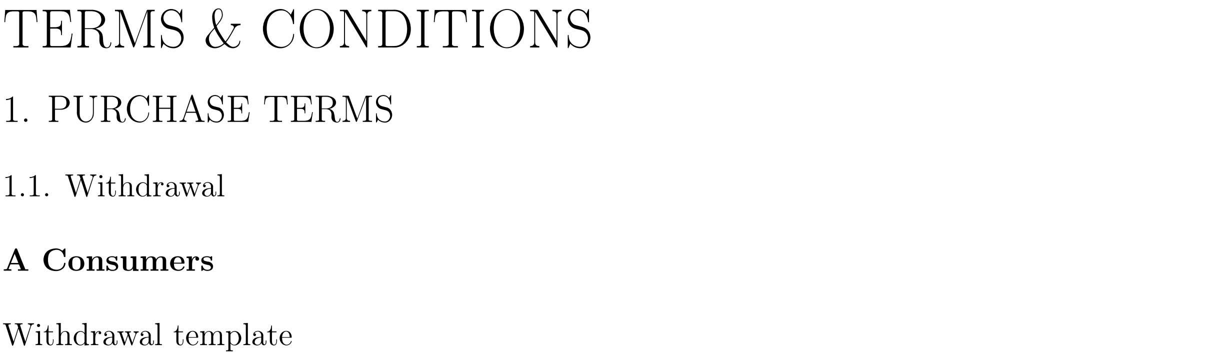
Listing 6.1: Excerpt from the T&C of a large German sportswear brand

6.2.3.1. Segmentation Rules

Although the text hierarchy is not explicitly given in Listing 6.1, it can be easily inferred from the different <h>-tags used for the markup of the headings. Unfortunately, we found that we can not rely on these annotations since many websites use <p> or <span> tags, which are styled with CSS, to markup heading in their T&C pages. Even the T&C shown in Listing 6.1 use <span> tags for the next lower level headings instead of <h4>. This corresponds with the findings of Manabe and Tajima (2015), who found that less than one-third of the headings they analyzed were marked-up with an <h> tag.

Manabe and Tajima (2015) suggest to use visual information, like font-size and text-decoration, to identify heading and their position in the hierarchy of a text, because, they argue, that headings on a higher level have a more “prominent visual style”, e.g., bigger font size. One of the assumptions they made is that “headings and non-heading segments never have the same visual style” (Manabe and Tajima, 2015). However, for T&C pages, we found that this is regularly not the case, especially for the lowest level of headings, which often use the same visual style as the content of the paragraphs. We also found that T&C pages use explicit numbering of their headings (1, 1.1, 1.1.1) much more often than other types of pages. Based on these observations, we developed an algorithm for the hierarchical segmentation of T&C pages that are based on the structure of the page, its visual style, and the content of the headings.

The algorithm takes as input the content extracted by Trafilatura, which, as mentioned before, preserves some of the structural elements of the pages (see Section 6.2.1). An (artificial) example of such input is shown in Listing 6.2. Headings which are marked with a <head> tag by Trafilatura, based on HTML <h> tags, can immediately be identified based on the existing structure. For each <p> element, we have to decide whether it represents a heading or a text paragraph. For this task, we identify six different styles of headings with regular expressions:



TERMS & CONDITIONS

1. PURCHASE TERMS

1.1. Withdrawal

A Consumers

Withdrawal template

Figure 6.11.: Example for different enumerations on different hierarchy levels

1. Arabic enumeration (e.g., 1, 1.1, 2., 2.1.)
2. Latin enumeration (e.g., I, II.I, X., XX.I.)
3. alphabetical enumeration (e.g., A, A.B, C., C.D.)
4. enumeration with leading section sign (e.g., §1, § A, § X.I.)
5. combinations of 1-4 (e.g., A.1, § X.2.A)
6. no enumeration (e.g., “Terms and Conditions”)

For the first five styles, this task is pretty straight forward. For headings without any enumeration (which are frequently used on the highest and/or lowest level), we check three different properties. If a `<p>` element starts with a capital letter, does *not* end with a punctuation mark, and contains less than ten words, we consider it to represent a heading. Once all headings are identified, all remaining text elements are assigned to the closest heading that precedes them.

```xml
<doc sitename="Example Website" title="Terms and Conditions" source="https://www.example.com/tc.html" categories="" tags="">
  <main>
    <head>Our terms and conditions</head>
    <p>Trafilatura includes a command-line interface and can be conveniently used without writing code.</p>
    <p>For the very first steps please refer to this nice step-by-step introduction and for general instructions
    <head>Quickstart</head>
    <p></p>
    <p><span></span></p>
  </main>
</doc>
```

Listing 6.2: Example result of the content extraction with trafilatura, which serves as input for the paragraph segmentation

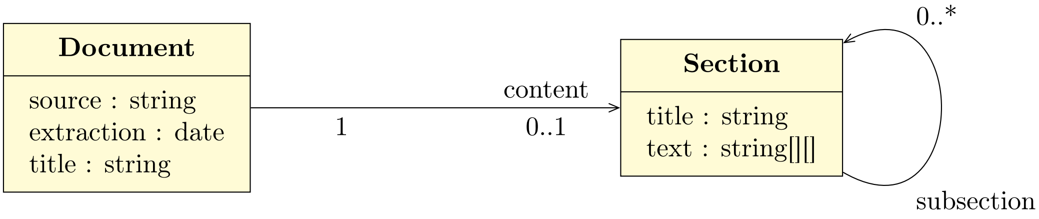


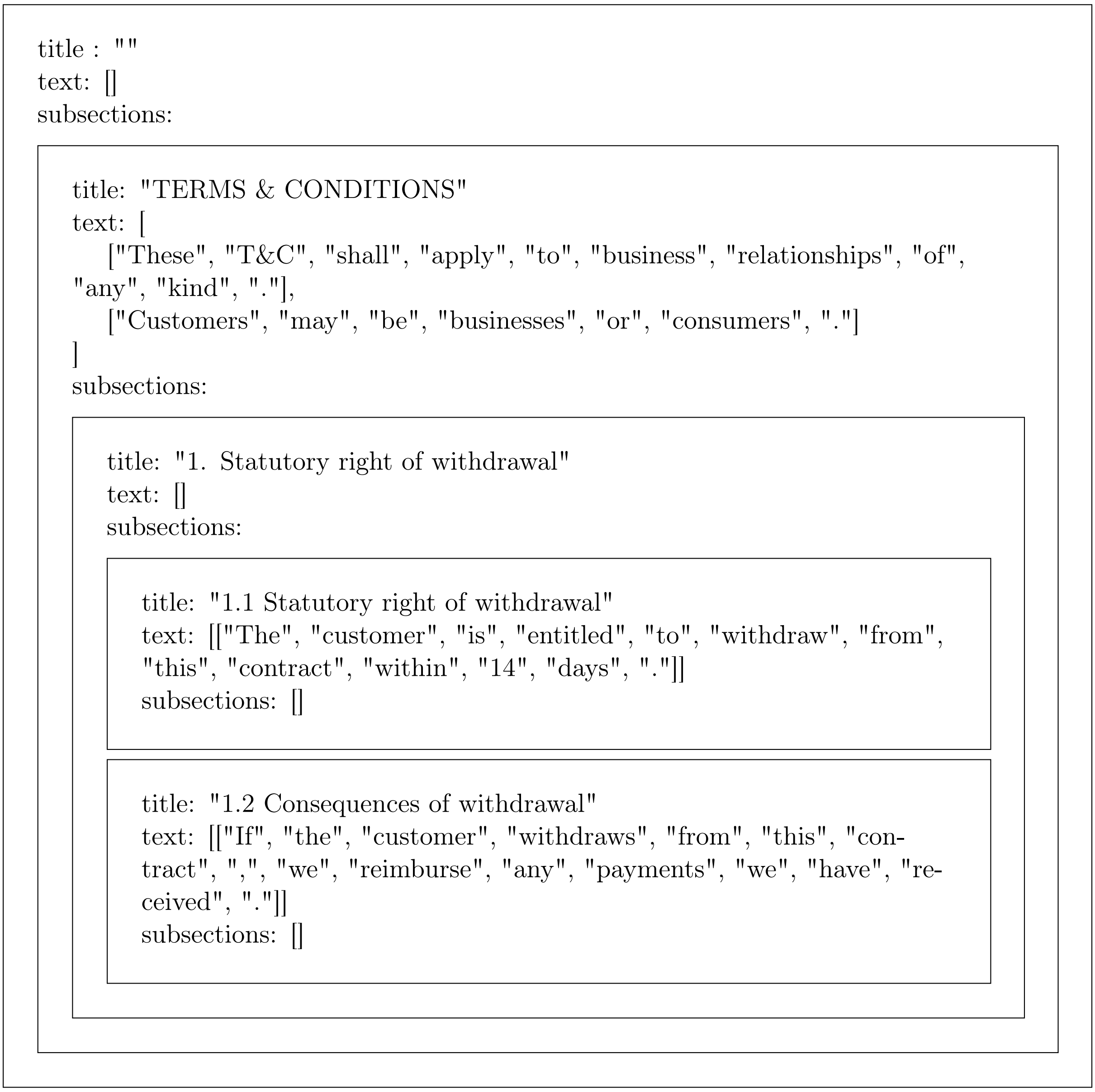
Figure 6.12.: Data model for mapping the paragraphs and their hierarchy

So far, we only retrieved the segmentation of the paragraphs, however, not yet their hierarchy. Whenever the enumeration contains an explicit hierarchy (1.2.3, A.1), we can easily map this hierarchy to the different paragraphs. This task is more difficult if the headings do not contain any enumeration or this enumeration does not explicitly reflect the hierarchy. Figure 6.11 shows an example where “A Consumers” is a subsection of “1.1. Withdrawal", however, does not explicitly reference its parent. “1. PURCHASE TERMS” is also a subsection of “TERMS & CONDITIONS”, without that being encoded in the enumeration. Moreover, both the highest level (“TERMS & CONDITIONS”) and the lowest level (“Withdrawal template”) of the hierarchy use not enumerated headings.

We can use the linear structure of the document and conventions for enumeration to extract the hierarchy in such cases. Whenever a new style of enumeration is introduced, which has not been seen before in a document, it has to be a subheading of the previous heading. Because if it were on the same level, it would have to use the same style, and if it was not seen before, it can also not be above the previous heading in the hierarchy. With this rule, we can easily identify that “1. PURCHASE TERMS” in Figure 6.11 is a subheading of “TERMS & CONDITIONS”. “1.1. Withdrawal” has the hierarchy explicitly encoded in the enumeration, which we can use. The newly introduced alphabetical enumeration (“A Consumers”) has to be a subsection of ‘1.1. Withdrawal”. However, in order to identify the lowest level correctly, we need an additional rule. While “Withdrawal template” uses the same enumeration style as “TERMS & CONDITIONS”, it is not on the same level of the hierarchy, on the contrary, it is on the different end of it. To capture such cases, for headings without enumeration, we apply the approach suggested by Manabe and Tajima (2015) and use the font-size to determine the position in the hierarchy.

6.2.3.2. Data Format

The hierarchical segmentation of the paragraphs is the last step in our pipeline in which we transform standard form contracts to machine-processable texts. Therefore, we now need a data format that can store the result of the pipeline until this point. Our data model consists of two elements shown in Figure 6.12: documents and sections. The document object holds meta-information that is not part of the extracted text. This includes the URL or name of the document the text was extracted from (source), the time and date the content was retrieved (extraction), and the title of the document or HTML page. If the document is not empty, it also contains a (single) section object.



title : ""
text: []
subsections:

title: "TERMS & CONDITIONS"
text: [
["These", "T&C", "shall", "apply", "to", "business", "relationships", "of", "any", "kind", "."],
["Customers", "may", "be", "businesses", "or", "consumers", "."]
]
subsections:

title: "1. Statutory right of withdrawal"
text: []
subsections:

title: "1.1 Statutory right of withdrawal"
text: [["The", "customer", "is", "entitled", "to", "withdraw", "from", "this", "contract", "within", "14", "days", "."]]
subsections: []

title: "1.2 Consequences of withdrawal"
text: [["If", "the", "customer", "withdraws", "from", "this", "contract", ",", "we", "reimburse", "any", "payments", "we", "have", "received", "."]]
subsections: []

Figure 6.13.: Graphical representation of a nested section structure

This section represents the first level in the content hierarchy of the document. All content of the standard form contract is a subsection of this first section. Each section can have a title, text, and subsections. The text is represented as an array of string arrays, representing the sentences and their tokens. Figure 6.13 shows a graphical representation of such a nested section and subsection structure. The top section will be empty in most cases, except for instances where there is text before the first heading appears.

We implemented this data model using JSON because it is supported by many programming languages and systems, a de-facto standard for data exchange through web APIs, and well suited to represent hierarchical structures. Listing 6.3 shows the corresponding JSON schema. We will later extend the schema to accommodate annotations and other additional information.

```
{
  "$id": "document.schema.json",
  "$schema": "http://json-schema.org/draft-07/schema#",
  "description": "Document representing a standard form contract",
  "title": "document",
  "type": "object",
  "properties": {
    "title": {
      "type": "string",
      "default": "",
      "examples": ["Terms and Conditions"]
    },
    "source": {
      "type": "string",
      "format": "uri",
      "default": "",
      "examples": ["https://www.example.com/tc.html"]
    },
    "extraction" : {
      "type": "string",
      "format": "date-time",
      "default": "",
      "examples": ["2018-11-13T20:20:39+00:00"]
    },
    "content" : { "$ref": "#/definitions/section" }
  },
  "definitions": {
    "section": {
      "type": "object",
      "properties": {
        "title": {
          "title": "title",
          "type": "string",
          "default": "",
          "examples": ["Terms and Conditions"]
        },
        "text": {
          "title": "text",
          "type": "array",
          "default": [],
          "items":{
            "type": "array",
            "default": [],
            "items":{
              "type": "string",
              "default": "",
              "examples": ["This is a sentence."]
            }
          }
        },
        "subsections" : {
          "type": "array",
          "default": [],
          "items": { "$ref": "#/definitions/section" }
        }
      }
    }
  }
}
```

Listing 6.3: JSON Schema for storing the processed standard form contracts.

6.3. Clause Topic Classification

One of the most important steps in the semantic analysis of contracts and their legal assessment is the classification of clause topics. Based on the clause topic, we decide which information we should extract from the clause and how we can assess its legality. In this section, we describe different approaches to the automatic classification of clause topics and subtopics, based on the taxonomy described in Section 4.4.1 and Appendix E.

The classification of the clause topics is not just a necessary step in our pipeline, but also in itself valuable information. As we will show in Chapter 10, the classification of clause topics can help consumer advocates to find and check relevant clauses faster. Moreover, the absence of a clause of a certain topic, e.g., online dispute resolution for consumer, can already constitute a violation of the law (Regulation (EU) No 524/2013 of the European Parliament and Council of the European Union (2013)), as we will explain in more detail in Chapter 7.

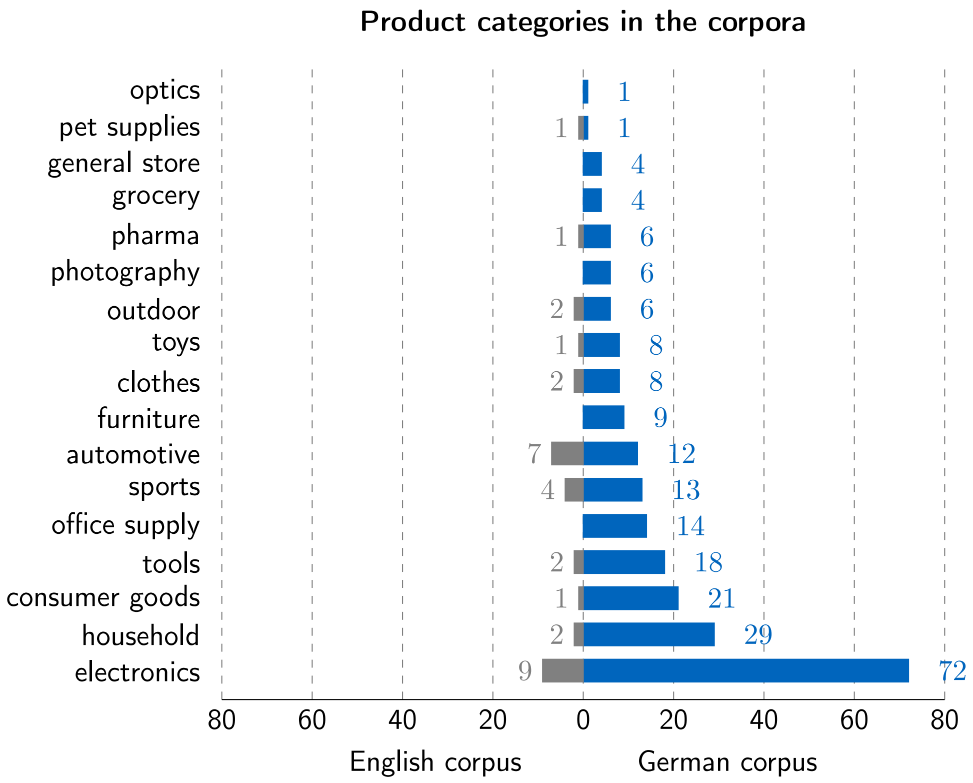


Figure 6.14.: Product categories offered by the online shops in the English and German corpus (more than one category per shop possible)

6.3.1. Corpus

Labeling each clause of T&C with their respective topic is a very time-consuming task. Therefore, we only used a subset of the corpus described in Section 6.1.1. For German, we randomly selected 142 T&C, and for English, we randomly selected 30 T&C. Each clause was manually copied into an Excel file, in which each row contains one clause. The file was used for the clause annotation (see Section 6.3.1.1). In addition to the text of the clause itself, each row contains a unique clause id, the id of the T&C the clause belongs to, (if existing) the title of the superordinate paragraph and (if existing) the title of the clause. The German corpus consists of 5,020 clauses, and the English corpus consists of 1,040 clauses. Figure 6.14 shows which product categories the shops sell, who's T&C are in our corpus.

The German corpus consists in total of 351,903 words, which is an average of 2,478 words per contract (see Table 6.9). The English corpus contains 55,392 words which equals to an average of 1,846 words per contract. The average number of clauses per contract is almost identical, with the average German contract containing 35 clauses and the average English containing 34 clauses, which means that the German clauses are, on average, significantly longer than the English ones. 5,013 clauses (or 99.9% of all clauses) in the German corpus have a paragraph

	contracts	clauses	words	∅ clauses/contract	∅ words/contract
German	142	5,020	351,903	35	2,478
English	30	1,040	1,846	34	1,846

Table 6.9.: Statistics on the German and English corpus

or clause title (or both), which we can possibly use for the topic classification. In the English corpus, that is the case for 989 clauses or 95.1%.

6.3.1.1. Topic Annotation

Each clause of both corpora was labeled with its topics according to the taxonomy described in Section 4.4.1. First, each clause was labeled by a student using only the classes from the first level (topics) of the taxonomy. Then, where applicable, classes from the second level of the taxonomy (subtopics) were added. In a second step, this process was repeated by the authors, i.e., each clause was again first labeled with classes from the first level and then, where applicable, with classes from the second level. A clause can be labeled with more than one topic or subtopic. A clause can also be assigned to a topic without necessarily having to be assigned to a subtopic of it. An example of such a clause from the corpus is "The warranty is subject to the relevant statutory provisions.", which is assigned the topic warranty, but none of its subtopics. Cases where the two annotators disagreed, were presented to (and finally decided by) consumer advocates from our collaboration partners. The inter-annotator agreement was relatively high at 87%, i.e., only 13% of all clauses had to be decided by consumer advocates. In this way, four people together spend more than 100 hours and generated more than 24,000 labels, which were consolidated to two corpora with 11,777 labels in total. The distribution of topics and subtopics is shown in Table 6.10.

The annotation also revealed some local specifics, e.g., almost none of the English T&C contained a model withdrawal form, the only two that did contain such a form were from companies based in Germany. On the other hand, clauses about loyalty schemes where almost non-existing in German T&C and far more popular in the English data set. It is worth reminding that our English data set was collected specifically from a United Kingdom (UK) perspective, i.e., the shops we collected the T&C from are either based in the UK or specifically targeted at the UK market. English T&C from other markets, like the USA or Australia, would look very different.

Label	German	English	Total
age	38	5	43
applicability	253	33	286
applicableLaw	137	23	160
arbitration	155	13	168
changes	13	12	25
codeOfConduct	55	1	56
conclusionOfContract	800	146	946

Table 6.10.: Distribution of topic and subtopic labels among the German and English corpus

Label	**German**	**English**	**Total**
conclusionOfContract:binding	328	39	367
conclusionOfContract:changeOfOrder	58	6	64
conclusionOfContract:definition	103	4	107
conclusionOfContract:restrictions	42	7	49
conclusionOfContract:steps	256	58	314
conclusionOfContract:withdrawal	95	20	115
delivery	839	164	1003
delivery:brokenPackaging	134	10	144
delivery:costs	247	57	304
delivery:customs	43	6	49
delivery:destination	96	16	112
delivery:methods	160	17	177
delivery:partial	32	5	37
delivery:time	143	41	184
description	86	30	116
disposal	51	16	67
intellectualProperty	45	24	69
language	124	11	135
liability	439	140	579
party	157	21	178
payment	898	112	1010
payment:fee	50	3	53
payment:late	48	1	49
payment:loyalty	7	22	29
payment:methods	435	53	488
payment:restraint	46	1	47
payment:vouchers	301	14	315
personalData	213	49	262
personalData:cookies	6	3	9
personalData:duration	8	1	9
personalData:information	48	12	60
personalData:reason	50	11	61
personalData:update	7	4	11
personalData:usage	57	16	73
placeOfJurisdiction	117	19	136
prices	158	56	214
prices:currency	17	13	30
prices:vat	119	24	143
retentionOfTitle	222	13	235
severability	42	12	54
textStorage	152	11	163
warranty	540	25	565
warranty:options	69	5	74

Table 6.10.: Distribution of topic and subtopic labels among the German and English corpus

Label	German	English	Total
warranty:period	155	10	165
withdrawal	484	202	686
withdrawal:compensation	94	27	121
withdrawal:effects	97	12	109
withdrawal:exclusion	100	27	127
withdrawal:form	131	37	168
withdrawal:model	41	2	43
withdrawal:period	126	40	166
withdrawal:shippingCosts	118	43	161
withdrawal:shippingMethod	74	13	87
Total lvl 1	6018	1138	7156
Total lvl 2	3941	680	4621

Table 6.10.: Distribution of topic and subtopic labels among the German and English corpus

6.3.2. Evaluation Approach

The task of automatically classifying the topics of a clause is a multi-label multi-class classification problem. In order to compare the different approaches we will describe in the following section, we will use four evaluation metrics. Since evaluation metrics for multi-class classification are slightly less intuitive than the standard versions, we will shortly introduce the metrics we are going to use.

Precision
Precision is a measurement for how many of the clauses that were assigned to a class actually belong to this class. In comparison to single-label classification, there are two different ways to calculate the precision of multi-label classification. Let us assume we have classified five true positives and one false positive for the topic withdrawal and one true positive and no false positive for the topic age. The precision for the class withdrawal would hence be $5/6 \approx 0.833$, and for age, it would be $1/1 = 1$. For the classification as a whole, we now have two ways of calculating the precision. The macro-average approach computes the overall precision as the average of the precision of the individual classes. In our example that would be $(0.833 + 1)/2 \approx 0.917$. The micro-average approach, on the other hand, computes the overall precision based on each individual classification independent from the classes. For our example that would mean the micro-average precision is $(5 + 1)/(5 + 1) + (1 + 0) \approx 0.857$. Especially on imbalanced data-sets where some classes are much more frequent than others (like in our case, for example, the class delivery occurs 1,003 times, while changes occurs only 25 times), the values of the macro-average and the micro-average approach can differ significantly. For our evaluation, we will use micro-average precision values because we believe they are better able to measure how useful a system would be in application. If one topic only occurs very sparsely, it would be less harmful if it would not be classified correctly.

Recall
Recall is a measurement for how many clauses that actually belong to a class were also assigned

to this class. Keeping in mind that we want to develop classification approaches that will mainly be used by domain experts, in the trade-off between precision and recall, recall should be prioritized higher. If a clause is assigned a topic it does not belong to, an expert using the system will notice that very quickly. However, if the system misses, e.g., a clause on withdrawal, it might be the case that the expert never sees the clause. As for precision, we also use the micro-average to calculate recall.

F1-score
The F1-score is the harmonic mean of precision and recall and hence represents a balance between the two values. It can be calculated based on the micro-average or macro-average of precision and recall. In our evaluation, we will use the micro-average of both values.

Accuracy
When it comes to multi-label classification, there are different evaluation metrics that might be denoted with "accuracy", because, other than single-label classification, a multi-label classification can be not only binary wrong or right but also partially right or wrong. One approach is to simply ignore the existence of partially correct classifications and calculate the exact match ratio, i.e., the share of all classifications that are completely correct. That means if a clause contains three different topics, and two out of three were classified correctly, the accuracy would be 0. (Sorower, 2010) A more commonly used approach is to calculate the ratio of correctly predicted labels with respect to the actual and predicted labels. If a clause, for example, belongs to the topics $T = withdrawal, warranty, payment$ and was predicted with $P = withdrawal, warranty, liability$, the accuracy would than be $|T \cap P|/|T \cup P| = 2/4 = 0.5$. We will use this notion of accuracy for our evaluation, which is sometimes also referred to as Hamming score (Godbole and Sarawagi, 2004) or Jaccard similarity coefficient (Niwattanakul et al., 2013).

6.3.3. Rule-based Topic Classification

As a baseline, we first developed an algorithm for the rule-based classification of clause topics. We decided to use a simple keyword-based approach. For each clause topic, we asked the consumer advocates to provide a list of keywords that are distinctive for the topic. The list can contain independent keywords (OR), keywords that should appear together (AND), and keywords that should not appear (together) (NOT).

We pre-processed the clauses using SoMaJo (see Section 6.2.2) to split the clauses into sentences and the sentences into tokens. Afterwards, we lemmatized all tokens using the Stanford Lemmatizer (Manning et al., 2014) for English and the Mate tools Lemmatizer (Björkelund et al., 2010) for German before applying the rules.

In German, we noticed very quickly that lemmatization (but also stemming) face big challenges, especially in the legal domain, when it comes to compound nouns, i.e., nouns that are combined to create new nouns, like "Vertragspartner" (contractual partner) is a combination of "Vertrag" (contract) and "Partner" (partner). Compound nouns can be inflected internally ("Vertrag**s**partner"), and splitting them into their constituents is not trivial. A "Druckerzeugnis"

		Input not empty				All clauses			
Language	**Input**	**A**	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**
German	Paragraph Title	0.482	0.705	0.603	0.650	0.481	0.705	0.602	0.649
	Clause Title	0.645	0.917	0.685	0.784	0.053	0.917	0.053	0.101
	Both Titles	0.488	0.706	0.612	0.656	0.487	0.706	0.611	0.655
	Clause Text	0.643	0.770	0.795	0.782	0.643	0.770	0.795	0.782
	Text and Titles	0.630	0.725	0.828	0.773	0.630	0.725	0.828	0.773
English	Paragraph Title	0.585	0.763	0.716	0.739	0.561	0.763	0.680	0.719
	Clause Title	0.886	0.969	0.912	0.939	0.054	0.969	0.054	0.103
	Both Titles	0.591	0.764	0.722	0.743	0.566	0.764	0.686	0.723
	Clause Text	0.481	0.600	0.708	0.649	0.481	0.600	0.708	0.649
	Text and Titles	0.519	0.587	0.818	0.684	0.514	0.588	0.804	0.679

Table 6.11.: Comparison of the rule-base clause topic classification results on different inputs in German and English only clauses where the input is not empty and all clauses (A = accuracy, P = precision, R = recall, F1 = F1-score)

(printed matter) could, for example, lexically speaking either be a “Druck-Erzeugnis” (print - matter) or a “Drucker-Zeugnis” (printer - certificate).

While there are existing approaches on how to automatically split compound words into their respective parts (e.g., by Baroni et al. (2002), Koehn and Knight (2003), Daiber et al. (2015), Sugisaki and Tuggener (2018), and Weller-Di Marco (2017)), the problem is far from being trivial and is not yet addressed by the implementation.

For both languages, we evaluated the rules for the topic (i.e., first level) classification on five different inputs:

- only paragraph titles
- only clause titles
- both titles
- clause text
- both titles and clause text

6.3.3.1. Topics

An overview of the classification performance for the topic classification is shown in Table 6.11. More detailed results, broken down to the individual topics, can be found in Appendix F.1.1.

For both languages, the best results (with regard to F1-score) were achieved when using only the clause title, i.e., the direct heading of a clause. In cases where such a title is present, we achieved an F1-score of 0.784 for German and 0.939 for English. However, since only a small proportion of all clauses have such a title, when measured on the complete data set, the F1-scores drop to 0.101 and 0.103. We identified two reasons for the performance gap between German and

English: It seems that in the German data set, there are more generic clause titles like “Hinweis” (hint) or “Allgemeines” (general remarks). Second, it seems to be more common in the German data set that clause titles are directly linked to the paragraph title and can not be understood without them. This is, for example, the case when the clause title would be “Frist” (period), which could be the withdrawal period or a payment period, depending on whether the paragraph (title) is about withdrawal or payment.

The paragraph titles as input perform considerably worse than the clause titles when considering only cases where the respective input is present (0.650 for German and 0.739 for English). This is not surprising, given that the same paragraph title applies to multiple clauses, and it is unlikely that they all have the same topic. Often paragraphs and therefore also their titles combine multiple topics, like “Payment and Delivery”, which generates many false positives since all clauses will be assigned both topics. However, since almost all clauses have a paragraph title, measured on the whole data-set, the performance is better than for the clause titles, with an F1-score of 0.649 on the German corpus and 0.719 on the English corpus. By concatenating the paragraph title with the clause title, the results can be improved slightly.

In summary, we can say that clause titles work fairly well (especially in English) where they exist. However, using titles as input has certain limitations, which can not be overcome, independent from the rules that are applied and even if more sophisticated technology is applied.

On the German corpus, we achieved the best results when applying the rules to just the clause text, which resulted in an F1-score of 0.782. For many individual topics, we were able to achieve F1-scores close to 1.0, despite very simple rules. For detecting clauses about the contract language, for example, we used only the keywords “Vertragssprache” and the combination of “Vertrag” and “Sprache” (as separate words) and achieved an F1-score of 0.988. For finding clauses about codices of conduct, we used the two keywords “Kodex” and “Zertifikat” and achieved an F1-score of 0.991. Arbitration clauses can be detected with an F1-score of 0.997 by just using the keywords “Schlichtung” and “Streitbeilegung”. In English, only one class (retention of title) achieved an F1-score of above 0.9. One of the reasons for this might be the frequent use of compound nouns in German T&C. With keywords like “Vertrassprache” or “Streitbeilegung”, false positives are very unlikely, while the combination of multiple independent words is potentially more likely to trigger false positives. The presence of “Vertragsparnter” (contractual partner), for example, almost certainly indicates a clause about the contracting party, the mention of “Vertrag” (contract) and “Partner” (partner) is much more likely to happen in different contexts (“Bei Problemen mit Ihrem Vertrag können Sie sich an einen unserer Partner wenden.” / “If you have problems with your contract, you can contact one of our partners.”). Combining the text of the clauses with their titles improved the recall on both corpora significantly, however, it also decreases the precision.

Based on our findings, we postulated the hypothesis that the classification performance of our rules is correlated to the length of the input, i.e., the number of words of the input text. In order to test this hypothesis, we calculated for each topic how many words the clause texts in this topic consist of on average. A first analysis based on a scatter plot (see Figure 6.15) supported the hypothesis for the German corpus, however not for the English corpus. This first impression was confirmed by further statistical analysis. We calculated the Spearman’s rank-order correlation and found a negative correlation between the average number of words and the

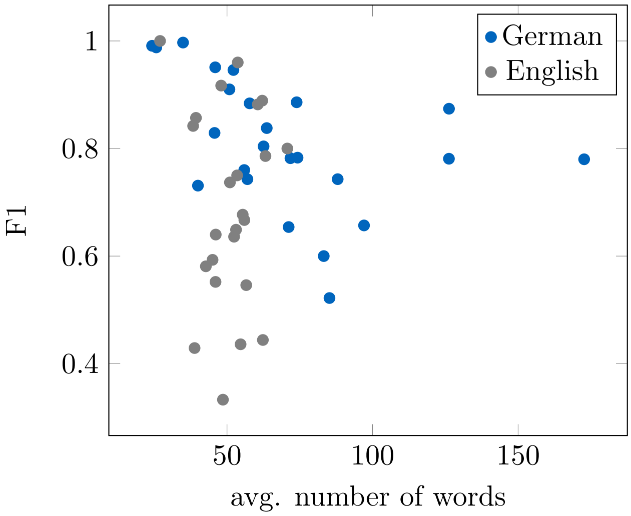


Figure 6.15.: Scatter plot showing the average number of words of clause texts per topic in relation to the achieved F1-score with the clause text as input

F1-score ($r_s = -0.57$) in the German corpus, which is statistically significant ($p < 0.005$). On the English corpus, we did not find any statistically significant correlation.

For the German corpus, that means that the rule-based topic classification performs best on short clauses, and the performance decreases, the longer the text gets. One possible explanation for this is that the likelihood of false positives increases with the number of words. Another possible explanation is that longer clauses represent more vague topics, which are therefore more difficult to identify. The fact that we do not see the same effects in English might simply be because the average length between the different topics differs only slightly, as can be seen in Figure 6.15.

6.3.3.2. Subtopics

For the subtopics (i.e., second level) classification, we decided to only use the clause texts because the paragraph headings and clause headings are not fine-grained enough to enable the classification of subtopics. The rules drafted for this task are independent of the first level classification. For the subtopic "cookies", which belongs to the topic "personal data", we use, for example, simply the keyword "cookie", without checking first whether a clause belongs to the topic "personal data". The results we achieved with this approach are shown in Table 6.12. More detailed results, broken down to the individual subtopics, can be found in Appendix F.2.1.

The results show clearly that the rules are not working well for the subtopic classification. Although an in-depth analysis shows that a few classes (e.g., "cookies" with $F1 = 0.857$ in English or "vouchers" with $F1 = 0.961$ in German) do perform well, the majority of all classes achieves F1-score below 0.5. There are multiple possible reasons for the bad performance. First, by their very nature, subtopics are more fine-grained than topics; therefore, it is not surprising they are more difficult to detect, especially with relatively simple rules. The boundaries between

Language	Input	A	P	R	F1
German	Clause Text	0.471	0.743	0.562	0.640
English	Clause Text	0.277	0.387	0.494	0.434

Table 6.12.: Comparison of the rule-based clause subtopic classification results in German and English (A = accuracy, P = precision, R = recall, F1 = F1-score)

different subtopics of the same topics can, even for humans, be sometimes not as clear as the boundaries between two main topics. Lastly, with 37 subtopics, there are simply more classes to classify than for the topic labels. We can see again that the rule-based classification works much better in German than in English. And again, one of the reasons for that is probably the use of compound nouns, which frequently occur in the rules for the subtopics, like "Widerrufsfrist" (withdrawal period) or "Rücksendekosten" (return costs).

A natural idea to improve classification performance would be to utilize the hierarchy of the taxonomy by first classifying the topics and then only applying subtopics that fit the main topic. However, there are some flaws to this approach. The best performing rule-based approach for topic classification in English (using both titles as input) has a much higher precision than recall, i.e., it will miss out on many labels that are actually correct. This error would be propagated to the subtopic classification, which would not be able to classify the correct subtopic, although it might have been if it would have been applied independently from the first level classifier. Although there would be other possible approaches (e.g., applying both classifiers independently and then try to merge their results based on a confidence value derived from the classification performance on individual classes), we decided to not pursue them because the results of both classifiers were simply not good enough and we wanted to focus on more promising approaches, which we are going to introduce in the next sections.

6.3.4. Logistic Regression

As the first stochastic approach, we trained a logistic regression classifier. This and the following approaches were implemented using Scikit-learn (Pedregosa et al., 2011). Unlike for the detection of T&C, we did not use a sparse count vector representation of the texts as input, but a Tf–idf vector representation, which does not assign each word its count, but its term frequency – inverse document frequency. In this way, words that frequently appear in all clauses, and are therefore less specific for any particular clause or class, get less weight. Before transforming the clauses into these vectors, we again remove the stopwords using "Stopwords ISO".

Since logistic regression in itself does not support multi-label classification (see also Section 2.1.1), we use a "one-vs-the-rest" approach. Instead of training one classifier, we train one classifier for each class, which performs a binary classification against all remaining classes (hence "one-vs-the-rest"), and combine all results to decide which labels are predicted for a given input.

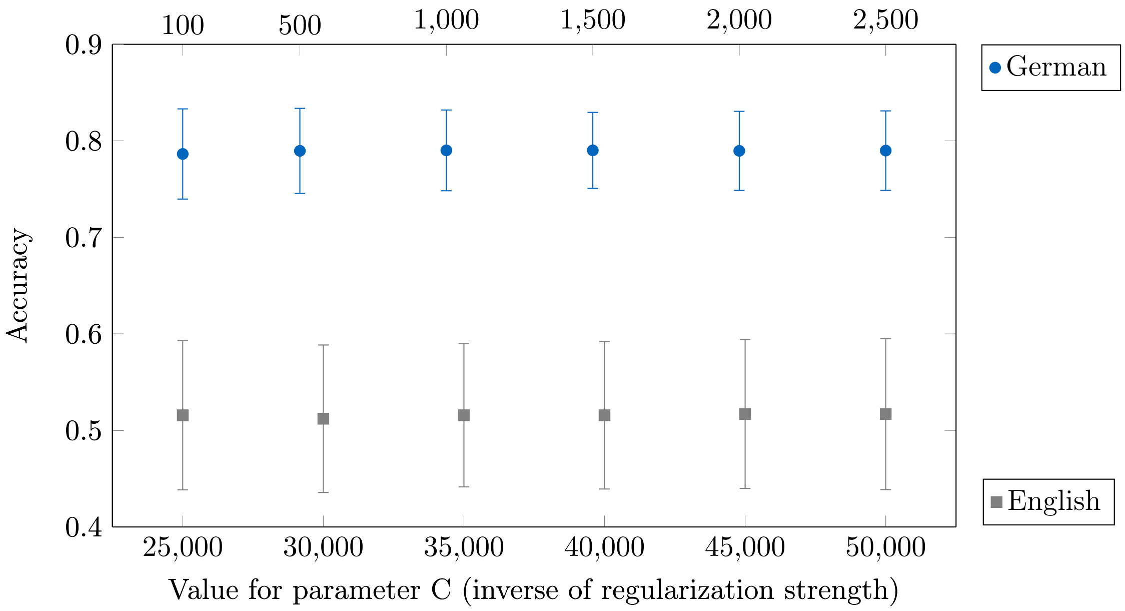


Figure 6.16.: Results of the grid search with 10-fold cross-validation to find the parameter for the regularization strength of the logistic regression for topic classification

6.3.4.1. Topics

We split the corpus into a training (80%) and a test (20%) set, using scikit-multilearn (Szymański and Kajdanowicz, 2017) to make sure the representation of labels is balanced between the training and the test set and reflect the original distribution. We used the training data set to perform a grid search with a 10-fold cross-validation to find the parameter for the regularization strength. We performed multiple iterations on both languages to narrow down the search space. The results of the final runs are shown in Figure 6.16. In German, we achieved the best results with $C = 1,000$ and in English with $C = 45,000$. The values are rather high for both languages, but especially for the smaller English data-set. Since C is the inverse of the regulator $(1/\lambda)$, a high value for C means a low value for λ and hence poses the risk of overfitting.

We evaluated the approach on the test set. For simplicity reasons, for this and the following approaches, we will only consider input configurations that can be applied to all clauses, i.e., the clause text and the combination of the clause text and titles. The results of the evaluation are shown in Table 6.13. More detailed results, broken down to the individual topics, can be found in Appendix F.1.2.

For both languages, the best performance was achieved when combining the titles and the text. While the performance in German almost did not change (+0.01), the performance increased slightly more in English (+0.03). The Logistic Regression classifier performed much better than the rule-based approach with regard to both languages, both input types, and all metrics. The F1-score went up by 0.1 in German and 0.12 in English. Overall, the approach performed well in both languages.

Language	Input	A	P	R	F1
German	Clause Text	0.76	0.95	0.79	0.86
	Text and Titles	0.77	0.95	0.80	0.87
English	Clause Text	0.67	0.85	0.70	0.77
	Text and Titles	0.71	0.88	0.73	0.80

Table 6.13.: Comparison of the Logistic Regression clause topic classification results on different inputs in German and English (A = accuracy, P = precision, R = recall, F1 = F1-score)

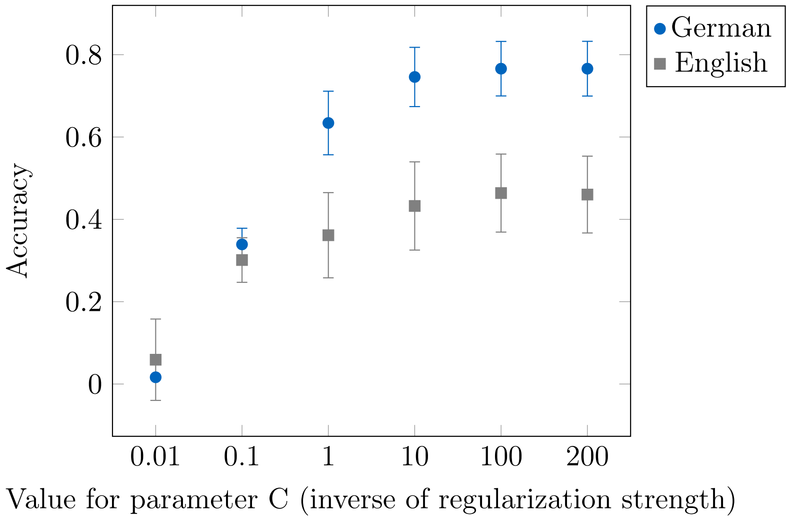


Figure 6.17.: Results of the grid search with 10-fold cross-validation to find the parameter for the regularization strength of the logistic regression for subtopic classification

6.3.4.2. Subtopics

We performed the same procedure for the classification of subtopics, just that we used the subtopic labels for the training and the evaluation this time, instead of the topic labels: We performed a stratified split into training and test data set and then performed multiple iterations of a grid search with a stratified 10-fold cross-validation. For the classification of the subtopics, we found that $C = 100$ performed best in both languages (see Figure 6.17), which is significantly lower than in the case of the topic classification.

The evaluation of the classification on the test set is shown in Table 6.14, more detailed results can be found in Appendix F.2.2.

In German, the classification of the subtopics performed almost as well as the classification of topics, while there is a more significant difference in English. Given the fact that the number of classes increased from 23 to 37 and the amount of training data available decreased by almost half at the same time, the results in German are surprisingly good, and the approach performed significantly better in both languages than the rule-based approach.

Language	Input	A	P	R	F1
German	Clause Text	0.74	0.91	0.77	0.83
	Text and Titles	0.75	0.91	0.78	0.84
English	Clause Text	0.55	0.80	0.58	0.67
	Text and Titles	0.54	0.80	0.59	0.68

Table 6.14.: Comparison of the Logistic Regression clause subtopic classification results on different inputs in German and English (A = accuracy, P = precision, R = recall, F1 = F1-score)

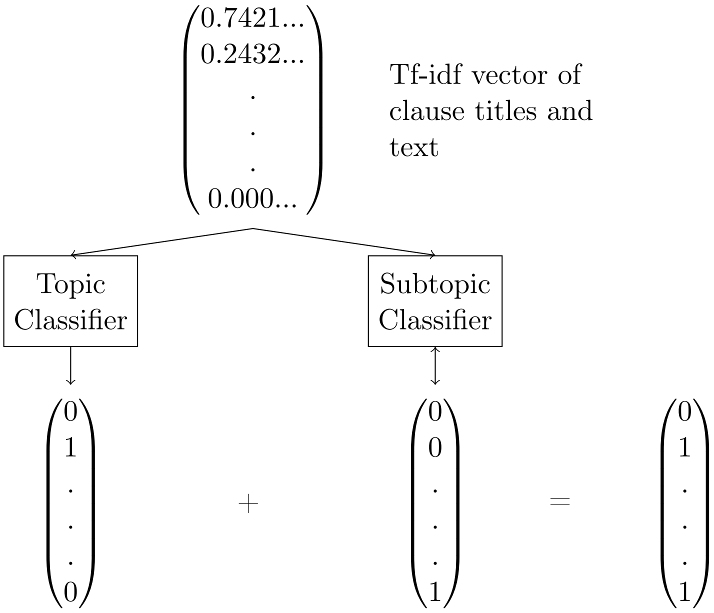


Figure 6.18.: Parallel classification approach

As for the rule-based approach, the classification of topics and subtopics happens independently: We apply the topic classifier and apply the subtopic classifier in parallel and independently from each other and ignore the hierarchical structure of the taxonomy. Each classifier will produce a sparse vector with the same dimensionality (see Figure 6.18), in which each component represents a topic or subtopic. However, the topic classifier will generate only vectors encoding topics, and the subtopic classifier will only generate vectors encoding subtopics. In the end, we can just add both vectors up to get the final classification containing topics and subtopics. This approach leads to the same problems described before; most notably, we can again end up with inconsistencies between topics and subtopics, and we do not utilize the hierarchy of the taxonomy.

Dealing with hierarchical data in machine and deep learning is still an open problem (see, e.g., Zhou et al. (2020), Rojas et al. (2020), and Aljedani et al. (2020) for recent work on the topic). In many cases, simply flattening the hierarchy, and hence losing the information encoded in it, is still the standard approach (Wang and Lu, 2010). So instead of training one classifier for each level of the hierarchy, we train one classifier on both levels, which then predicts both topics and subtopics (see Figure 6.19).

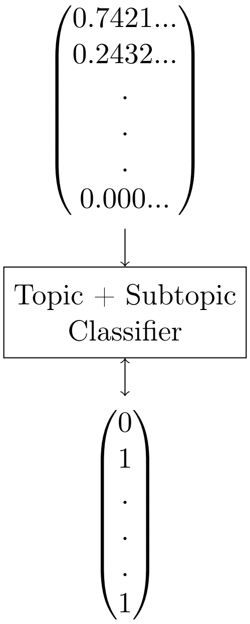


Figure 6.19.: Flat classification approach

Flattening the hierarchy to a certain degree always means losing information. For some classifiers, that is less problematic because they can re-learn the relation between labels from the data. For logistic regression, however, this is not the case. Since we apply a “one-vs-the-rest” approach, i.e., train an individual classifier for each label, the classifier does not learn anything about the relationship between labels. Hence, we do not gain anything compared to the parallel classification approach. This is also visible in the results in Table 6.15. When we apply two independent classifiers for topics and subtopics and combine their results, the classification performance overall labels (topics and subtopics) differs only slightly from the performance of a classifier that was trained on both types of labels at the same time.

Another approach could be a sequential classification process, which consists of two stages. In the first stage, we classify the topics for each clause as described in Section 6.3.4.1. In the second stage, we have one classifier for every topic (that has subtopics), which is only trained on clauses from this topic to distinguish between the subtopics of this particular topic. The result of the topic classification determines which subtopic classifiers will be involved. Every classifier generates a sparse (mutual exclusive) vector for its specific range of subtopics. In the end, we sum up all vectors from the second stage and the topic vector from the first stage to retrieve the final classification (see Figure 6.20).

With this approach, we can avoid inconsistencies, i.e., that a clause is assigned a subtopic but not the associated topic. While this approach does not influence the first level classification performance, it could, in theory, increase accuracy on the second level because the classifier only has to learn to distinguish between the subtopics of one topic instead of all subtopics. However,

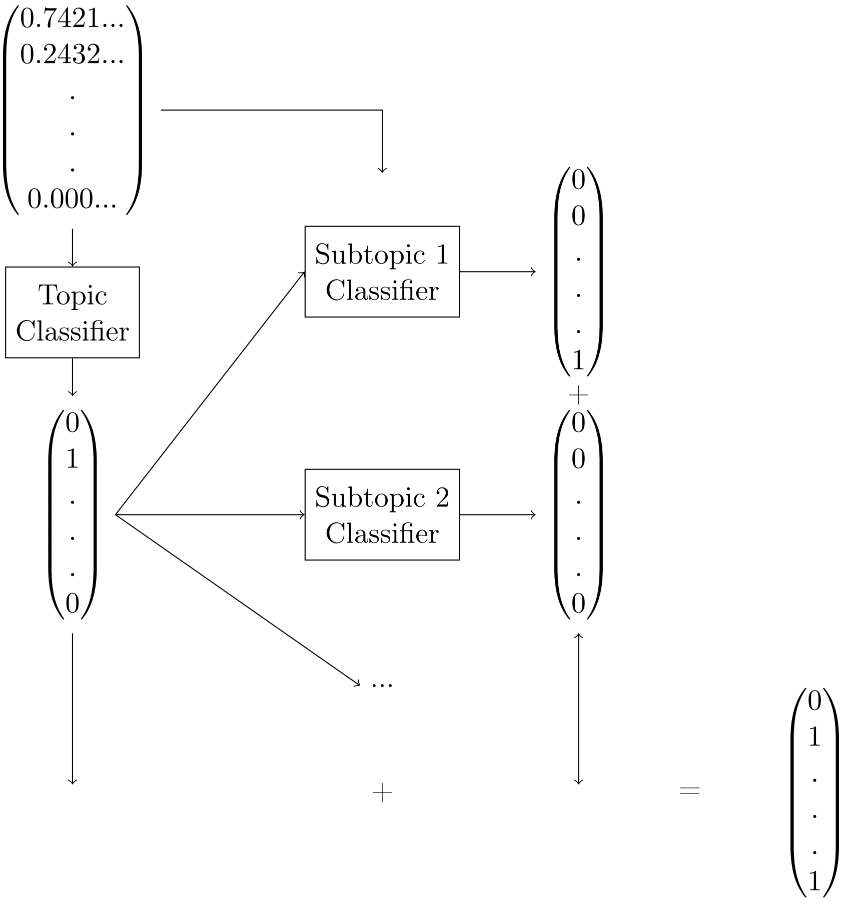


Figure 6.20.: Sequential classification approach

this approach certainly would decrease recall because classification errors in the first stage are propagated to the second stage: When the topic of a clause was misclassified, the subtopics of the clause can never be classified correctly. The propagation of error could even harm second level accuracy. If, e.g., a withdrawal clause is wrongly classified as a warranty clause, the classifier for warranty clauses will be applied to it on the second level. Since this classifier never saw a withdrawal clause before, there is a possibility that it will be misclassified. And indeed, an evaluation on our data set showed that the approach could increase the accuracy but decreased recall on both languages (see Table 6.15).

Lastly, we developed an approach for a joined classification by two classifiers. For this, we use the two original independent classifiers for topics and subtopics in parallel. However, instead of joining their output to the final classification, we developed a module we call balancer that uses both outputs and their confidence in their classification, which, in the case of logistic regression, is expressed as the distance of a data point to the hyperplane that represents the decision boundary. If a point is very close, the classifier is very sure it belongs to a class. If it is very far away, the classifier is very sure it does not belong to the class. Unlike in a classical voting approach, the two classifiers do not predict the same target, however, their prediction targets

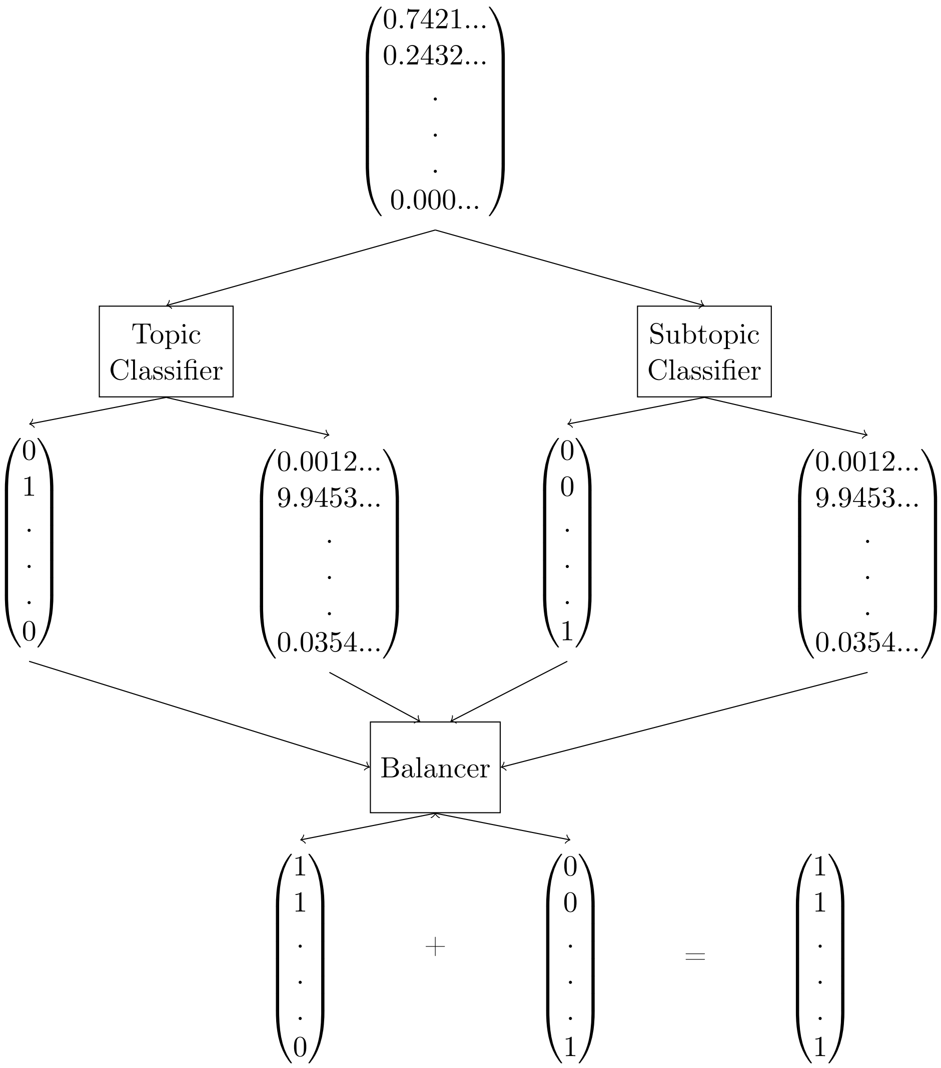


Figure 6.21.: Joined classification approach

Language	Approach	A	P	R	F1
German	Parallel	0.64	0.94	0.77	0.85
	Flat	0.65	0.94	0.79	0.86
	Sequential	0.61	0.92	0.76	0.83
	Joined	0.65	0.94	0.79	0.86
English	Parallel	0.37	0.88	0.61	0.72
	Flat	0.39	0.87	0.64	0.74
	Sequential	0.36	0.88	0.52	0.65
	Joined	0.38	0.88	0.62	0.73

Table 6.15.: Comparison of the Logistic Regression clause subtopic classification results on different inputs in German and English (A = accuracy, P = precision, R = recall, F1 = F1-score)

are related, and we can use this fact. If, for example, the subtopic classifier is very certain that a clause belongs to the subclass “restrictions”, which belongs to the topic “conclusion of contract”, however, the topic classifier was just shy of predicting this topic, the balancer could add the label “conclusion of contract” to the topic vector. And the other way around, if the topic classifier is very sure that a clause (only) belongs to “conclusion of contract” but the subtopic classifier has added a subtopic from a different topic with a relatively low confidence, the balancer could remove that from the subtopic vector. In the end, both vectors are summed up again to retrieve the final classification (see Figure 6.21).

More technically speaking, the balancer iterates over the subtopic vector and checks for each subtopic with the value 1 whether the value for the corresponding topic in the topic vector is also 1. If that is not the case, the balancer looks up the confidence (i.e., distance) for the subtopic and the respective topic. If $distance(topic) < threshold_a$, then the balancer will add the topic that belongs to the subtopic to the vector of topics. If $distance(subtopic) > threshold_b$ (with $threshold_a < threshold_b$), then the balancer instead removes the subtopic from the vector of subtopics. Please note that we can not draw any conclusions from the opposite situation. When a topic label was assigned but none of its subtopics, because according to our label procedure, it is not mandatory to assign any subtopics because they do not represent an exhaustive division of their parent topic. The appropriate thresholds can be learned on the training data. On our data, we found $threshold_a = 0.005$ and $threshold_a = 0.01$ to work well in both languages.

Tested on the data, the joined approach worked as expected and improved accuracy, precision, recall, and F1-score, compared to the parallel approach, in both languages. However, we found that only approximately 1% of all clauses in both languages were labeled inconsistently by the parallel approach, which is not surprising given the high precision of the approach. However, that means that the balancer, on our data, only gets active in 1% of all cases. This means that, although we were able to fix almost all inconsistencies correctly, the accuracy increased by only 0.003 in both languages and is therefore almost not visible in Table 6.15 due to rounding. However, we have shown that this approach, which utilizes the hierarchy of the taxonomy, is able to successfully correct cases of inconsistency, which might have an even bigger impact on other data sets.

6.3.5. Random Forest

Logistic regression is computationally efficient and generalizes well, and is, therefore, a good baseline. However, in our specific use-case, its inability for "real" multi-label classification is a considerable drawback. As mentioned before, logistic regression performs a binary classification and therefore does not "naturally" support multi-label classification. Decision trees, on the other hand, do inherently support multi-label classification and also are inherently explainable. However, they are not as efficient as logistic regression and more prone to overfitting (see also Section 2.1.2). We conducted a series of experiments to evaluate how decision trees perform on our task, especially with regard to jointly classifying topics and subtopics, because, other than logistic regression, Decision Trees are able to learn about correlations between labels. Instead of training just one decision tree, we use a random forest approach, where multiple independent randomized decision trees are trained, and a majority vote is used for classification.

Figure 6.22 shows the root of a decision tree which was built for topic classification. At each node, one dimension of the input Tf-idf-vector is checked, which equals the word that is printed in the first row. If the Tf-idf value is smaller or equal, then the value given there, the tree continues in the left branch. Otherwise, it continues in the right branch. In this way, in each step, the number of clauses that belong to one branch is reduced. The value array at the bottom represents how many clauses in the branch belong to a given topic (in alphabetical order). At the root node, we have 828 clauses. Of these 828, four belong to the topic "age", and 824 do not belong to this topic; hence the first value is `[824, 4]`, 26 belong to the topic applicability, and so on.

If the Tf-idf-value for "delivery" is lower than 0.11, i.e., if the word does not frequently occur in the clause, the tree branches to the left, otherwise to the right. In the left branch, 28 out of 689 clauses (or 4%) belong to the class "delivery". In the right branch, 103 out of 139 clauses (or 74%) belong to this class. However, 16% also still belong to the class "withdrawal"; therefore, the next test in this branch is for the word "returns". If the word occurs frequently, only one out of 14 samples still belongs to the class "delivery" and all others to the class "withdrawal". In this way, step by step, the algorithm goes through the tree until it reaches a leave and suggests a classification.

6.3.5.1. Topics

For the classification of clause topics with the random forest classifier, we first again performed a grid search with stratified 10-fold cross-validation to find the best performing values for the hyper-parameters: number of estimators (i.e., the numbers of trees), the maximum depth of the trees, the minimum number of samples per internal node that is needed for a split, and the minimum number of samples per leaf. As usual, we performed several iterations to narrow down the search space before, in the final iteration, we found the following values to perform best. In German: number of estimators = 2,000, maximum depth = ∞, samples per node = 2, samples per leaf = 1 and in English: number of estimators = 1,000, maximum depth = 100, samples per node = 2, samples per leaf = 1.

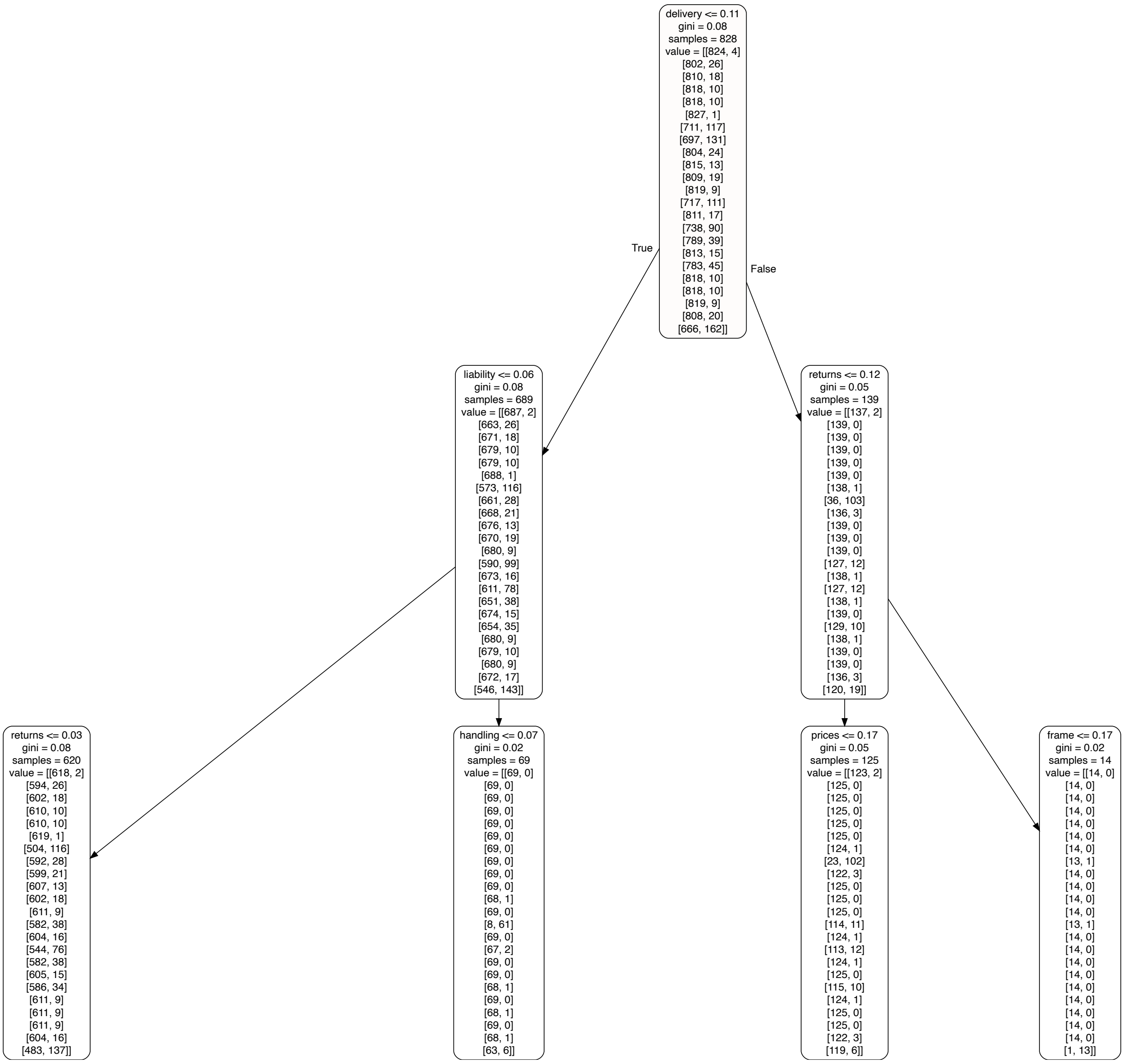


Figure 6.22.: Root of a decision tree for topic classification

Language	**A**	**P**	**R**	**F1**
German	0.73	0.97	0.72	0.83
English	0.57	0.88	0.58	0.70

Table 6.16.: Comparison of the Random Forest clause topic classification results in German and English (A = accuracy, P = precision, R = recall, F1 = F1-score)

Language	**A**	**P**	**R**	**F1**
German	0.68	0.91	0.67	0.77
English	0.44	0.85	0.43	0.57

Table 6.17.: Comparison of the Random Forest clause subtopic classification results in German and English (A = accuracy, P = precision, R = recall, F1 = F1-score)

The results of the evaluation on the test data are shown in Table 6.16. We used both titles and the clause text as input, a short evaluation with only the clause text showed that performance decreases significantly when removing the titles from the input.

The results show that the random forest classifier performs even better than the logistic regression classifier when it comes to precision, however, much worse when it comes to recall. That indicates that, despite an expensive grid search to optimize the parameters, the classifier is overfitting. A closer inspection of the trees generated confirms our earlier hypothesis that the significantly better performance on the German corpus is, at least partially, caused by the extensive usage of compound nouns. In the first layers of the German trees, almost all of the tested features relate to compound nouns, like "Widerrufsrecht" (right of withdrawal), "Vertragssschluss" (conclusion of contract), and "Versandkosten" (shipping costs).

6.3.5.2. Subtopics

The evaluation of the subtopic classification performance, which is shown in Table 6.17, shows a very similar result. While the random forest classifier performs better with regard to precision (+0.05 in English), it performs much worse with regard to recall (-0.11 in German and -0.16 in English), and therefore also with regard to F1-score and accuracy.

Finally, we also tested the performance for the joined classification of topics and subtopics. The results are shown in Table 6.18. Compared to just classifying subtopics, we see an increase in precision and F1-score, but a decrease in accuracy and recall, similar to the logistic regression classifier. The change is also of a similar magnitude as before, which suggests that the classifier is not much influenced by the correlation between the labels. Since we saw almost no inconsistency between topics and subtopics when they were classified independently, this result is not surprising.

Language	A	P	R	F1
German	0.57	0.94	0.69	0.80
English	0.26	0.87	0.51	0.58

Table 6.18.: Comparison of the Random Forest joint clause topic and subtopic classification results in German and English (A = accuracy, P = precision, R = recall, F1 = F1-score)

6.3.6. Mutlilayer Perceptron

So far, we achieved the best results with logistic regression. However, especially in English, the results have not yet been satisfying. In this and the following sections, we evaluate more advanced deep learning technologies in order to find out whether we can improve the results with these technologies. Training neural networks is computationally much more expensive than, e.g., logistic regression and at the same time depends on more hyper-parameters. In order to manage the increasing complexity, we changed our approach to hyper-parameter tuning by reducing the 10-fold cross-validation to a 5-fold cross-validation. Additionally, we fixed hyper-parameters that always performed best or almost best, independent from other parameters, as early as possible in order to reduce the search space and converge faster to a local optimum.

6.3.6.1. Embeddings

In addition to the new classifiers in the following sections, we will also introduce new inputs by using different kinds of word embeddings instead of Tf-idf vectors. We will continue to use the clause text, as well as the paragraph and clause title to calculate the embeddings. The embeddings we will use are:

- German
  - Word2Vec embeddings with 300 dimensions based on the German Wikipedia[18]
  - GloVe embeddings with 300 dimensions based on the German Wikipedia[19]
  - Word2Vec embeddings with 300 dimensions we trained on the T&C corpus presented in Section 6.1.1
  - A combination of the above approaches, where we used the pre-trained Word2Vec and GloVe embeddings respectively as initial weights and intersected them with the weights learned on our corpus
- English
  - Word2Vec embeddings with 300 dimensions based on the Google News Corpus (Mikolov et al., 2013)

[18] https://gitlab.com/deepset-ai/open-source/word2vec-embeddings-de

[19] https://gitlab.com/deepset-ai/open-source/glove-embeddings-de

- GloVe embeddings with 300 dimensions based on Wikipedia and Gigawords 5 (Pennington et al., 2014)
- Word2Vec embeddings with 300 dimensions we trained on the T&C corpus presented in Section 6.1.1
- A combination of the above approaches, where we used the pre-trained Word2Vec and GloVe embeddings respectively as initial weights and intersected them with the weights learned on our corpus

Since the core promise of word embeddings is that they capture semantic meaning, and especially similarity, the natural way to compare them, independent from a concrete task, is to test how well they capture the similarity between words. Therefore, in both languages, we chose three domain-relevant words and searched for the ten closest words (measured by cosine similarity) in the different models.

In English, we inspected, for example, the words similar to “deliver”, the results we got were:

- T&C: ship, inspect, delivered, collection, returning, collect, dispatch, return, supply, pass
- Word2Vec: delivering, delivered, delivers, Delivering, provide, to deliver, produce, deliver, achieve
- GloVe: delivering, delivered, delivers, provide, promised, able, promise, produce, promises, bring
- Word2Vec + T&C: receive, get, collect, take, ensure that, send, qualify for, need, dispatch, provide
- GloVe + T&C: delivering, delivered, ship, arrive, supply, sell, bring, collect, send, shall pass

At first glance, we can see that for Word2Vec and GloVe, the first words are very similar to the original word, in the sense that they are grammatical derivations from the same stem. For the model we trained solely on our T&C corpus, on the other hand, we see that already the second entry is semantically not very similar to “deliver”, however, it co-occurs very often with it in T&C. For the GloVe model, we can see that after half of the list (promised, able, ...), the meaning seems to change from the literal, physical delivery to a meaning that Merriam-Webster (2020c) describes as “to produce the promised, desired, or expected results”. The intersection between the T&C model and the other models seems to be worse in capturing semantic meaning than GloVe or Word2Vec alone, which is not surprising, given the results produced by the T&C model.

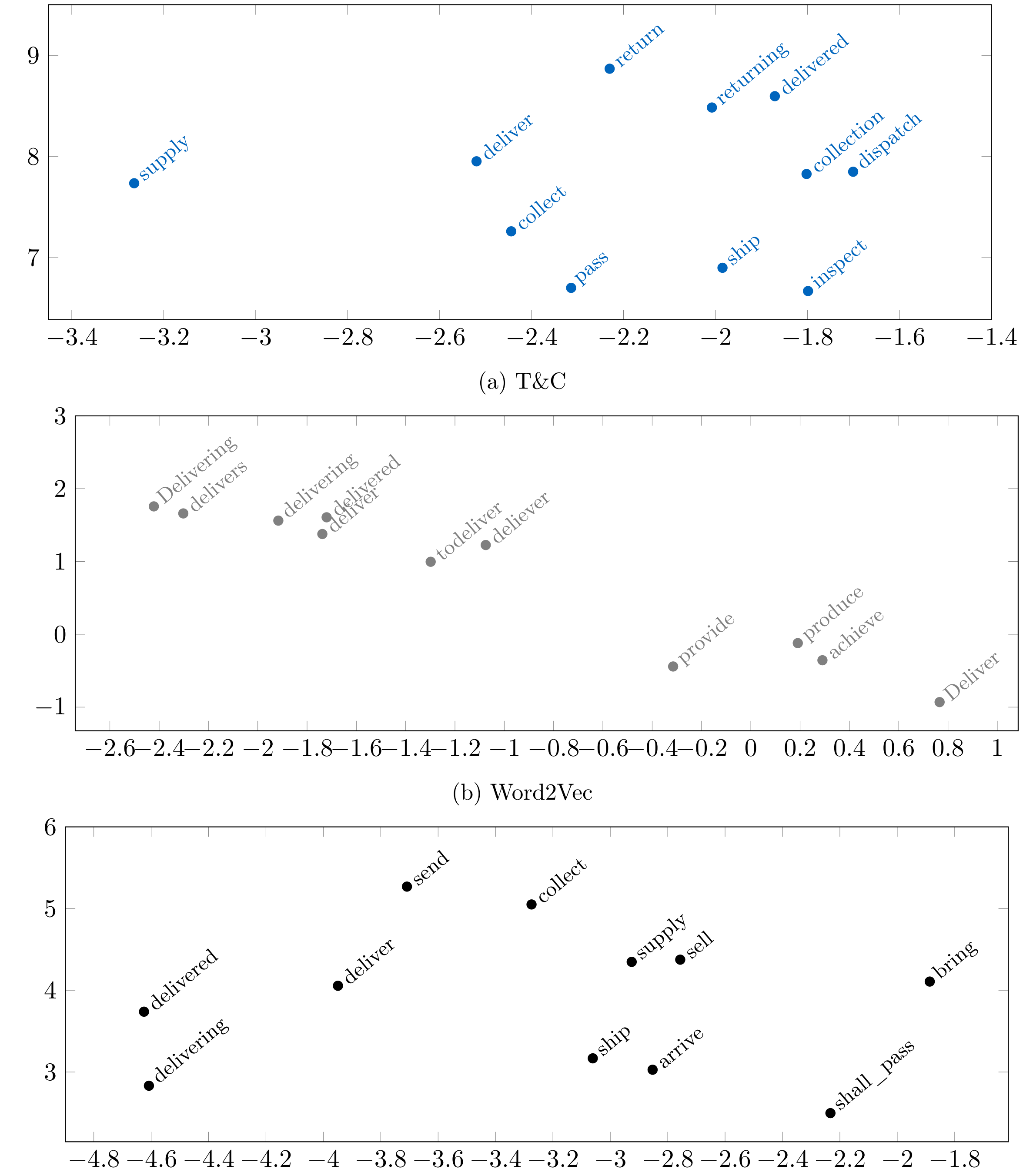
When we inspect not only the words, but also their distances (see Figure 6.23), we can also see that words that are not very similar, like “ship” and “inspect” are very close in the T&C model (Figure 6.23a), while words that are really close in the Word2Vec model tend to be very similar in their semantic meaning (“delivering” and “delivers” or “delivered” and deliver”, Figure 6.23b).

While we only described the results for “deliver” in detail, for reasons of clarity and comprehensibility, we could observe the same pattern for other words like “liability”, “warranty” or "withdraw". In summary, we can say that the embedding model trained on our English T&C corpus does not capture semantic meaning very well, which is most likely caused by the small size of the corpus.

In German, we inspected, among others, the words similar to "Widerrufsrecht" (right of withdrawal), which is one example of a compound noun that is very important in the domain of standard form contracts and has little ambiguity. The results were:

- T&C: Rückgaberecht (right to return), gesetzliches Widerrufsrecht (statutory right of withdrawal), Rücktrittsrecht (right to resign), Gewährleistungsrecht (warranty rights), Widerrufsbelehrung (cancellation policy) Widerrufsrecht (right of withdrawal), Widerrufsbelehrung (cancellation policy), Wiederrufsrecht [sic] (right of withdrawal misspelled), 14-tägiges Widerrufsrecht (14 day right of withdrawal), freiwilliges Rückgaberecht (voluntary right to return), Widerrufsformular (withdrawal form)
- Word2Vec: Rücktrittsrecht (right to resign), Widerrufsbehlehrung (cancellation policy), Rückgaberecht (right to return), Kündigungsrecht (right of termination), allgemeinen Geschäftsbedingungen (terms and conditions), Sonderkündigungsrecht (special right of termination), Pfändungsschutzkonto (garnishment protection account), Rechtsgeschäft (legal transaction), § ff, P-Konto (abbreviation for garnishment protection account)
- GloVe: Rücktrittsrecht (right to resign), Rückgaberecht (right to return), HWiG (abbreviation for the German "doorstep cancellation law"), Verbrauchers (consumer's), Widerruf (withdrawal), BGB (abbreviation for the German civil code), Vorkaufsrecht (preemption), Vorrecht (prerogative), erlischt (expires), Verbrauchern (consumer)
- Word2Vec + T&C: Rückgaberecht (right to return), gesetzliches Widerrufsrecht (statutory right of withdrawal), Rücktrittsrecht (right to resign), Gewährleistungsrecht (warranty rights), Widerrufsbelehrung (cancellation policy) Widerrufsrecht (right of withdrawal), Widerrufsbelehrung (cancellation policy), Wiederrufsrecht [sic] (right of withdrawal misspelled), 14-tägiges Widerrufsrecht (14 day right of withdrawal), freiwilliges Rückgaberecht (voluntary right to return), Widerrufsformular (withdrawal form)
- GloVe + T&C: Rückgaberecht (right to return), gesetzliches Widerrufsrecht (statutory right of withdrawal), Widerrufsbelehrung (cancellation policy), Rücktrittsrecht (right to resign), Widerrfusrechts (genitive form of right of withdrawal), freiwilliges Rückgaberecht (voluntary right to return), gesetzliche (statutory), Regelung (regulation), Widerruf (withdrawal), Verbraucher (consumer)

This time, the picture is different: The lists produced by Word2Vec and GloVe contain words that are, semantically, very far from the original word, like "Pfändungsschutzkonto" (garnishment protection account) or "Vorkaufsrecht" (preemption). Additionally, in the list produced by Word2Vec, we find an entry for "allgemeinen Geschäftsbedingungen" (terms and conditions), and in the list produced by GloVe, we find an entry for the German civil code. In a general domain corpus, like the ones the Word2Vec and GloVe embeddings are trained on, the word "Widerrufsrecht" tends to correlate with the name of the civil code, because it provides legal provision about the right of withdrawal, or with T&C, because they (almost) always contain information about the right of withdrawal. This is one aspect where the German compound



(a) T&C

(b) Word2Vec

(c) GloVe + T&C

Figure 6.23.: Words closest to “deliver” in different word embedding models

nouns again have a special position. The English word “withdrawal” (and even more so the word “right”) can occur in very different contexts. One can withdraw money, withdraw from an organization, or withdraw troops. Therefore, on a general domain corpus, it is less likely that we end up with biases as the ones described before. In German, the nouns “Widerruf” (withdrawal) and “Recht” (Right) can also occur in very different contexts. The compound “Widerrufsrecht”, however, hast virtually no use outside the context of legal contracts.

Figure 6.24 shows the distance between the different words again. The different graphs confirm what the lists have already shown, the Word2Vec (Figure 6.24b) and GloVe (Figure 6.24c) model do not capture the semantic meaning very well in German, while the model trained on the (larger) German T&C corpus does a better job in doing that. We found the same to be true for other words important in the domain, like “Garantie” (warranty), “Haftung” (liability), or “Gewährleistung” (warranty).

For our classification problem, that means that we expect our own embedding models to perform better than Word2Vec and Glove in German, however, not in English. In order to confirm this hypothesis, we will train the different classifiers with input vectors created by the respective models.

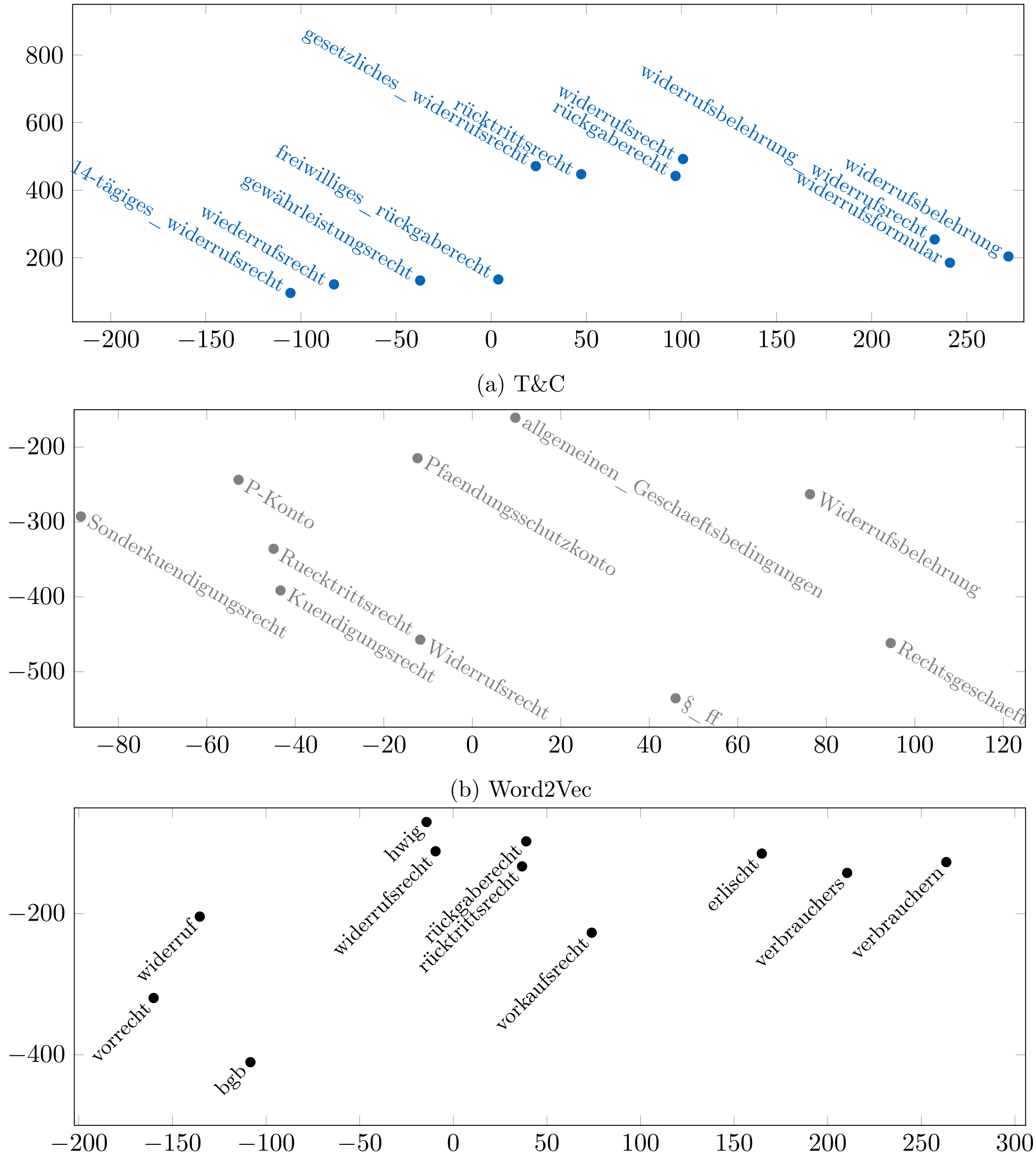
Using different input types also means that we have to find hyper-parameters for every single input type. In order to keep the text readable, we will from now on only report the final results of the hyper-parameter tuning, i.e., the values we used in the end (see, e.g., Table 6.19), and not report unsuccessful combinations anymore. Additionally, in the appendix, we will only report detailed results for the best-performing input type.

The advantage of word embeddings is that they can be trained in an unsupervised fashion, which allows us to use the much bigger corpus described in Section 6.1.1 to train them. This is also why we decided against training paragraph embeddings because we could only train them on the much smaller corpus described in Section 6.3.1, which is separated by paragraphs. In order to aggregate word-level embeddings to paragraph-level (or in our case clause-level) embeddings, there are mainly three approaches: calculating the average of all vectors, summing up all vectors, and using a concatenation of a minimum vector, which consists of the minimum values for all features, and a maximum vector, which contains the maximum values for all features. De Boom et al. (2015) have shown that the min-max approach works especially well on short texts, like our clauses, and a test on our corpus also suggested that the approach performs at least as good or even better than the other approaches. Therefore, we decided to use the approach for all the following experiments.

6.3.6.2. Topics

The results of the hyper-parameter-studies for the topic classification are shown in Table 6.19. It shows the best performing hyper-parameters for the different input types. The results are based on a stratified 5-fold cross-validation, as mentioned before.

The results we achieved on the test set using these hyper-parameters are shown in Table 6.20. In German, we achieved the best results using the Tf-idf vectors as input, which performed comparably to the logistic regression classifier (accuracy 0.76) with an accuracy of 0.75. The



(a) T&C

(b) Word2Vec

(c) GloVe

Figure 6.24.: Words closest to “Widerrufsrecht” in different word embedding models

Language	Input	Layers	Neurons	Dropout	Batch	Epochs
German	Tf-idf	3	250, 150, 250	0.3	50	300
	T&C	2	200, 200	0.4	100	500
	Word2Vec	3	200, 150, 200	0.4	100	300
	Word2Vec + T&C	1	100	0.3	200	300
	GloVe	1	110	0.3	200	200
	GloVe + T&C	1	150	0.2	300	300
English	Tf-idf	2	200, 150	0.3	50	300
	T&C	2	200, 150	0.2	20	300
	Word2Vec	3	200, 150, 200	0.3	100	500
	Word2Vec + T&C	3	200, 150, 200	0.4	200	300
	GloVe	3	100, 200, 100	0.2	50	400
	GloVe + T&C	3	200, 150, 200	0.2	200	300

Table 6.19.: Hyper-parameters used for the topic classification with the Multilayer Perceptron on different inputs (activation function for all was `tanh` and optimizer `adam`)

second-best performance was achieved with the T&C embeddings we trained on our own corpus (accuracy 0.72), which outperformed the Word2Vec and GloVe embeddings as expected. The combined embeddings performed worse than the respective individual embeddings, which we also expected based on the previous findings.

In English, the Word2Vec embeddings as input performed best (accuracy 0.73), however, only slightly better than the Tf-idf vectors (accuracy 0.72). Both perform comparable to the logistic regression classifier again (accuracy 0.71). As we expected, in English, the embeddings trained on our own corpus performed worse; the same applies again for the combined embeddings.

In summary, the simple feed-forward neural network did not perform better on the task of clause topic classification than the logistic regression approach. Depending on the language, either domain-specific word embeddings (German) or more general embeddings (English) performed better. In both cases, the combination of both performed worse; therefore, we will exclude them as input for the next steps.

Intuitively, one might have expected the neural network to outperform the logistic regression classifier. One possible explanation can be found by looking back at the results of the rule-based topic classification. The rule-based topic classification has shown that the topics can be classified quite well with very few keywords or phrases. If we use a Tf-idf vector as input that contains these key-phrases, we have a linearly separable problem. Thus, we could not see a big improvement from the logistic regression to the MLP. This is supported by the fact that the subtopics can not be as easily separated by simple key-phrases, and we hence saw a more significant improvement from the rule-based to the logistic regression approach, and we will show in the next section that the MLP significantly improved the classification performance for subtopics in English again.

Language	Input	A	P	R	F1
German	Tf-idf	0.75	0.89	0.74	0.81
	T&C	0.72	0.87	0.72	0.79
	Word2Vec	0.71	0.85	0.70	0.77
	Word2Vec + T&C	0.68	0.83	0.68	0.75
	GloVe	0.67	0.84	0.67	0.75
	GloVe + T&C	0.62	0.84	0.61	0.71
English	Tf-idf	0.72	0.79	0.73	0.76
	T&C	0.67	0.74	0.68	0.71
	Word2Vec	0.73	0.80	0.74	0.77
	Word2Vec + T&C	0.57	0.65	0.58	0.61
	GloVe	0.63	0.69	0.64	0.67
	GloVe + T&C	0.54	0.63	0.54	0.59

Table 6.20.: Comparison of the Multilayer Perceptron clause topic classification results on different inputs in German and English (A = accuracy, P = precision, R = recall, F1 = F1-score)

Language	Input	Layers	Neurons	Dropout	Batch	Epochs
German	Tf-idf	3	250, 150, 250	0.3	80	300
	T&C	1	150	0.3	500	400
	Word2Vec	3	100, 200, 100	0.4	300	500
	GloVe	1	80	0.3	300	500
English	Tf-idf	3	200, 150, 200	0.2	32	400
	T&C	1	150	0.3	500	400
	Word2Vec	3	100, 200, 100	0.4	150	300
	GloVe	3	200, 150, 200	0.3	150	400

Table 6.21.: Hyper-parameters used for the subtopic classification with the Multilayer Perceptron on different Inputs (activation function for all was `tanh` and optimizer `adam`)

6.3.6.3. Subtopics

For the classification of subtopics, we proceeded in the same way, i.e., we performed a stratified 5-fold cross-validation in order to determine the hyper-parameters (see Table 6.21) and then used these parameters to evaluate the classifier with the different inputs.

The results are shown in Table 6.22. The best result for subtopic classification so far was achieved with logistic regression, with an accuracy of 0.75 in German and 0.54 in English. The best performing configuration in German, using Tf-idf vectors as input, was not able to improve the accuracy. In English, however, we saw a significant increase (+0.13), most likely because the multi-layer perceptron is better in handling the sparse training data than the one-vs-rest logistic regression classifier. If we compare the performance between the different inputs, we see the same patterns as before in the topic classification. In both languages, the Tf-idf vectors performed very well. However, this time, the GloVe embeddings outperformed the custom T&C

Language	Input	A	P	R	F1
German	Tf-idf	0.73	0.86	0.66	0.75
	T&C	0.67	0.84	0.60	0.70
	Word2Vec	0.65	0.78	0.58	0.67
	GloVe	0.70	0.84	0.63	0.72
English	Tf-idf	0.63	0.72	0.63	0.67
	T&C	0.58	0.73	0.58	0.64
	Word2Vec	0.67	0.76	0.66	0.71
	GloVe	0.55	0.65	0.54	0.59

Table 6.22.: Comparison of the Multilayer Perceptron clause subtopic classification results on different inputs in German and English (A = accuracy, P = precision, R = recall, F1 = F1-score)

embeddings in German, most likely because the subtopic classification does not rely as much on single keywords as the topic classification.

Finally, we also investigated a joined-classification approach again by flattening the hierarchy of the labels and training an MLP on that new input. However, we quickly found out that this approach is not very promising. For German Tf-idf vectors as input, we could achieve an accuracy of 0.62 and an F1-score of 0.79, which is considerably worse than the result we would achieve by performing the two steps independently and joining the results. This is not surprising since the amount of training data does not increase by joining the classification tasks; however, the maximum number of labels that have to be predicted almost doubles. This high-dimensional input also leads to a hyper-parameter configuration that is very different from what we saw before. For the mentioned example of German Tf-idf vectors as input, the best hyper-parameter configuration we found in our grid search is a network with three hidden layers with 900, 400, and 100 neurons, respectively.

6.3.7. Convolutional Neural Network

Since the “simple” feed-forward neural networks could not significantly increase performance for topic classification, and only in English for subtopic classification, we also wanted to test more complex networks. We started with a CNN.

CNNs are most popular in computer vision because they are good at dealing with two-dimensional data. In NLP, however, we usually deal with sequential input data. As usual, we will look at topic- and subtopic classification separately and first perform a grid search with 5-fold cross-validation to determine the hyper-parameters for the different inputs and then evaluate them. The hyper-parameters we optimized are the number of convolutional layers, the number of filters and the kernel size per convolutional layer, the number of dense layers and the number of neurons per dense layer, the dropout, the batch size, and the number of epochs. In total, we explored 160 configurations per input. Given a 5-fold cross-validation and four different inputs per language, that means more than 6,000 training and test runs for the topic classification and the same amount again for the subtopic classification.

Lang.	Input	Conv. Layers	Filters	Kernel Size	Dense Layers	Neurons	Dropout	Batch	Epochs
DE	T&C	2	128, 128	3, 3	1	100	0.3	150	10
	W2V	2	256, 128	5, 5	1	100	0.3	50	5
	GloVe	2	128, 128	3, 3	1	100	0.4	150	10
EN	T&C	2	128, 128	3, 3	1	200	0.4	20	10
	W2V	1	128	5	1	100	0.4	20	10
	GloVe	2	128, 128	3, 3	1	100	0.4	20	10

Table 6.23.: Hyper-parameters used for the topic classification with the Convolutional Neural Network on different Inputs (activation function for all was `tanh` and optimizer `adam`)

In order to not further increase the search space, we again fixed some parameters early on, which proved to perform best independently from the rest of the configuration. Therefore, the activation function was again fixed to tanh, the optimizer to adam. We also fixed the length of the input windows to 150.

6.3.7.1. Topics

The best performing hyper-parameters per input for the topic classification are shown in Table 6.23. Independent from the input, we found that the network always performed better if we add one dense layer at the end with 100 neurons. Except for GloVe vectors as input in English, where we found five epochs to perform best, we also found that ten epochs gave us the best results for all other configurations.

The results of the evaluation on the test set are shown in Table 6.24. In general, the results of the CNN does not differ significantly from the results of the MLP. For most inputs, the accuracy and the F1-score are within the range of +/- 0.02, compared to the MLP results. Only for GloVe vectors, which showed the poorest performance before, we could see a more significant improvement (+0.05 F1-score in German and +0.08 in English). Due to this increase, the classification quality for GloVe inputs is now on the same level as the other two. The fact that the performance of the different inputs has leveled when using a CNN for classification might be caused by the fact that, unlike before, we now consider the individual embeddings within a larger window, while before we just used one accumulated vector for the whole sentence. If just one vector in a sentence is “off”, this error would propagate through to the whole sentence. However, if we look at individual vectors, the difference might well be balanced by the rest of the sentence.

Language	Input	A	P	R	F1
German	T&C	0.70	0.92	0.68	0.78
	Word2Vec	0.69	0.90	0.69	0.78
	GloVe	0.71	0.92	0.70	0.80
English	T&C	0.62	0.84	0.63	0.72
	Word2Vec	0.64	0.88	0.66	0.75
	GloVe	0.65	0.85	0.67	0.75

Table 6.24.: Comparison of the Convolutional Neural Network clause topic classification results on different inputs in German and English (A = accuracy, P = precision, R = recall, F1 = F1-score)

Lang.	Input	Conv. Layers	Filters	Kernel Size	Dense Layers	Neurons	Dropout	Batch	Epochs
DE	T&C	2	128, 128	3, 3	1	100	0.3	150	10
	W2V	2	256, 128	5, 5	1	100	0.3	150	10
	GloVe	2	128, 128	5, 5	1	100	0.3	100	10
EN	T&C	1	128	5	1	100	0.4	20	10
	W2V	2	128, 128	3, 3	1	100	0.3	20	10
	GloVe	1	128	5	1	100	0.4	20	5

Table 6.25.: Hyper-parameters used for the subtopic classification with the Convolutional Neural Network on different Inputs (activation function for all was `tanh` and optimizer `adam`)

6.3.7.2. Subtopics

Table 6.25 shows the results of the hyper-parameter optimisation for the subtopic classification with a CNN. As it was the case for the topic classification, we found again that an added dense layer with 100 neurons performed best in all configurations, as well as a training of ten epochs, except for English GloVe vectors, where five epochs again brought the best results. For the convolutional layers, however, we found slightly different hyper-parameters to work better.

The classification results achieved with these hyper-parameters are shown in Table 6.26. In German, we can see similar effects as we have seen for the topic classification. Except for GloVe as input, the results differ only slightly (+/- 0.02 F1-score) from the MLP. For GloVe, we see a slightly larger improvement (+0.03 F1-score), which, in total, levels the differences between the different inputs. For English, we see a significant decrease through all input types.

6.3.8. Recurrent Neural Network

Since the application of a CNN did not provide better results than the simple MLP, we next tried to apply an LSTM network, which, unlike CNNs, which were developed for computer vision

Language	Input	A	P	R	F1
German	T&C	0.66	0.93	0.59	0.72
	Word2Vec	0.67	0.89	0.60	0.72
	GloVe	0.62	0.91	0.57	0.70
English	T&C	0.41	0.79	0.38	0.52
	Word2Vec	0.54	0.77	0.52	0.62
	GloVe	0.38	0.91	0.33	0.49

Table 6.26.: Comparison of the Convolutional Neural Network clause subtopic classification results on different inputs in German and English (A = accuracy, P = precision, R = recall, F1 = F1-score)

Lang.	Input	Sequ. Length	LSTM Layers	Neurons	Dense Layers	Neurons	Dropout	Batch	Epochs
DE	T&C	35	1	300	2	200, 50	0.3	300	30
	W2V	50	1	300	1	50	0.3	40	30
	GloVe	50	1	300	1	50	0.3	40	13
EN	T&C	100	1	300	2	200, 50	0.3	150	100
	W2V	40	1	45	0		0.6	10	50
	GloVe	65	1	200	1	65	0.7	15	100

Table 6.27.: Hyper-parameters used for the topic classification with the LSTM on different inputs (activation function for all was `tanh` and optimizer `adam`)

tasks, is optimized for processing textual input. We applied the same procedure as before: We first conducted a grid-search using a five-fold cross-validation to optimize the hyper-parameters for the different inputs on the training data and then evaluated the classification performance on the test data.

6.3.8.1. Topics

Table 6.27 shows the hyper-parameters we identified. Unlike for the CNN, we found that the optimal hyper-parameters for different inputs are much more diverse in all aspects, from the length of the input sequence to the number of dense layers and the dropout rate.

Table 6.28 shows the results of the evaluation of the topic classification. The overall performance is very much comparable with the performance that both the CNN and the MLP showed. At first, it might seem surprising that we do not see a significant increase in the performance; however, the results are in line with the findings of Yin et al. (2017), who found that LSTM networks do not provide advantages in cases where the classification is mainly dependent on individual key-phrases, which is the case for the classification of clause topics in standard form contracts, as we have discussed before.

Language	Input	A	P	R	F1
German	T&C	0.73	0.90	0.72	0.80
	Word2Vec	0.71	0.89	0.71	0.79
	GloVe	0.72	0.89	0.72	0.79
English	T&C	0.72	0.81	0.71	0.76
	Word2Vec	0.70	0.79	0.69	0.74
	GloVe	0.72	0.80	0.74	0.77

Table 6.28.: Comparison of the LSTM clause topic classification results on different inputs in German and English (A = accuracy, P = precision, R = recall, F1 = F1-score)

Lang.	Input	Sequ. Length	LSTM Layers	Neurons	Dense Layers	Neurons	Dropout	Batch	Epochs
DE	T&C	35	1	250	1	50	0.4	15	20
	W2V	45	1	200	1	50	0.4	15	20
	GloVe	45	1	105	1	50	0.3	15	10
EN	T&C	45	1	200	1	50	0.3	20	15
	W2V	50	1	250	1	50	0.4	25	25
	GloVe	40	1	150	1	50	0.3	15	10

Table 6.29.: Hyper-parameters used for the subtopic classification with the LSTM on different inputs (activation function for all was `tanh` and optimizer `adam`)

6.3.8.2. Subtopics

The results of the hyper-parameter optimisation for the sub-topic classification are shown in Table 6.29. In most of the cases, the best performing architectures for the sub-topic classification are more simple than the ones for the topic classification, i.e., they have fewer neurons in the LSTM layer and less dense layers with fewer neurons.

The evaluation of the classification performance is shown in Table 6.30. We can see that, especially in English, the LSTM outperforms the CNN and reaches results comparable to those of the MLP. In German, the LSTM also outperforms the CNN, however, only by a very slim margin and does not quite reach the quality of the MLP.

6.3.9. BERT

From the results of the previous evaluations, it became clear that the limiting factor was not so much the classifier we use but rather the input encoding. As we will discuss in more detail in Section 6.3.11, for the task at hand, the word embeddings did not manage to capture the semantics of the clauses as much as we would have hoped for. Therefore, we decided to test using a transformer model, more specifically, the Bidirectional Encoder Representations from

Language	Input	A	P	R	F1
German	T&C	0.69	0.85	0.63	0.72
	Word2Vec	0.68	0.85	0.61	0.71
	GloVe	0.69	0.86	0.62	0.72
English	T&C	0.64	0.73	0.63	0.68
	Word2Vec	0.67	0.78	0.65	0.70
	GloVe	0.63	0.73	0.61	0.66

Table 6.30.: Comparison of the LSTM clause subtopic classification results on different inputs in German and English (A = accuracy, P = precision, R = recall, F1 = F1-score)

Transformers (BERT) language model (Devlin et al., 2019). We used the HuggingFace transformers library (Wolf et al., 2019) to fine-tune the pre-trained language models and implement the classification.

For English, we used the "bert-base-uncased" pre-trained model, provided by the original authors Devlin et al. (2019). The model, which is trained on lower case English texts, has 12 hidden layers with a size of 768, 12 attention heads per attention layer, and 110 million parameters. For German, we used the "bert-base-german-cased" model from Chan et al. (2020). It is trained on cased German texts and, like the original model, has 12 hidden layers with a size of 768, 12 attention heads per attention layer, and 110 million parameters.

The original BERT language model was trained on the English Wikipedia and the BookCorpus by Zhu et al. (2015), which consists of 11,038 fiction books that are available for free on the internet. The German language model we are using was pre-trained on a more diverse set of sources, among which are the German Wikipedia and a web corpus gathered by Suárez et al. (2019), which account for more than 90% of the data the model was trained on. However, the model was also trained on the Open Legal Data set from Ostendorff et al. (2020), which consists of more than 100,000 German court decisions.

We also briefly evaluated a multilingual approach with the Multilingual Universal Sentence Encoder transformer model we used for the automated detection of T&C in Section 6.1, however, first tests on German and English were not promising, so we discard the thought.

6.3.9.1. Topics

We used our data set to fine-tune both language models, the English and the German, for the topic classification task. In order to find the best hyper-parameters for the fine-tuning, we split 20% off the training data as validation set. We started our search with the values suggested in the original BERT paper: batch size 16 or 32, learning rate 5e-5, 3e-5 or 2e-5, and 2, 3 or 4 epochs (Devlin et al., 2019). However, the authors also note that the optimal hyper-parameters are task-specific and that small data sets (which they define as less than 100,000 labels) are more sensitive to the choice of parameters than larger ones. For our data sets and task, we found a smaller batch size with a slightly higher number of epochs to work better than the suggested parameters in both languages (see Table 6.31).

Language	Batch	Epochs	Learning Rate
German	8	8	5e-5
English	8	6	5e-5

Table 6.31.: Hyper-parameters used for fine-tuning the BERT language models for the topic classification (all other parameters were kept equal to the pre-trained model)

Language	A	P	R	F1
German	0.84	0.93	0.89	0.91
English	0.79	0.89	0.82	0.85

Table 6.32.: Comparison of the BERT clause topic classification results in German and English (A = accuracy, P = precision, R = recall, F1 = F1-score)

The results of the evaluation, after fine-tuning the models with these parameters, are shown in Table 6.32. BERT outperforms all other approaches we have tested before. In German, it increases the accuracy. In comparison, the previously best-performing approach logistic regression, by nearly 10%. In English, the increase is even higher, with more than 11%. This increase is largely caused by an increase in the recall, which is, as mentioned before, especially desirable for the kind of system we want to build.

6.3.9.2. Subtopics

We performed the same procedure for the classification of subtopics. We again found a smaller batch size and a larger number of epochs to work best. In the case of English, we found a significantly larger number of epochs to work best for the task of subtopic classification (see Table 6.33).

The evaluation of subtopic classification shows similar results than the topic classification: BERT outperformed all previously tested approaches again, although the performance increase is smaller this time, with an increase of 5% in German and just 1.5% in English (see Table 6.34).

We also evaluated the performance of a joined classification of topics and subtopics, but, like before, we found that a joined classification decreases the quality of both the topic and the subtopic classification.

Language	Batch	Epochs	Learning Rate
German	8	6	5e-5
English	8	21	5e-5

Table 6.33.: Hyper-parameters used for fine-tuning the BERT language models for the subtopic classification (all other parameters were kept equal to the pre-trained model)

Language	A	P	R	F1
German	0.79	0.89	0.83	0.86
English	0.69	0.79	0.68	0.73

Table 6.34.: Comparison of the BERT clause subtopic classification results in German and English (A = accuracy, P = precision, R = recall, F1 = F1-score)

6.3.10. Transferability

As we have pointed out before, this part of the semantic analysis of standard form contracts is the one where we expect to have the lowest transferability between T&C from online shops and other types of standard form contracts because the taxonomy we use for the classification is domain-specific. In order to test this assumption, we again turned to the general business condition from banks, that we already used in Section 6.1.5.

We chose three of the ten contracts (Commerzbank, Deutsche Bank, and Sparkassen) and manually separated them by clauses, which we then annotated in the same way we annotated the T&C before: First, a student and the author annotated the clauses independently with their topics and subtopics, then conflicting labels were resolved by the team of experts. We used the taxonomy described in Section 4.4.1 for the classification, however, we added an additional class "n.a." to mark clauses that cover a topic that is not represented by any of the classes in the taxonomy.

Among the three contracts, there were 214 clauses, roughly 71 per contract. In comparison: The German T&C on average consisted of 35 clauses, the English of 34. The 214 clauses consist of 13,681 words, which equals to 4,560 words per contract. The German T&C in our corpus consist of an average of 2,478 words, the English ones of 1,846. These numbers tell us already that, if we have almost twice the number of clauses per contract, we probably have a lot of clauses that will not be covered by our taxonomy.

The results of the annotation process, which are shown in Table 6.35, confirm this assumption. Of the 214 clauses, only 71 (or 33%) are concerned with a topic that is covered by our taxonomy. The other clauses are concerned with a wide range of banking specific topics, from deposit protection funds to banking confidentiality. This means that, even if we would correctly classify all the other clauses, we could never achieve a recall above 0.33. This already shows that the taxonomy we developed can not simply be applied to other types of standard form contracts.

Since the taxonomy is still able to cover one-third of the banking contracts, we wanted to test how well the classifiers we trained would perform on this data-set. Therefore, we took the best performing topic classifier, i.e., the BERT model we fine-tuned, and applied it to the new corpus. The evaluation of the results is shown in Table 6.36. We can see that, for some of the topics, e.g., applicability, applicableLaw, arbitration, and changes, the performance is actually very good, despite the fact that our model has never seen this type of contract before.

We can see that, especially for the more "technical" clauses, i.e., clauses that govern the contract itself, rather than the actual subject of the transaction, models can be transferred between different types of standard form contracts. During the annotation of the corpus, we also fund

Topic	#clauses
applicability	3
applicableLaw	3
arbitration	2
changes	5
liability	9
n.a.	143
payment	14
placeOfJurisdiction	6
prices	1
withdrawal	27

Table 6.35.: Topics of clauses in the general business conditions of banks

Topic	P	R	F1
applicability	0.60	1.00	0.75
applicableLaw	1.00	1.00	1.00
arbitration	1.00	1.00	1.00
changes	1.00	0.83	0.91
liability	0.41	1.00	0.58
payment	0.16	1.00	0.28
placeOfJurisdiction	0.00	0.00	0.00
prices	0.00	0.00	0.00
withdrawal	0.54	1.00	0.70
TOTAL	0.28	0.97	0.43

Table 6.36.: BERT clause topic classification results on the banking corpus (A = accuracy, P = precision, R = recall, F1 = F1-score)

that the banking contracts are even more similar to one another than the T&C from online shops. Therefore, it would be very easy to define a common taxonomy for them. So even if the actual taxonomy and models can not be transferred between different types of standard form contracts, the approach we have taken can be, and we believe that it can not just be applied to online shopping and banking, but every type of consumer standard form contract within a highly regulated domain, like insurances, housing, and employment.

6.3.11. Summary

In Section 6.3, we have presented an extensive comparative study of different classifiers and different input format for the classification of topics and subtopic of clauses from standard form contracts. We tested tens of thousands of combinations of different inputs and hyper-parameters to find the optimal results using hundreds of hours of computing power.

Table 6.37 shows the best results for topic classification we could achieve with each approach, in-

Approach	A	P	R	F1
BERT	0.84	0.93	0.89	0.91
Log. Regression	0.77	0.95	0.80	0.87
Random Forest	0.73	0.97	0.72	0.83
MLP	0.75	0.89	0.74	0.81
LSTM	0.73	0.90	0.72	0.80
CNN	0.71	0.92	0.70	0.80
Rule-based	0.64	0.77	0.80	0.78

(a) German

Approach	A	P	R	F1
BERT	0.79	0.89	0.82	0.85
Log. Regression	0.71	0.88	0.73	0.80
LSTM	0.72	0.80	0.74	0.77
MLP	0.72	0.79	0.73	0.76
CNN	0.65	0.85	0.67	0.75
Rule-based	0.57	0.76	0.69	0.72
Random Forest	0.57	0.88	0.58	0.70

(b) English

Table 6.37.: Best clause topic classification results for each approach, ordered by F1-score (A = accuracy, P = precision, R = recall, F1 = F1-score)

Approach	A	P	R	F1
BERT	0.79	0.89	0.83	0.86
Log. Regression	0.75	0.91	0.78	0.84
Random Forest	0.68	0.91	0.67	0.77
MLP	0.73	0.86	0.66	0.75
LSTM	0.69	0.85	0.63	0.72
CNN	0.67	0.89	0.60	0.72
Rule-based	0.47	0.74	0.56	0.64

(a) German

Approach	A	P	R	F1
BERT	0.68	0.79	0.68	0.73
MLP	0.67	0.76	0.66	0.71
LSTM	0.67	0.78	0.65	0.70
Log. Regression	0.54	0.80	0.59	0.68
CNN	0.54	0.77	0.52	0.62
Random Forest	0.44	0.85	0.43	0.57
Rule-based	0.28	0.39	0.49	0.43

(b) English

Table 6.38.: Best clause subtopic classification results for each approach, ordered by F1-score (A = accuracy, P = precision, R = recall, F1 = F1-score)

dependent from the input and hyper-parameters, Table 6.38 shows the same for the classification of sub-topics.

BERT performed best on both tasks and languages, and the rule-based classifier performed worse overall. While these two results might not be very surprising, the ranking in between might be. Except for the sub-topic classification in English, the classical logistic regression outperformed the computing-intensive neural networks. However, what we see here is, we believe, much more a ranking of input dimensionality than a ranking of classifiers. The logistic regression classifier achieved these results by using a Tf-idf vector as input that has as many dimensions as there are different words in the corpus, so did the random forest classifier, which performed third-best in German for both tasks. The MLP achieved the best results using a Tf-idf vector with 2000 dimensions as input. For the LSTM and the CNN, we could not test such high-dimensional inputs because of the computational complexity. Therefore, we believe the comparably poor (absolutely still good) performance is caused by the relatively low dimensional input (300 dimensions).

Word embeddings generally promise to be able to represent semantic meaning in a much smaller dimensional space than, e.g., Tf-idf or count-vectors. In our case, however, neither the pre-trained nor the task-specific embedding models seem to have been able to fulfill that promise sufficiently, especially in German. As we have pointed out before, we believe that one of the reasons for this is the fact that, in German, the topic classification largely depends on domain-specific composite nouns, which the embedding models probably did not capture well.

Finally, we want to point out that, while BERT was able to significantly increase performance compared to the logistic regression, this increase was bought with computational complexity. Training the initial Bert model took the authors four days on (Devlin et al., 2019) 16 specialized chips (Tensor Processing Unit (TPU)) and the fine-tuning took us again about one hour on a GPU, not including the hyper-parameter search, compared to the logistic regression classifier which can be trained within a few minutes on a CPU. Once the model is trained, the classification with BERT is still computationally more expensive and time-consuming.

6.4. Information Extraction

In the last step of the semantic analysis, we extract relevant information from the different clauses in a structured format, which can be used for the subsequent processing steps, i.e., the legal assessment and summarization of the standard form contracts. Which information is deemed as relevant is mostly influenced by which information is needed for the legal assessment presented in Chapter 7, but also by the preferences given by consumers in Section 4.2.

First, we asked the experts from the consumer protection agencies to provide us for each topic and subtopic in the taxonomy with the information they think should be extracted from the respective clauses. Then we tried to assign each information with a corresponding data type. The result of this process is shown in Table 6.39. In total, the experts defined 21 data points which they would like to see extracted, which cover 13 out of the 23 topics. For the remaining ten topics, the experts either deemed them not to be relevant for further analysis (e.g., severability[20]), not the focus of the work (e.g., personal data) or they could not think of a way to structure the contained information, because they are too diverse or vague (e.g., conclusion of contract). However, just because we do not extract information from a clause type does not necessarily mean that we will not be able to detect void clauses of this type, it just means that we will need a different assessment strategy. We will use an ML-approach based on a transformer model for such clauses, as we did for the topic classification.

In this section, we will first introduce the corpus that we created for this task and then describe two different approaches for the automatic extraction of information, one using ML and one using rules. While there is a wide range of open source libraries for Named Entity Recognition (NER), there are significantly less for more complex and structured Information Extraction (IE). The goal of NER, in short, is to differentiate different types of entities. In the sentence "The law of the Federal Republic of Germany shall apply in principle, excluding the UN Convention on Contracts for the International Sale of Goods.", an NER algorithm would, in the best case, detect that two entities of type Law are present, "law of the Federal Republic of Germany" and "UN convention on Contracts for the International Sale of Goods". However, they appear in

[20] This does not mean that there are no void severability clauses, on the contrary, a lot of them are actually void. However, they almost never lead to negative consequences for consumers and are also almost never complained about and therefore not very relevant for the work of the consumer advocates.

very different roles, one as applicable, one as not applicable, a piece of information that is not captured by NER, but that we need for further analysis.

Subtopic	**Information and Data Type**	**Description**
age		
-	minimum age: number	Minimum age to enter into a contractual relationship.
applicability		
applicableLaw		
-	applicable law(s): list of strings, not-applicable law(s): list of strings	List of which laws are applicable and which are not applicable.
arbitration		
-	link: URL	Extracts the link to the arbitration platform.
changes		
codeOfConduct		
-	name: String, link: URL	Name of the code of conduct the shop submits to and link to the code itself.
conclusionOfContract		
definition	term: String	Term that is defined.
delivery		
costs	costs: amount of money	Delivery charges.
time	time: timespan	Delivery time.
description		
disposal		
intellectualProperty		
language		
-	language: list of strings	Languages in which the contract can be closed.
party		
-	name: String, address: String	Name and address of the contracting party.
payment		
fee	fee: money value	Fee charged for payment methods.
late	fee: money value, interest: percentage	Fees and interest reates for late payments.
personalData		
placeOfJurisdiction		

Table 6.39.: Relevant information in different clause types and their data type

Subtopic	Information and Data Type	Description
-	place: String	Place of Jurisdiction.
prices		
currency	currency: currency	Currency in which prices are given.
retentionOfTitle		
severability		
textStorage		
warranty		
period	period: timespan	Warranty period.
withdrawal		
form	allowed: list of string, forbidden: list of string	List of allowed and forbidden forms of withdrawal.
period	period: timespan	Withdrawal period.

Table 6.39.: Relevant information in different clause types and their data type

6.4.1. Corpus

First, we needed an additional corpus to evaluate rule-based information extraction approaches and train and evaluate stochastic approaches. We used the relevant clauses from the topic classification corpus described in Section 6.3.1, i.e., clauses which belong to one of the (sub)topics we want to extract information from. We then labeled all of these clauses with the relevant information. We again used two annotators (a student and the authors) to independently label the corpus twice. Conflicts were again presented to the experts and decided by them. We had a very high inter-annotator agreement of more than 92% and therefore only had to present very few clauses to the experts.

We used a web-based graphical interface for the annotation of intent and entities provided by RASA (see Section 2.2.1) for the initial annotation and presented the conflicts as a spreadsheet to the experts. Listing 6.4 shows the data format in which the annotations are stored by the tool provided by RASA: Each clause is an object which contains its topic (or class), its content in the form of an array of sentences, which themselves consists of an array of tokens, and an array of annotations, which describe the passage in the text which holds the information based on the token number.

```
{
    "class": "applicableLaw",
    "sentences": [
        [
            "(",
            "3",
            ")",
            "The",
            "law",
            "of",
            "the",
            "Federal",
            "Republic",
            "of",
            "Germany",
            "shall",
            "apply",
            "in",
            "principle",
            ",",
            "excluding",
            "the",
            "UN",
            "Convention",
            "on",
            "Contracts",
            "for",
            "the",
            "International",
            "Sale",
            "of",
            "Goods",
            "(",
            "CISG",
            ")",
            "."
        ],
        [
            "All",
            "contracts",
            "are",
            "concluded",
            "in",
            "English",
            "."
        ]
    ],
    "annotations": [
        {
            "label": "applicable",
            "sentence": 0,
            "start": 4,
            "end": 10
        },
        {
            "label": "notApplicable",
            "sentence": 0,
            "start": 18,
            "end": 30
        }
    ]
}
```

Listing 6.4: Data format of the information extraction corpus

Table 6.40 gives an overview of the annotated data set, which consists of 1,819 German clauses and 380 English clauses. In the German data set, we have 2,357 instances of annotated information and, in the English data set, 474 instances. One clause can contain multiple instances of information that we want to extract. Withdrawal clauses, for example, frequently state the withdrawal period repeatedly: "You have the right to cancel this contract within 14 days without giving any reasons. The 14 days cancellation period starts from the day on which you acquire physical possession of the goods." However, a clause can also not contain relevant information to extract. Clauses about the delivery costs, for example, frequently just mention that the costs of delivery will be shown to the customer during the checkout process.

The overview in Table 6.40 already shows us that some contracts in the corpus do not fulfill all regulations. We know, for example, that the German corpus consists of 142 contracts, but we found only 94 links to arbitration platforms. However, as many courts have ruled, e.g., the Landgericht Dortmund (2020), online shops need to provide a (clickable) link to the Online Dispute Resolution (ODR) platform of the EU. At least 48 of the 142 contracts do not meet this information obligation.[21]

We found the information provided in the corpus the be quite diverse. Delivery times in the corpus, for example, span from one day to three months (for bespoke goods), reminder fees for late payments range from 70 Cents to 40 Euros, and interest rates from 5% to 10% above the base interest rate, which is illegal. We also found a number of consumer-friendly regulations. While most shops offer the legally required withdrawal period of 14 days, we found that a significant number of shops offer increased withdrawal periods, from 30 days up to 6 months.

6.4.2. MITIE

The MIT Information Extraction (MITIE) library is a machine learning library developed at the Massachusetts Institute of Technology (MIT). It is build based on the Dlib-ml library by King (2009), a high-performance ML library, and uses word embeddings and Structural Support

[21] The number is actually even higher, because some T&C contain, in addition to the link to the EU platform, links to private ODR platforms.

Language	Topic	Information	# Instances	# Clauses
DE	age	age	28	38
	applicableLaw	applicable	189	137
	applicableLaw	notApplicable	112	137
	arbitration	link	94	155
	codeOfConduct	name	68	55
	codeOfConduct	link	61	55
	conclusionOfContract:definition	term	194	103
	delivery:costs	costs	81	247
	delivery:time	time	77	143
	language	language	134	124
	party	name	186	157
	party	address	122	157
	payment:fee	fee	24	50
	payment:late	interest	25	48
	payment:late	fee	14	48
	placeOfJurisdiction	place	166	117
	prices:currency	currency	17	17
	warranty:period	period	223	155
	withdrawal:form	allowed	126	131
	withdrawal:period	period	351	126
	Total		2357	1819
EN	age	age	5	5
	applicableLaw	applicable	29	23
	applicableLaw	notApplicable	9	23
	arbitration	link	8	13
	codeOfConduct	name	1	1
	codeOfConduct	link	1	1
	conclusionOfContract:definition	term	12	4
	delivery:costs	costs	16	57
	delivery:time	time	35	41
	language	language	11	11
	party	name	16	21
	party	address	17	21
	payment:fee	fee	1	3
	payment:late	interest	1	1
	payment:late	fee	0	0
	placeOfJurisdiction	place	33	19
	prices:currency	currency	14	13
	warranty:period	period	10	10
	withdrawal:form	allowed	18	37
	withdrawal:period	period	60	40
	Total		474	380

Table 6.40.: Data set for the extraction of information

Vector Machines.[22] There are three pre-trained language models available for MITIE. In addition to German and English, there is also a model for Spanish. Out-of-the-box, MITIE supports the extraction of standard entity types, like place and persons, but also supports the training of custom extractor models. MITIE is also used by the version of RASA we evaluated in Section 2.2.1, which outperformed other available NLU services. We use the information from the previous steps of the pipeline, most importantly the detected topic and subtopic, as input for the information extraction, in order to reduce the likelihood of false positives during the extraction.

6.4.2.1. Training

We split the data set described in Section 6.4.1 into a training (80%) and a test set (20%) while remaining the original distribution of labels. The training process of MITIE is fully automated and consists of two steps: The first step is the training of a segmenter, which segments the input sentences into text units that (possibly) represent pieces of information that are to be extracted. The second step is the training of a classifier, which classifies for each of these segments, whether they belong to a relevant class, i.e., whether they represent information that should be extracted. In both steps, MITIE first performs a hyper-parameter tuning step to find suitable values for the hyper-parameter C (inverse of the regulator ($1/\lambda$), see also Chapter 2).

We trained two separate custom extraction models for German and English on a machine with a 2.9 GHz Dual-Core processor and 16 GB of DDR3 RAM. The first training step of the German extraction model took about 1 hour and 15 minutes to complete. The hyper-parameter C was set to 28.1818. The second training step took considerably longer, with a total time of 23 hours and 10 minutes. C was set to 104.424. The final model has a size of around 370 MB. On the much smaller English data set, the first step took just 9 minutes, and the hyper-parameter C was set at 11.4378. The second step took about 12 hours, and C was set to 300. The final model has a size of around 350 MB.

6.4.2.2. Evaluation

In order to evaluate the extraction model we trained, we feed it with the sentences from our test data set. For each sentence, the model returns an array of extracted entities, which consist of a range of tokens which represent the entity, the type of the entity, and a score that represents the likelihood that an entity belongs to the given class. The score is given by the SVM that MITIE uses for the classification and is based on the distance from the hyperplane that is used for the classification. As such, the score has no fixed domain, but the larger the score is, the more certain is the classification.

Evaluating the information extraction is not a binary task because extracted entities can also be partially correct, e.g., when the correct fee would be “5 EUR”, but only “5” would be extracted. We will therefore report two different results for each metric, one which only considers exact matches to be true and one which considers any overlap to be true. Other than that, we will

[22] https://github.com/mit-nlp/MITIE

Information	P	R	F1
address	0.67	0.08	0.14
age	0.80	0.67	0.73
allowed	1.00	0.04	0.08
applicable	0.88	0.76	0.82
costs	0.62	0.31	0.41
currency	1.00	0.33	0.50
fee	0.75	0.38	0.50
interest	1.00	0.60	0.75
language	0.96	0.89	0.92
link	1.00	0.90	0.95
name	0.75	0.29	0.42
notApplicable	1.00	0.55	0.71
period	0.98	0.40	0.57
place	1.00	0.64	0.78
term	0.96	0.69	0.80
time	1.00	0.14	0.25
TOTAL	0.93	0.49	0.64

(a) Exact matches

Information	P	R	F1
address	0.67	0.08	0.14
age	0.80	0.67	0.73
allowed	1.00	0.04	0.08
applicable	0.89	0.84	0.86
costs	0.67	0.38	0.48
currency	1.00	0.33	0.50
fee	0.80	0.50	0.62
interest	1.00	0.60	0.75
language	0.96	0.89	0.92
link	1.00	0.90	0.95
name	0.78	0.35	0.48
notApplicable	1.00	0.55	0.71
period	0.98	0.40	0.57
place	1.00	0.67	0.80
term	0.96	0.69	0.80
time	1.00	0.21	0.35
TOTAL	0.93	0.51	0.66

(b) Any overlap

Table 6.41.: Results of the Information Extraction with MITIE in German for exact matches and any overlay (A = accuracy, P = precision, R = recall, F1 = F1-score)

follow the evaluation approach used in the CoNLL-2003 shared task on NER (Tjong Kim Sang and De Meulder, 2003), i.e., we will consider it a true positive if an entity is extracted which fully (exact match) or partially (any overlap) matches an entity of the same type in the data, a false positive if an entity is extracted which does not match any entity or an entity of another type, and a false negative if an entity is in the data but no entity of the same type is extracted that fully (exact match) or partially (any overlap) matches that entity. Nadeau (2007) presents a comprehensive overview of NER evaluation strategies, which also includes other approaches.

Table 6.41 shows the result of the evaluation on the German corpus. For some types, the information extraction performed very well, e.g., for the extraction of links (F1 = 0.95) or languages (F1 = 0.92). For others, it performed very poorly, e.g., the allowed forms of withdrawal (F1 = 0.08) or addresses (F1 = 0.14). Surprisingly MITIE also performed rather poorly on extracting the information for delivery costs. One would think this would be an easy task because we basically have to extract money values. However, most of the information that MITIE missed were instances of free delivery, where the costs are not described as a number but with words. The same applies to the information on the period of withdrawal and warranty. The fact that they can appear as either number or word seems to have confused the extractor. Generally speaking, MITIE achieves a very high precision in German (0.93), however, at the cost of a low recall (0.49), which only increases slightly (+0.02) if we also consider partial overlaps.

Table 6.41 shows the result of the evaluation on the English corpus. Due to the smaller size of the corpus, some information types are not part of the evaluation because if there were less than five instances of a type in total in the corpus, none of them was added to the test set. As we would expect, based on the size of the corpus, MITIE performed worse on the English corpus, however, the difference is relatively small with -0.10 for the F1 score for exact matches and -0.06 for partial matches. The recall for partial matches is even closer (-0.01.) to the performance on the German corpus. Most likely, the language model, which is used by MITIE to train the classifier, is stronger in English than it is in German. We, therefore, assume that, given a data set of the same size, MITIE would outperform the results on the German set in English.

Information	P	R	F1
address	1.00	0.33	0.50
allowed	0.00	0.00	0.00
applicable	0.50	0.17	0.25
costs	0.33	0.33	0.33
currency	1.00	0.67	0.80
language	1.00	1.00	1.00
link	1.00	0.50	0.67
name	1.00	0.33	0.50
notApplicable	1.00	1.00	1.00
period	0.60	0.64	0.62
place	0.00	0.00	0.00
time	1.00	0.57	0.73
TOTAL	0.71	0.43	0.54

(a) Exact matches

Information	P	R	F1
address	1.00	0.33	0.50
allowed	0.00	0.00	0.00
applicable	0.67	0.33	0.44
costs	0.50	0.67	0.57
currency	1.00	0.67	0.80
language	1.00	1.00	1.00
link	1.00	0.50	0.67
name	1.00	0.33	0.50
notApplicable	1.00	1.00	1.00
period	0.62	0.71	0.66
place	0.00	0.00	0.00
time	1.00	0.71	0.83
TOTAL	0.74	0.50	0.60

(b) Any overlap

Table 6.42.: Results of the Information Extraction with MITIE in English for exact matches and any overlay (A = accuracy, P = precision, R = recall, F1 = F1-score)

6.4.3. Rule-based

The rule-based information extraction approach we implemented is based on the analysis of the dependency trees we introduced in Section 2.2.2. Our literature review has shown that this is a commonly used approach, which we already applied successfully in previous work too (Braun et al., 2018c).

The dependency trees themselves are generated with the Stanford CoreNLP library (Manning et al., 2014), which includes a high-performance dependency parser that is based on a neural network with one hidden layer (Chen and Manning, 2014) and comes with language models for English, German, Chinese, and Arabic.

Before we take a look at the individual rules which we apply, we want to explain with two short examples why we use dependency parsing, rather than more simple and computationally less expensive approaches like regular expressions or part-of-speech of tagging. Let us assume we have the clause “We charge 7 EUR for delivery.” and know that the topic of the clause is “delivery”, and the subtopic is “costs”. It would be very easy to write a regular expression that is able to extract the delivery costs from such a clause (look for a number followed by a currency name, symbol, or code). However, extracting information from clauses that use number words ("We charge seven Euros for delivery.") is not feasible with regular expressions. We could handle such cases with PoS tagging, which would recognize both "7" and "seven" as numbers. However, as soon as grammatical constructs become more complex, this approach fails too. If the clause were, for example, "For orders under 40 EUR, we charge 7 EUR for delivery.", with PoS tags alone, we would not be able to figure out which of the numbers is the delivery charge. With a dependency tree, however, we know that "7 EUR" as a nominal phrase is the object to the verb "charge" and hence the number we are looking for.

Extracting and using these dependencies is even more crucial for more complex clauses. Clauses about the possible forms of withdrawal, for example, sometimes include a listing of forms that are allowed or even mandatory and forms that are forbidden: "You may revoke your contractual declaration within two weeks without providing reasons in written form (but not electronically) or by calling our customer support." The challenge here, which neither regular expressions nor PoS tagging can handle reliably, is to determine which parts of the sentence the negation applies to. In a dependency tree, that is a trivial task because each part of the sentence that is affected by the negation has a direct negation-relation to the negation word.

6.4.3.1. Development

We developed the rules together with the experts from the consumer protection agencies and fine-tuned them using the training data of the previously described data-split. In essence, we are trying to find a sub-tree within the dependency tree, which we want to extract and label with the type of information it represents.

Figure 6.25 shows the dependency tree for a sentence from a withdrawal form clause (see Appendix A for an overview of the annotations used). The two sub-trees we would want to extract in this case both only consist of a root: email and letter. The rules we developed analyze the existing relations (i.e., dependencies) to identify sub-trees that should be extracted. Any relation can be characterized by up to five features: the type of relation, the text and PoS of the source node, and the text and PoS of the target node. Our rules differentiate between three types of relations: Relations that belong to the sub-tree that should be extracted, relations that have to be presented in order for a rule to apply but should not be extracted, and relations that may not be present in order for a rule to apply.

Since we use the information from the previous stage, we already know that the sentence shown in Figure 6.25 is concerned with the form of withdrawal. The form of withdrawal is usually described by a noun ("email", "letter", "call"), sometimes in combination with an adjective ("written statement", "clear statement"). Therefore, the rule for the detection of the sub-tree is that the root of the sub-tree must be a noun that modifies the verb of the sentence, and any adjective that modifies the noun is also part of the sub-tree that should be extracted. This defines the boundaries of the sub-trees that should be extracted, however, it is not yet sufficient to make sure that we are really only extracting forms of withdrawal. For this, we use the second categories of relation our rules check for, a relation that has to be present but is not part of the extraction. In this case, we want to make sure that the subject of the sentence is some form of cancellation or withdrawal. i.e., we search for a relation of type "nsubj*" which has a target

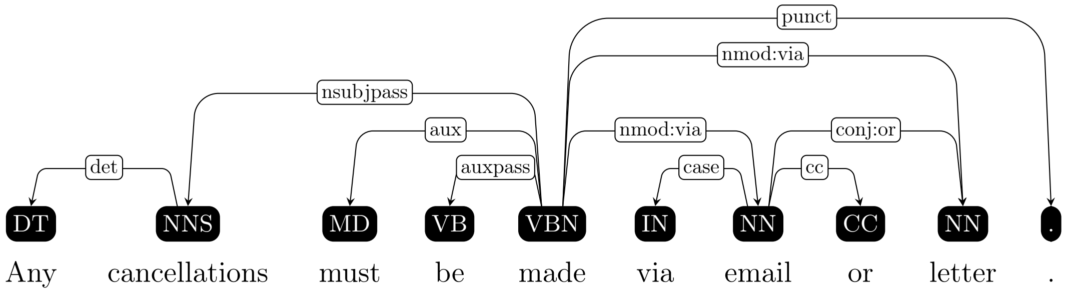


Figure 6.25.: Dependency tree with PoS tags for the sentence "Any cancellations must be made via email or letter."

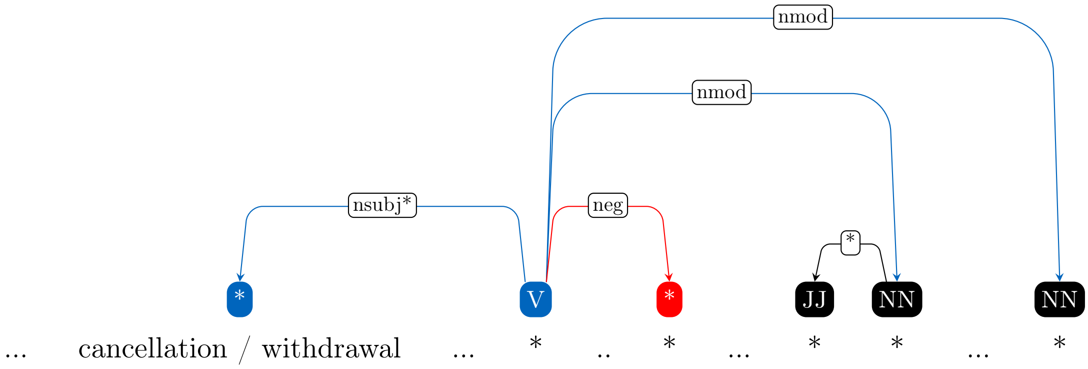


Figure 6.26.: Graphical representation of the rule for the extraction of allowed withdrawal forms, black indicates information that should be extracted, blue indicates relations that have to be presented but are not extracted, and red indicates relations that may not be present

node that contains cancellation or withdrawal (we use the lemmatized texts of the nodes for this check). If we want to extract the allowed forms of withdrawal, these rules are still not enough. If we would change the sentence to "Any cancellations must not be made via email or letter.", the rule would still extract the same information. Therefore we should also make sure that there is no negation relation existing. On the other hand, if we would like to extract forbidden forms of withdrawal, we should make sure that such a relation is present.

Figure 6.26 shows a graphical representation of the rule with all its elements: the black parts are the information that is going to be extracted, the blue parts are relations that have to be present but will not be extracted, and the red part represents the relations that may not be present for the rule to apply. This is not the only rule we developed to extract allowed withdrawal forms. We also use a second rule for cases where cancellation or withdrawal is not used as a noun, but the verbs withdraw or cancel are used, e.g., together with the subject contract (or any other subject). For both versions, we also have a rule in place with negation as a mandatory relation, which is used to extract forbidden forms of withdrawal.

In total, we developed 33 rules, which means that, on average, we need just two rules per information we want to extract. Another advantage of rules that are based on dependency

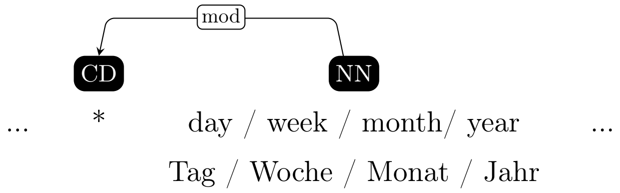


Figure 6.27.: Graphical representation of the rule for the extraction of time periods

trees is that, at least for German and English (and probably also other Germanic languages), the rules are, to a certain extend, language independent. For the rule in Figure 6.26 to also work in German, we just have to replace “cancellation” and “withdrawal” with their German counterparts “widerrufen” and “zurücktreten”, the relations themselves as well as the PoS tags remain unchanged.

While we do not want to go into detail here for each of the 33 rules, we want to give another example of a rule, which is even more simple than the rule for the form of withdrawal, yet, it is not completely language independent. Figure 6.27 shows a graphical representation of a rule for the extraction of time periods for both withdrawal and warranty periods. The rule is pretty straightforward: If a noun describing a time period (day, week, month, year, or their German equivalents) is modified by a number, they together form a time period. This rule is sufficient for English. However, for German, we have to consider a (frequently occurring) edge case: According to § 476 of the German civil code BGB, the warranty (“Gewährleistung”) period can be reduced to one year, in case of used goods. In English, the rule in Figure 6.27 is capable of detecting this period without any problems. In German, however, we are facing some limitations of the PoS tagging because of ambiguity.

The word “ein” in German is the indefinite singular article for male and neuter nouns in nominative (as in “It was a good year” / “Es war ein gutes Jahr”), but also a numeral (as in “One Euro is missing.” / “Ein Euro fehlt.”).[23] During the development of the rules, we found that in German in the context of “ein Jahr Gewährleistung” (“one year warranty”), the numeral “ein” was consistently labeled as determiner. We, therefore, decided to add an extra rule for German, which is shown in Figure 6.28, to handle this case.

Theoretically speaking, these rules would wrongly claim that the sentence “The withdrawal period is not 14 days.” grants a withdrawal period of 14 days. When we design rules that use grammatical structures, we often think about such possibilities. However, in practice, we found that they have only very limited relevance. A sentence like this did not occur in any of the thousands of clauses we investigated during the course of this thesis. Our evaluation will show that with these very basic rules, we achieve an F1-score of 0.92 for the extraction of time periods in German.

[23] The indefinite articles in English (“a” and “an”) etymologically also originate from “one” (Merriam-Webster, 2020a,b).

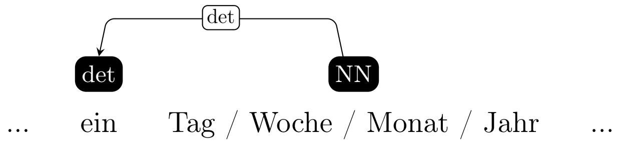


Figure 6.28.: Graphical representation of the additional German rule for the extraction of time periods

6.4.3.2. Evaluation

Table 6.43 shows the results of the evaluation of the rule-based information extraction on the German corpus, and Table 6.44 shows the results on the English corpus. On a high level, we can immediately see that the rule-based approach outperforms the extraction with MITIE significantly with regard to the F1-score in both languages for exact matches and partial matches. While in English, also both precision (+0.16 for exact matches) and recall (+0.18 for exact matches) increased, in German, the increased F1-score is caused by an increased recall (+0.14 for exact matches), while precision decreased (-0.14 for exact matches). Since we want to build an assistance system for domain experts, we deliberately designed our rules in a way that priorities recall over precision, as such, these results are expected. The increase in performance was even higher with regard to partial matches (Recall +0.29 in German and +0.36 in English).

Since our approach is not limited anymore by the scarcity of training data for English, we now perform better in English (F1=0.72 for exact matches and F1=0.88 for partial matches) than in German (F1=0.70 for exact matches and F1=0.81 for partial matches). This is, at least partially, caused by the fact that the dependency trees generated by the Stanford CoreNLP library are more prone to contain errors in German[24] and our rules can only be as good as the dependency tree they are applied to.

If we inspect the results more closely, on the level of individual types of information, we can see that for a number of information types, the F1-score for exact matches is zero (address in both languages and notApplicable in English) or very low (applicable in German and place in both languages), while it is much better for partial matches. These instances are also mostly caused by shortcomings of the dependency trees. For address, for example, they usually occur in a format like this:

```
Company Name
Building Name
0 Street Name
XX 000 XX Place
Country
```

When such a piece of text is given to the dependency parser, it will be flattened to a string that is separated by white-spaces. Since such a string does not form a grammatical sentence, the dependency parser will almost certainly fail to produce a sensible dependency tree from it.

[24] Which is at least partially caused by the fact that there is fewer data available in German to train the dependency parser on.

Information	P	R	F1
address	0.00	0.00	0.00
age	0.40	0.33	0.36
allowed	0.83	0.80	0.81
applicable	0.44	0.21	0.28
costs	0.76	0.81	0.78
currency	1.00	0.67	0.80
fee	0.86	0.75	0.80
interest	1.00	1.00	1.00
language	0.96	0.96	0.96
link	0.97	0.88	0.92
name	0.48	0.25	0.33
notApplicable	0.92	0.55	0.69
period	0.91	0.94	0.92
place	0.71	0.36	0.48
term	0.77	0.62	0.69
time	0.67	0.86	0.75
TOTAL	0.79	0.63	0.70

(a) Exact matches

Information	P	R	F1
address	0.50	0.32	0.39
age	0.57	0.67	0.62
allowed	0.83	0.80	0.81
applicable	0.75	0.79	0.77
costs	0.76	0.81	0.78
currency	1.00	0.67	0.80
fee	0.88	0.88	0.88
interest	1.00	1.00	1.00
language	0.96	0.96	0.96
link	0.97	0.94	0.95
name	0.71	0.65	0.68
notApplicable	0.95	0.82	0.88
period	0.91	0.95	0.93
place	0.83	0.73	0.78
term	0.79	0.69	0.74
time	0.68	0.93	0.79
TOTAL	0.83	0.80	0.81

(b) Any overlap

Table 6.43.: Results of the rule-based Information Extraction in German for exact matches and any overlay (A = accuracy, P = precision, R = recall, F1 = F1-score)

Information	P	R	F1
address	0.00	0.00	0.00
allowed	0.67	0.50	0.57
applicable	1.00	0.50	0.67
costs	1.00	1.00	1.00
currency	1.00	0.67	0.80
language	1.00	1.00	1.00
link	1.00	1.00	1.00
name	1.00	1.00	1.00
notApplicable	0.00	0.00	0.00
period	0.76	0.93	0.84
place	1.00	0.14	0.25
time	1.00	0.43	0.60
TOTAL	0.87	0.61	0.72

(a) Exact matches

Information	P	R	F1
address	1.00	0.67	0.80
allowed	0.67	0.50	0.57
applicable	1.00	0.83	0.91
costs	1.00	1.00	1.00
currency	1.00	1.00	1.00
language	1.00	1.00	1.00
link	1.00	1.00	1.00
name	1.00	1.00	1.00
notApplicable	1.00	1.00	1.00
period	0.76	0.93	0.84
place	1.00	0.57	0.73
time	1.00	1.00	1.00
TOTAL	0.91	0.86	0.88

(b) Any overlap

Table 6.44.: Results of the rule-based Information Extraction in English for exact matches and any overlay (A = accuracy, P = precision, R = recall, F1 = F1-score)

While it is relatively easy to detect individual parts of such an address (e.g., with a regular expression for the zip code), it is almost impossible to reliably find the beginning and end based on a dependency tree.

Nevertheless, with regard to exact matches, in English, MITIE performs better for only three types of information (allowed, notApplicable, and time) and not better for any type of information with regard to partial matches. In German, MITIE performs better with regard to exact matches for seven types of information (address, age, applicable, name, notApplicable, place, and term) and three with regard to partial matches (age, applicable, and term). For these types, we can apply MITIE for extracting the necessary information while using the rule-based approach for the others.

6.5. Summary

In this chapter, we went through all stages of our NLP-pipeline and compared different approaches for each individual step, from the detection of relevant documents to the extraction of structured information. We also introduced a number of corpora in order to conduct these comparisons. Since this chapter included a lot of information, we shortly want to summarise the insights we gained.

6.5.1. Corpora

We created and applied the following corpora in the course of this chapter:

- **Detection and Segmentation Corpus:**
  A corpus of 15,986 web pages from online shops, 14,408 of which are German and 1,578 of which are English. The pages were crawled using price comparison websites and are annotated with a label that indicates whether the pages is a T&C page (about one-third of the corpus) or not.
- **Topic Classification Corpus:**
  A corpus of 172 T&C (142 German and 30 English, a sub-set of the detection and segmentation corpus), which manually have been segmented in clauses, which have been annotated with topics and subtopics according to our taxonomy. The corpus contains 7,156 topic labels and 4,621 sub-topic labels, which is a total of 11,777 labels.
- **Information Extraction Corpus:**
  A corpus of 2,199 clauses (1,819 German and 380 English, a subset of the topic classification corpus), which we manually annotated with extractable information. In total, the corpus is labeled with 2,831 instances of such information.

6.5.2. NLP-Pipeline

These are the results we gathered for the comparison of different approaches to different stages of the NLP pipeline:

- **Automatic Detection:**
  We compared four approaches for the automatic detection of T&C pages in online shops: rule-based analysis of URLs, rule-based analysis of link-texts, logistic regression on a bag-of-words model, and a multilingual transformer model. We found the rule-based URL classifier to perform best for this stage with an F1-score of 0.95 for German and 0.89 for English.

- **Content Extraction:**
  We compared four different approaches to the extraction of content from T&C pages: two different models implemented by the Boilerpipe library and the two libraries jusText, and Trafilatura. We found Trafilatura to work best in both languages, extracting exactly the correct content in 73% of all cases in German and 82% of all cases in English.

- **Sentence Segmentation:**
  In a comparison of three different libraries for sentence segmentation (NLTK, spaCy, and SoMaJo), SoMaJo was the only library that was able to extract all sentences correctly in English and German.

- **Paragraph Segmentation:**
  We presented a rule-based approach for paragraph segmentation that uses HTML-tags as well as the syntax of headings and their style to segment paragraphs and build a document hierarchy from them.

- **Topic Classification:**
  For the clause topic and subtopic classification, we compared permutations of seven classifiers (rule-based, logistic regression, random forest, MLP, CNN, LSTM, and a fine-tuned BERT language model) and six input encodings (Tf-idf, Word2Vec, GloVe, and domain-specific embeddings. We found that BERT performed best with an F1-score of 0.91 and 0.85 for German and English topic classification and 0.86 and 0.73 for German and English subtopic classification.

- **Information Extraction:**
  We compared the MITIE library with a rule-based approach that analysis dependency trees we implemented for the extraction of information from T&C. We found that, overall, our rule-based approach performed better, however, for some selected types of information, MITIE performs better. Therefore, we suggest to combine both approaches and decide, based on the information type, which one to use.

6.5.3. General Insights

Beyond insights on the different tasks, in the course of the experiments we conducted, we also gained some more general insights into the documents we are dealing with and the state-of-the-art in NLP:

- We were able to achieve very good results for complex tasks, like topic classification, while the available libraries for many allegedly easy tasks early in the pipeline (e.g., content extraction) proved to be far from perfect.
- There is very little work existing that tries to preserve and leverage the internal hierarchy of text documents.
- Composit nouns play a very important role in German T&C.
- While in theory, neural networks could leverage a hierarchical taxonomy and learn interdependencies between labels from different levels, in practice, the joined classification of topics and subtopics always decreased the classification performance on both in our experiments.
- Multilingual approaches bring little advantages, at least for the tasks we evaluated.
- For our classification task, the encoding of the input turned out the be more crucial than the applied classifier.

CHAPTER 7

Legal Assessment of Standard Form Contracts

After we have semantically analyzed the standard form contracts, we want to assess them legally in order to decide whether the clauses contained in them adhere to the existing legal regulations. We will call a clause "void" when our legal experts have reasonable doubt that, based on existing legislation and court rulings, a clause could be upheld if it would be challenged in a court of law. However, we can never be completely sure if that would be the case until a clause is actually tried. It is important to keep in mind that we want to build a system that assists consumer advocates, every clause that will be marked as "void" will be re-checked by an expert. Therefore, we also instructed our experts during the annotation (see Section 7.2) that, in doubt, we rather want to have one clause too much annotated as void than missing one.

Although the UK was still part of the EU when we collected our corpus, and therefore many of the regulations governing the drafting of consumer standard form contracts were based on common regulations between the UK and Germany, there were still nation-specific differences. Since we only worked together with experts for German law, we treated the English standard form contracts as if they were governed by German law, although this will most likely mean that we will label some clauses as void, which would not have been void under the law at the time we collected them and vice versa.

As we have pointed out before, the regulations for consumer standard form contracts are much stricter than regulations for business contracts. Some of the T&C in our corpus contain clauses for both consumers and businesses. Since we want to further the cause of consumer protection, and since our experts did not feel sufficiently confident to assess business clauses, we will only consider clauses that affect consumers in our legal assessment.

In order to automate the legal assessment, we will again use different approaches: a rule-based and an ML approach. For the rule-based approach, we first had to define which legal provisions we want to check for. Based on the relevance criteria we set out in Chapter 4. Based on these provisions, we then had to define rules which can check their adherence. For this, we build a legal knowledge-base that contains all the relevant information and is easy to maintain in case legal provisions change (see Section 7.3). For the ML approach, we treated the task as a binary classification problem, in which we classify each clause as either void or valid. We evaluated both approaches on a corpus that was annotated by the experts from the consumer protection agencies (see Section 7.2).

7.1. Relevant Legal Provisions

One of the drawbacks of a rule-based approach to the legal assessment is that we have to define beforehand which aspects of the T&C we want to assess. The decision is mainly based on the relevance of different aspects of T&C for consumers and consumer advocates, which we investigated in Chapter 4, but also on the technical feasibility of implementing them. Therefore, in consultation with the experts, we decided to focus on the following nine legal provisions.

- **Withdrawal Period:** According to §355 No. 2 BGB, consumers have the right to withdraw from distance contracts within fourteen days, without providing any reasons. There are only very few exceptions from this rule, which are listed in §312g BGB and include tailored products, perishable goods, and unsealed audio, video, and software volumes.
- **Withdrawal Form:** According to §355 No. 1 BGB the declaration of the intention to withdraw from a contract must be "unambiguous" ("eindeutig"), but there are no further formal requirements. In practice, this means that companies can not stipulate the form of withdrawal to, e.g., letters or emails, but have to accept all forms of textual communication and even phone calls[1] (Sommer and von Stumm, 2017, p. 475).
- **Warranty Period:** For new products, consumers always have a warranty ("Gewährleistung") period of at least two years and for used products at least one year, according to §476 No. 2 BGB.
- **Payment Fees:** Since 2018, online shop providers are not allowed to charge fees from consumers for payments made via bank transfer or payment cards like credit cards (§270a BGB). The Oberlandesgericht München (2019) ruled that this, however, does not apply to payment service providers like PayPal[2] or Sofortüberweisung, overruling a previous decision by the Landgericht München I.
- **Dunning Costs:** In case of a default in payment from the customer, many T&C include a flat fee for dunning costs, instead of calculating the actual costs. According to §309 No. 5 lit. a BGB, such flat fees are only allowed if they reflect the usually occurring costs. Particularly, it is not allowed to factor in overhead costs (like salaries and rents), but only actually occurring costs, which in the case of online shopping, are usually limited to printing costs and postage (Bundesgerichtshof, 2019). It is difficult to pinpoint which

[1] Consumer advocates, however, advise against using the telephone to withdraw from a contract, because it might be difficult for consumers to later proof they indeed withdrew in time from the contract.

[2] PayPal, however, forbids online shop providers in their own T&C to charge consumers a fee for using PayPal as payment method.

flat fee is permissible and which is not. In some cases, courts have ruled that fees of more than 1.20 EUR are too high (Oberlandesgericht München, 2011), while in other cases, fees of 2.50 EUR have been ruled as still permissible (Amtsgericht Brandenburg an der Havel, 2007). According to our experts, 2.50 EUR is a good rule of thumb, above which it is very likely that a flat fee for dunning costs is void. In addition, according to §309 No. 5 lit. b BGB, whenever a flat fee is used, the consumer always have to be given the option to prove that the costs that actuality occurred were lower, and this option has to be mentioned explicitly, if not, the clause is also void.

- **Default Interest:** In case of a default in payment from the customer, the seller is owed interest for the defaulted amount. According to §288 No. 1 BGB, the interest rate is set at five percentage points above the base interest rate. Generally speaking, a company would be allowed to set a higher interest rate in their T&C, however, the aforementioned §309 No. 5 BGB would also apply here, i.e., it would only be allowed to set a higher interest rate if that would reflect the actually occurring costs. Since, in general, interest rates are currently very low, virtually all clauses that set higher interest rates are usually void. (IWW Institut, 2011)
- **Delivery Time:** If information about delivery times are given in the T&C, they have to be "sufficiently concrete" (§308 No. 1 BGB). Somewhat ironically, "sufficiently concrete" itself is rather unspecific and one of the vague legal concepts we discussed in Section 2.3. However, courts have ruled that consumers can usually expect that their good will be shipped immediately, and if that is not the case, this information cannot be hidden in the T&C but has to be given on the product page (Landgericht Koblenz, 2019).
- **Individual Arrangements:** Although most companies do not want to deviate from their original T&C, because it is the very idea and most important benefit of a standard form contract that it is a "one size fits all" solution, it is possible for the consumer to make individual arrangements with the seller, which deviate from the T&C. The legislator gives such individual arrangements always priority over the standard form contract, independent from their form. A clause that makes it mandatory for such individual agreements to be in written form is therefore void. (Basedow, 2019, Rn. 13-20)
- **Online Dispute Resolution:** As we have already mentioned in Section 6.4.1, online shops need to provide a (clickable) link to the ODR platform of the EU (Landgericht Dortmund, 2020), if they do not provide such a link, the clause is void.

While we believe that these provisions cover a major share of the actually occurring void clauses that are relevant for consumers, they are by no means exhaustive. Most notably, they do not cover the clauses that fall under §305c BGB, because they are "'surprising or ambiguous". As we have mentioned in Section 3.1.1, such clauses are not the focus of our work, however, we do consider them with our second, ML-based approach.

Additionally, we have identified three information obligations that operators of online shops have towards their customers. They differ from the legal provisions above since they do not originate from the laws governing standard form contracts and are, therefore, as our experts have told us, not usually a part of their assessment of T&C. These three information obligations, which are listed below, have to be fulfilled, however, they do not necessarily have to be fulfilled as part of the T&C. In theory, the shop operator could also fulfill them by sending a separate email to the customer. In practice, we believe it is very unlikely that they are fulfilled if they are not part of the T&C. However, our experts still preferred to exclude these information obligations.

- **Online Dispute Resolution:** The consumer has the be informed about the aforementioned ODR platform of the EU (European Parliament and Council of the European Union, 2013).
- **Storage of the Contract Text:** Based on §312i BGB, the online shop provider has to inform the consumer in which way the contract text will be stored.
- **Contract Language:** The same regulation as for the storage of the contract text also applies to the contract language, i.e., the shop owner has to inform the consumer in which languages he can conclude the contract.

Technically speaking, it would be very easy to detect if these information obligations are fulfilled since our taxonomy contains separate topic classes for these three points. If a clause of the related topic is present, the information obligation is most likely fulfilled if no such clause is present, the information obligation is most likely not fulfilled.

7.2. Data Corpus

Building a corpus to train and develop rule-based and ML-based approaches for the automated legal assessment of T&C is far from trivial. We need a sufficient number of void clauses, especially for the cases described in Section 7.1. At the same time, we also need T&C that are labeled completely in order get an approximation of the distribution of void clauses in the "real world" and in order to be able to detect missing clauses. However, if we would only use completely labeled T&C, we would need thousands of contracts to find a sufficient number of void clauses.

7.2.1. Sources

We decided to combine three approaches for gathering data for our German corpus:

- We took 78 clauses from the JURA database, described in Section 5.2, where the consumer protection agencies document clauses that have been successfully challenged legally. This means that all of the clauses from this source are void.
- From our previous corpus, we selected clauses from a set of relevant topics (e.g., all clauses about late payments or payment fees) and in addition, the experts specifically searched for void clauses from these topics on the internet. In total, this amounted to 140 clauses.
- We took 24 complete T&C which together consisted of 968 clauses.

Overall, the German corpus consists of 1,186 clauses. Since our experts did not feel comfortable legally assessing English texts, we provided them with translations of English T&C. This made the process even more laborious and also influenced the validity of the results. Therefore we decided to keep the English corpus rather small and instead focus on the German contracts. Additionally, we also did not have a source for already challenged void clauses, as we did for German. Therefore, we decided to only use one approach to gather data for the English corpus by having our experts annotating ten complete T&C which included 193 clauses.

7.2.2. Annotation

For the annotation of the previous corpora, where we mainly annotated clause topics, we could reasonably let non-experts do the annotation in the first stage, based on a taxonomy that we developed together with the experts, and just present conflicting cases to the experts, in order to optimize the usage of their valuable time. However, for this corpus, we need to assess each clause legally, which is a task that can only be performed by experts. Therefore, we had each clause labeled independently by two experts with "void" or "valid". We also asked the experts to shortly justify their assessment in a commentary. We then compared the annotations and provided the experts with a list of the conflicting annotations, which they then resolved together by agreeing on one common assessment. The 78 clauses which were extracted from the JURA database were not labeled again because they already have been classified as void by successful legal proceedings.

For the other clauses, we found the old prejudice of "two lawyers, three opinions" to carry a certain amount of truth. The inter-annotator agreement (before the resolution phase) was between 76% (for the annotation of complete T&C) and 64% (for the annotation of the pre-selected clauses).

For the English corpus, we would like to point out again the already mentioned limitations of the annotation process: Since we did not have access to experts for British law and also no English speaking legal experts, the annotation is based on German law, and the experts made their assessment based on a translated version of the clauses. These limitations might raise the question of why at all use such a corpus, however, we believe that, despite its limitations, the corpus is still useful to evaluate whether the technologies we use can also be applied to English texts. And at least two of the contracts in the English corpus specifies that "The Law of the Federal Republic of Germany applies exclusively."

7.2.3. Analysis

Table 7.1 shows how many clauses of each topic are in both corpora and how many clauses of each topic in the corpora are void. When looking at these figures, there are two things one should keep in mind: a clause can belong to multiple topics, therefore, the sum of the counts will be larger than the actual number of clauses, and for some topics, we actively searched for void clauses in German. The fact that more than 41% of all payment clauses were void, but just about 12% of all delivery clauses, gives no indication about whether payment clauses are generally more likely to be void.

Topic	#clauses	#void
age	12	0
applicability	22	1
applicableLaw	12	1
arbitration	18	1
changes	3	0
conclusionOfContract	135	8
delivery	117	14
description	8	0
disposal	8	0
intellectualProperty	21	0
language	9	1
liability	99	43
party	26	0
payment	305	126
personalData	64	1
placeOfJurisdiction	11	2
prices	38	9
retentionOfTitle	26	4
severability	13	6
textStorage	10	0
warranty	43	9
withdrawal	209	26

(a) German

Topic	#clauses	#void
applicability	5	0
applicableLaw	5	0
arbitration	5	0
changes	1	0
conclusionOfContract	27	0
delivery	34	1
description	4	0
disposal	2	0
intellectualProperty	9	0
language	1	0
liability	21	2
party	3	0
payment	16	0
personalData	11	0
placeOfJurisdiction	3	0
prices	12	0
retentionOfTitle	4	0
severability	1	0
textStorage	3	0
warranty	4	0
withdrawal	41	4

(b) English

Table 7.1.: Distribution of clause topics and void clauses in the corpora

Therefore, we want to focus only on data from T&C that were annotated completely for a moment. In German, the experts annotated 24 complete T&C, in these 24 T&C, they found 73 void clauses, about three clauses per contract. In English, the experts annotated ten complete T&C and found seven void clauses, less than one per contract. However, the German contracts also consist of 50 clauses per contract on average, while the English contracts on average only consist of 19 clauses.[3] If we count the ratio of void clauses to total clauses, we find that in the German T&C, about 6% of all clauses are void, while in the English T&C it is about 4% of all clauses. Even our experts were surprised that the ratio of void clauses is that high. They said they never before analyzed all aspects of such a large number of T&C and would not have expected to find so many void clauses, and also decided to take actions about some of the clauses they found during the annotation process. So at least at this very small level, our work already had an impact and helped to protect consumers better. We should also note that the data in Table 7.1 only covers clauses that were present and void. The fact that the German corpus consists of 24 T&C, but we found only 18 arbitration clauses implies that six companies most likely did not fulfill there legal obligation to inform consumers about the EU ODR platform.

[3] This is in line with our previous corpus, where we also found that the English T&C were on average much shorter than the German T&C (see Section 6.3.1).

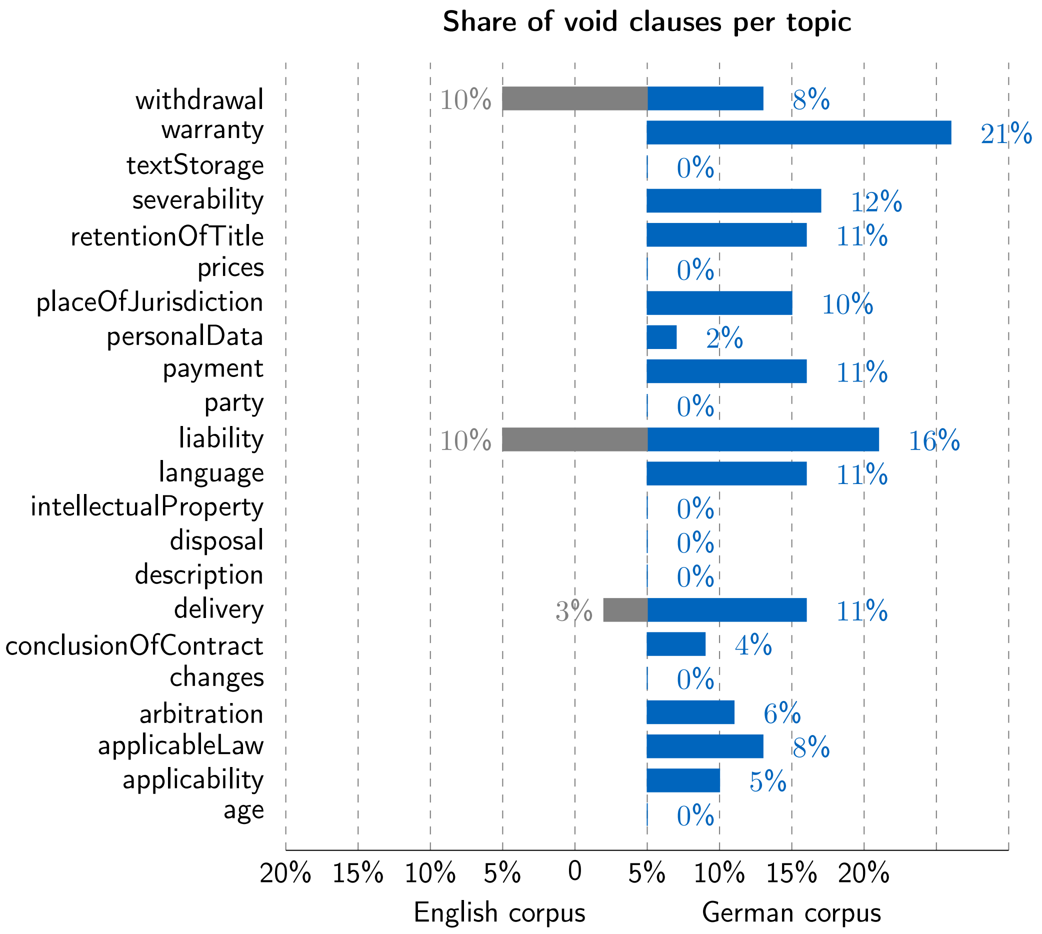


Figure 7.1.: Share of void clauses per topic in the German (right) and English (left) corpus

Figure 7.1 shows how many percent of all clauses were void for all topics in both corpora. In the German corpus, warranty clauses were most frequently void. The reason in most cases were illegal deadlines that were set by companies until which consumers should claim damages. Second most frequently, liability clauses were void because companies illegally tried to limit their liability for certain damages, a pattern that was also frequently found in the English corpus. Other frequently occurring reasons for clauses to be void include: too high default interest and dunning costs, too unspecific information about delivery times, illegal exclusion of consumer protection regulations from the jurisdiction of the consumer, and illegal replacement clauses for void clauses.

7.3. Knowledge-base

We wanted to build a knowledge-base that does not only contain the necessary legal knowledge outlined in Section 7.1, but also all the other legal and linguistic knowledge that is necessary for the rule-based analysis of standard form contracts, from the taxonomy we developed and

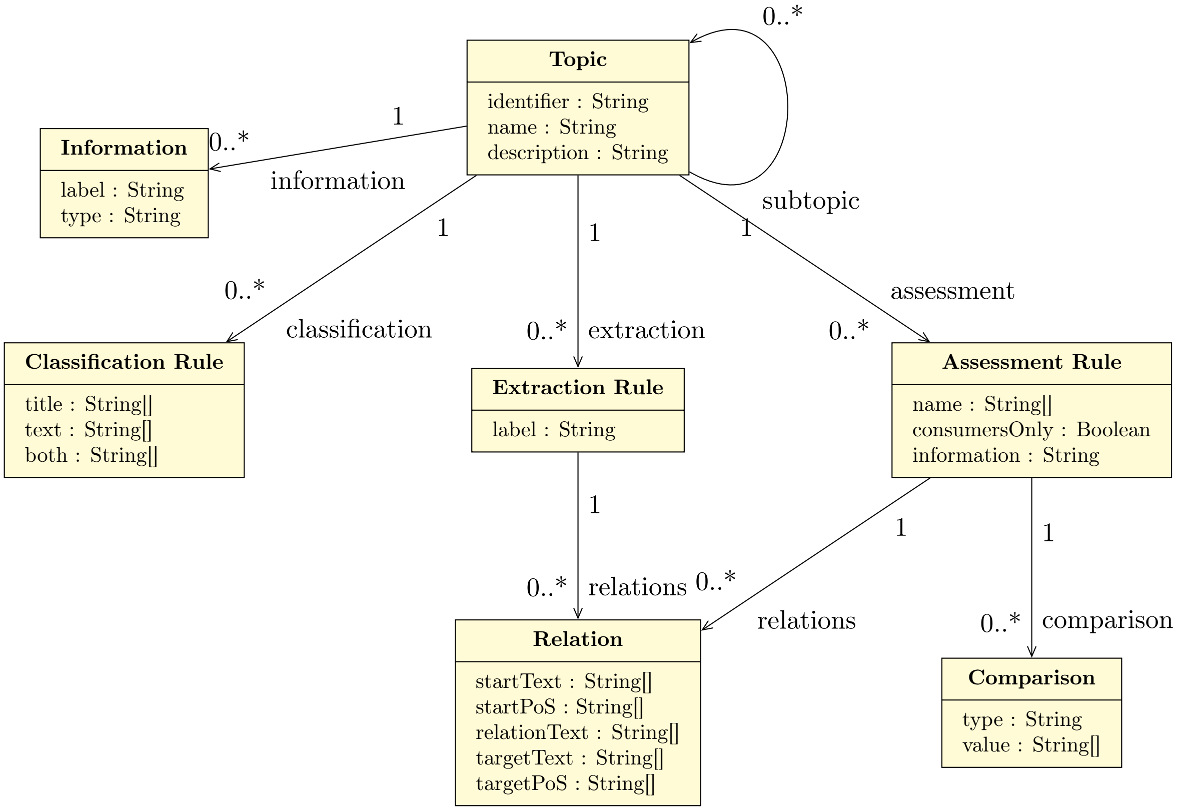


Figure 7.2.: Data schema of the knowledge-base

its hierarchy, to the keywords used for the topic classification, the rules for the information extraction, and finally, of course also the rules for the legal assessment. As we already concluded in our analysis of the existing formats for the representation of knowledge in the legal domain in Section 3.2, none of the existing formats is flexible enough to map these requirements, which are less driven by legal theory and more by practical applicability. Therefore, we decided to use the JSON-format to store our knowledge-base, which provides the necessary flexibility but also the compatibility that allows easy and direct integration of the knowledge-base into different kinds of NLP-pipelines. In Sections 7.3.1 to 7.3.4, we describe the format of the knowledge-base using examples and the JSON Schema description language (Pezoa et al., 2016). Figure 7.2 shows a graphical representation of the data schema of the knowledge-base in UML.

7.3.1. Taxonomy

The taxonomy and the topics it contains are the main building blocks of our knowledge-base because all other information and rules are connected to the topic they belong to. Listing 7.1 shows how a topic is represented in JSON format. Each topic has an identifier, which we already have been using throughout this thesis, as well as a name and a description. A topic can also have a number of subtopics. Additionally, the topic object holds references to all the information that will follow in the next sections: the topic classification rules, the extraction rules, and the rules for the legal assessment. The object also has a field called `information`, which, as we will see in Section 7.3.3, holds a list of information that can be extracted for the topic. This information can be used in both the extraction and assessment rules.

```
{
    "identifier": "payment",
    "name": "Payment",
    "description": "Clauses covering payments.",
    "information": [],
    "classificationRules": [],
    "extractionRules": [],
    "assessmentRules": [],
    "subtopics": []
}
```

Listing 7.1: JSON example for topics

Listing 7.2 shows the respective JSON schema. Most notably, the array of subtopics consists of topic objects (i.e., a recursive definition), which means that we can not just construct taxonomies with two levels (topic and subtopic), like the one we currently use, but can, theoretically, create taxonomies of arbitrary depth.

```
{
    "$id": "topic",
    "description": "Topic object",
    "type": "object",
    "required": [
        "identifier",
        "name",
        "description"
    ],
    "properties": {
        "identifier": {
                "$id": "#/properties/identifier",
                "type": "string"
        },
        "name": {
                "$id": "#/properties/name",
                "type": "string"
        },
        "description": {
                "$id": "#/properties/description",
                "type": "string"
        },
        "information": {
                "$id": "#/properties/information",
                "type": "array",
                "items": { "$ref": "information" }
        },
        "classificationRules": {
                "$id": "#/properties/classificationRules",
                "type": "array",
                "items": { "$ref": "classificationRule" }
        },
        "extractionRules": {
                "$id": "#/properties/extractionRules",
                "type": "array",
                "items": { "$ref": "extractionRule" }
        },
        "assessmentRules": {
                "$id": "#/properties/assessmentRules",
                "type": "array",
                "items": { "$ref": "assessmentRule" }
        },
        "subtopics": {
                "$id": "#/properties/subtopics",
                "type": "array",
                "items": { "$ref": "topic" }
        }
    }
}
```

Listing 7.2: JSON schema for topics

7.3.2. Topic Classification

As we described in Section 6.3.3, the rules we use for the classification of (sub-) topics, three operators: “and”, “or”, and “negation”. Already then, we decided that, because we want consumer advocates to be able to define their own topics and rules, we do not want to use a powerful but complex language like regular expressions, but rather something more simple. A rule can also either be applied to the title of a clause, the text of a clause, or both, and that also has to be reflected in the knowledge base. Listing 7.3 shows how we encode the information in JSON, using three fields of string arrays. We represent “or” with different elements of an array, i.e., Listing 7.3 shows a rule that checks whether the words withdrawal **or** cancellation are in the title. We represent “and” with words in one string that are separated by a white-space and negations with an exclamation mark. So for the text of a clause, we see a rule in Listing 7.3 that checks if the word “withdraw” is present **and** the word “cash” is **not** present.

This means that we cannot nest “or”-statements in other “or”-statements‘: `A or (B and (C or D))`, for example, can not be expressed in this way. While this limits the comfort in drafting expressions, it does not limit the expressiveness because we can rewrite the statement to `A or (B and C) or (B and D)`.

```
{
    "title": ["withdrawal", "cancellation"],
    "text": ["withdraw !cash"],
    "both": []
}
```

Listing 7.3: JSON example for topic classification rules

Listing 7.4 shows the corresponding JSON schema, which is referenced by the topic JSON schema.

```
{
    "$id": "classificationRule",
    "description": "Rules for the classification of clause topics",
    "type": "object",
    "properties": {
        "title": {
                "$id": "#/properties/title",
                "type": "array"
        },
        "text": {
                "$id": "#/properties/text",
                "type": "array"
        },
        "both": {
                "$id": "#/properties/both",
                "type": "array"
        }
    }
}
```

Listing 7.4: JSON schema for topic classification rules

7.3.3. Information Extraction

In Section 6.4.3, we described how the rules for the information extraction are constructed. At the heart of the rules is the way in which we can specify dependency relations with them. Either by the text or PoS of the start node of the relation, or by the dependency type of the relation, or by the text or PoS of the end note of the relation (or any combination of those). Because crafting these rules needs linguistic knowledge and can not be done by the consumer advocated without assistance, we decided to use regular expressions for the five aspects in order to make the rules as expressive as possible. Listing 7.5 shows the schema for the representation of such rules in JSON.

```
{
    "$id": "relation",
    "description": "Description of a dependency relation or a set of
    ↪ dependency relations",
    "type": "object",
    "properties": {
        "startText": {
                "$id": "#/properties/startText",
                "type": "array",
                "items": { "type": "string" }
        },
        "startPoS": {
                "$id": "#/properties/startText",
                "type": "array",
                "items": { "type": "string" }
        },
        "relation": {
                "$id": "#/properties/relation",
                "type": "array",
                "items": { "type": "string" }
        },
        "targetText": {
                "$id": "#/properties/targetText",
                "type": "array",
                "items": { "type": "string" }
        },
        "targetPoS": {
                "$id": "#/properties/targetText",
                "type": "array",
                "items": { "type": "string" }
        },
    }
}
```

Listing 7.5: JSON schema for dependency relation descriptions

All that the parent element that carries the rules has to encode is whether the rule is part of the rules that define the sub-graph which has to be extracted, mandatory relations that have to be present but are not extracted, or a relation that should not be present (see also Section 6.4.3) and which label the information should be labeled with, once they are extracted.

```
{
    "$id": "extractionRule",
    "description": "Description of a set of rules for the extraction of
    ↪ information",
    "type": "object",
    "required": [ "label" ],
    "properties": {
        "label": {
                "$id": "#/properties/label",
                "type": "string"
        },
        "extractionRelations": {
                "$id": "#/properties/extractionRelations",
                "type": "array",
                "items": { "$ref": "relation" }
        },
        "existingRelations": {
                "$id": "#/properties/existingRelations",
                "type": "array",
                "items": { "$ref": "relation" }
        },
        "excludedRelations": {
                "$id": "#/properties/excludedRelations",
                "type": "array",
                "items": { "$ref": "relation" }
        }
    }
}
```

Listing 7.6: JSON schema for information extraction rules

A complete entry for an extraction rule in the knowledge-base could look like the example given in Listing 7.7, which shows a rule for the extraction of time periods: The start node is one of the nouns “year”, “month”, “week”, and “day”, and the target node is a cardinal (“CD”), which is connected by a modifier relation. No other relations need to be present or absent.

```
{
    "label": "period",
    "extractionRelations": [
        {
            "startText": ["year", "month", "week", "day"],
            "startPoS": ["NN"],
            "relation": ["mod"],
            "targetText": [],
            "targetPoS": ["CD"]
        }
    ],
    "existingRelations": [],
    "excludedRelations": [],
}
```

Listing 7.7: Example for an information extraction rule

In order to make the extracted information directly usable for the legal assessment (see Section 7.3.4), we use `information` objects to model which information can potentially be extracted from a given topic. In addition to the name of the label under which the information will be extracted, the object also holds information about the types of the extracted information, which will also be used for the legal assessment. The rule in Listing 7.7, for example, extracts a number together with a unit under the label "period". Listing 7.8 shows how we would encode this information in a JSON object.

```
{
    "label": "period",
    "type": "numberWithUnit"
}
```

Listing 7.8: Example for an information object

The extraction for payment fees, for example, also extracts information of the same types, only the unit, in this case, does not describe a unit of time, but a currency (i.e., a unit of money). The only other type we currently use for the rule-based legal assessment of clauses are enumerations, where a list of strings is extracted. These can be used very differently, as we will show in the next section. The JSON schema for the information object is shown in Listing 7.9.

```
{
    "$id": "information",
    "description": "Description of the information that can be extracted from
    ↪  clauses of a given topic",
    "type": "object",
    "required": [
        "label",
        "type"
    ],
    "properties": {
        "label": {
                "$id": "#/properties/label",
                "type": "string"
        },
        "type": {
                "$id": "#/properties/type",
                "type": "string"
        }
    }
}
```

Listing 7.9: JSON schema for an information object

7.3.4. Legal Assessment

The legal assessment of clauses is mainly based on the information that is extracted by the rules described in the previous section. However, in some cases, these rules alone are not sufficient, and we need some additional information to decide which assessment rules to apply. For all rules, we need to make sure not to apply them to clauses that only apply to companies because the legal provisions we base our assessment on only apply to consumers. In some cases, we also might find two different values, e.g., in the case of the warranty period, which may be different for new goods and used goods. For such cases, our legal assessment rules need additional information. In order to store these additional rules in the knowledge-base, we can use the “relation” type, introduced in Listing 7.5. Similar to the extraction rules schema, we might have relations that have to be present, i.e., existing relations, and relations that should not be present, i.e., excluded relations. Since the distinction of whether a clause only applies to companies is so universal and applies to almost all rules, instead of just manually adding it each time as an excluded relation, we add a binary switch to the schema for legal assessment rules, whether to check for this relation or not.

The complete schema is shown in Listing 7.10. Each legal assessment rule has a name, which we can later use (e.g., during the summary generation which we describe in Chapter 8) to refer to the assessment rule and its results. The “information” field defines which extracted information to use for the assessment by referring to the label field of the information object. Each assessment rule also contains an array of objects from the type “comparison”. These comparison objects are the legal regulations that will be encoded in a machine-readable way.

```
{
    "$id": "legalAssessmentRule",
    "description": "Description of a rule for the legal assessment of a
   ↪ clause",
    "type": "object",
    "required": [
        "name",
        "consumersOnly",
        "comparison"
    ],
    "properties": {
        "name": {
                "$id": "#/properties/name",
                "type": "string"
        },
        "consumersOnly": {
                "$id": "#/properties/consumersOnly",
                "type": "boolean"
        },
        "information": {
                "$id": "#/properties/information",
                "type": "string"
        },
        "existingRelations": {
                "$id": "#/properties/existingRelations",
                "type": "array",
                "items": { "$ref": "relation" }
        },
        "excludedRelations": {
                "$id": "#/properties/excludedRelations",
                "type": "array",
                "items": { "$ref": "relation" }
        },
        "comparison": {
                "$id": "#/properties/comparison",
                "type": "array",
                "items": { "$ref": "comparison" }
        }
    }
}
```

Listing 7.10: JSON schema for a legal assessment rule

As we described in the previous section, the information we extract can belong to two types, enumerations or numbers with units. These types set the basis for our comparison. For numbers with values, we can define the six typical mathematical comparisons: less ($<$), less or equal ($\leq$), equal ($=$), unequal ($\neq$), greater or equal ($\geq$), and greater ($>$). Since we can add multiple comparisons to a legal assessment rule, this also allows testing whether a number is in a certain range by using two of these comparisons in one rule. Listing 7.11 shows how a comparison object inside a rule for the legal assessment of clauses about the withdrawal period would look like, if the period is equal to 14 days or more, the clause is legal.

```
{
    "type": ">=",
    "value": ["14 days"]
}
```

Listing 7.11: JSON comparison object for the legal assessment of withdrawal clauses

For enumerations, we defined four comparison types: contains, does not contain, equal, and unequal. For the legal assessment of clauses about the form of withdrawal, for example, we have to make sure that, e.g., if mandatory forms are defined, “Schriftform” (written form) is not one of them and if forms are excluded “Textform” (text form) is not one of them. Listing 7.12 shows how such a rule would look like in our knowledge-base.

```
{
    "name": "forbiddenWithdrawalForms",
    "consumersOnly": true,
    "information": "forbidden",
    "comparison": [
        {
            "type": "notContains",
            "value": ["text form"]
        }
    ]
}
```

Listing 7.12: JSON object for a legal assessment rule for forbidden forms of withdrawal

There is one additional type of comparison that we need for our legal assessment, for the special cases where we do not extract any information but want to make sure that a certain clause is present at all, e.g., to check whether there is a clause about the storage of the contract text. For this, we simply use a comparison type “existing” without an “information” in the assessment rule object and without a value in the comparison object, as shown in Listing 7.13. Since the legal assessment rules are stored within the topic object, we also do not need an explicit reference to the clause topic. In the future, we might also use a type “nonExisting” if we want to make sure that certain clauses are not present in a contract.

```
{
    "name": "contractText",
    "consumersOnly": true,
    "comparison": [
        {
            "type": "existing"
        }
    ]
}
```

Listing 7.13: JSON object for a legal assessment rule for contract text storage

Listing 7.14 shows the JSON schema for comparison objects that is able to encode all of the above-mentioned cases.

```
{
    "$id": "comparison",
    "description": "Description of a comparison rule for the legal assessment
    ↪  of clauses",
    "type": "object",
    "required": [
        "type"
    ],
    "properties": {
        "type": {
                "$id": "#/properties/type",
                "type": "string"
        },
        "value": {
                "$id": "#/properties/value",
                "type": "array",
                "items": { "type": "string" }
        }
    }
}
```

Listing 7.14: JSON schema for comparison objects

7.4. Approaches

In the following two sections, we will describe both approaches we evaluated for the legal assessment of clauses, starting with the rule-based approach, followed by the ML approach.

7.4.1. Rule-based

For each of the legal provisions outlined in Section 7.1, we drafted rules using the format described in Section 7.3.4. For most provisions, one rule was sufficient, however, for some, two rules were needed. In the following enumeration, we will briefly describe the rules we drafted for each of the nine provisions.

- **Withdrawal Period:** The rule for the assessment of clauses which set the withdrawal period is one of the easiest rules because all we have to do is check whether the information extracted with the label "period" from clauses with the subtopic "withdrawal:period" is larger or equal to fourteen. A smaller technical challenge that might occur is that we might have to convert written-out numerals (like "fourteen") into numbers, but we can easily do that by stemming them and then matching them against a list. Another challenge is that we might have to convert between units. If we would just compare if the value is larger or equal to fourteen, we would mark "two weeks" as void, although it is not. The conversion between weeks and days is straightforward; if we find the unit "week", we multiply the numeral with seven because one week always has seven days. The situation becomes less obvious when we talk about months or years. If we would, for example, check for a period that has to be at least 30 days, a clause granting a period of "one month" could be void. A period of one month, according to the German law (§188 BGB), ends with the day in the next month that has the same number as the start of the period. If a period of one month starts, e.g., on the 15th of November, it ends on the 15th of December. If the next month does not have that many days, the period ends on the last day of the month. If a period starts, e.g., at the 31st of January, it ends at the 28th of February (or 29th in case of a leap year) and therefore after less than 30 days. To avoid such problems when converting from months or years to days, we use lower bounds, i.e., we calculate one month with 28 days and one year with 365 days.
- **Withdrawal Form:** As we outlined in Section 7.1, a company basically has to accept all forms of withdrawal, as long as they form a "unambiguous" statement. Therefore, the only "mandatory" label that may occur in a clause of the subtopic "withdrawal:form" without

the clause being void is an (unambiguous) declaration of will/intent. Analogous, almost nothing can be labeled as “forbidden”, except for “sending back goods”, which a company can specifically exclude without the clause being void.

- **Warranty Period:** In order to check clauses about the warranty period, we need two rules. And for the first time, we can not just rely on the extracted information. For used goods, the warranty period has to be at least 12 months. For new goods, it has to be at least 24 months. We can easily distinguish these two cases by using the excluded and existing relations field of the data model for our rules: For the rule for used goods, a relation to “used” (“gebraucht” in German) has to exist, for the rule for new goods, a relation to “used” has to be excluded.
- **Payment Fees:** For payment fees, we just want to check fees for bank transfer or payment cards. Therefore, as soon as a label “fee” is extracted from a clause with the subtopic “payment:fee”, which is larger than zero, and a relation exists to “bank transfer”, “credit card”, “visa” or “master card” (in German “SEPA”, “Überweisung”, “Lastschrift”, “Kreditkarte”, “visa”, “master card”), the clause is void.
- **Dunning Costs:** The rule for the assessment of dunning costs is very easy again, the information extracted with the label “fee” from clauses of the subtopic “payment:late” has to be below or equal to 2.50 EUR.
- **Default Interest:** Similarly, for interest rates, the information with the label “interest” from clauses with the same subtopic has to be below or equal to five.
- **Delivery Time:** For the delivery time, we want to check whether the given times are too vague. Therefore if we can extract a piece of information with the label “time”, the clause is most likely not void. However, if we do not find such a piece of information but a relation to a vague word like “approximately”, “several” or “few” (or in German “ungefähr”, “zirka”, “einige”, “wenige”), we will label the clause as void.
- **Individual Arrangements:** The assessment of clauses about individual arrangements is the only instance in which we do not use any extracted information at all, since we do not even extract any information from the relevant clause (topic “changes”). Instead, we simply check whether the phrases “written form” or “text form” (“Schriftform” or “Textform”) appear in the clause, which would make the clause void
- **Online Dispute Resolution:** If an arbitration clause does not contain a link, it is automatically marked as void. Additionally, we check whether at least one of the links in the clause points to the ODR platform of the EU (`www.ec.europa.eu/consumers/odr`).

7.4.2. BERT

Because BERT proved to be most successful in classifying the topics and subtopics of clauses, we decided to also apply it for the legal assessment of the clauses. As we have mentioned in the introduction to this chapter, from a technical perspective, we are looking at a binary classification problem, as we try to find out whether a clause is void (1) or valid (0). The general advantage of this stochastic approach is that we are not limited to certain types of void clauses, as a drawback, however, we also will not be able to provide any further explanation, why a certain assessment was made.

As usual, we split our corpus into a training (80%) and a test set (20%) and first perform a stratified five-fold cross-validation on the training set to identify the best performing hyperparameters. Unlike for the topic classification (see Section 6.3.9), we found that this time, hyperparameters which were suggested in the original BERT paper by Devlin et al. (2019) performed best in German: batch size 16, learning rate 3e-5, and number of epochs 3. The English corpus contains so few void clauses (seven) that there is no sensible way in which we could have applied this approach to it. Therefore, we applied BERT only to the German corpus.

7.5. Evaluation

In the following two sections, we describe the results of the evaluation we performed with both approaches, using the corpus described in Section 7.2.

7.5.1. Rule-based

In order to evaluate the rule-based approach, it only makes sense to look at clauses that can actually be assessed by the rules, i.e., clauses that fall under one of the aspects covered by the legal provisions described in Section 7.1. The following topics and subtopics fall in this category and are therefore used for the evaluation:

- arbitration
- changes
- delivery:time
- language
- payment:fee
- payment:late
- textStorage
- warranty:period
- withdrawal:form
- withdrawal:period

In the part of the German corpus which only consists of the clauses from completely annotated T&C, these topics and subtopics cover 144 of 968 (around 15%) and 22 of the 73 void clauses (30%). In the English corpus, the share is 17% and 57%, respectively. Even if our rules would work perfectly, we would still only be able to assess around 15% of the contracts, which is a significant limitation of this approach.

Topic	TP	TN	FP	FN	A	P	R	F1
arbitration	1	17	0	0	1.00	1.00	1.00	1.00
changes	0	3	0	0	1.00	n.a.	n.a.	n.a.
delivery:time	4	20	7	2	0.73	0.36	0.67	0.47
payment:fee	18	1	0	17	0.53	1.00	0.51	0.68
payment:late	60	41	2	24	0.80	0.97	0.71	0.82
warranty:period	0	9	0	0	1.00	n.a.	n.a.	n.a.
withdrawal:form	2	19	3	0	0.88	0.40	1.00	0.57
withdrawal:period	0	30	0	1	0.97	n.a.	0.00	n.a.
TOTAL	85	140	12	44	0.80	0.88	0.66	0.75

Table 7.2.: Results of the evaluation of the rule-based legal assessment of German clauses (A = accuracy, P = precision, R = recall, F1 = F1-score)

Topic	TP	TN	FP	FN	A	P	R	F1
arbitration	0	5	0	0	1.00	n.a.	n.a.	n.a.
changes	0	1	0	0	1.00	n.a.	n.a.	n.a.
delivery:time	1	7	0	0	1.00	1.00	1.00	1.00
withdrawal:form	1	6	0	0	1.00	1.00	1.00	1.00
withdrawal:period	1	7	0	2	0.80	1.00	0.33	0.50
TOTAL	3	26	0	2	0.94	1.00	0.60	0.75

Table 7.3.: Results of the evaluation of the rule-based legal assessment of English clauses (A = accuracy, P = precision, R = recall, F1 = F1-score)

Table 7.2 shows the result of the evaluation on the German corpus. While the overall precision is quite high, with 0.88, the recall is considerably lower with 0.66, and we have to keep in mind that this is already just a subset of about 15% of all clauses. So the rules cover only a very small fraction of all clauses, on these, however, they work quite well.

The results do not necessarily represent an evaluation of the quality of the rules: Since a clause can have multiple topics, a false negative in the evaluation does not necessarily mean that the rule overlooked something, it can also mean that the clause was void for a reason that is simply not covered by the rules. We saw that frequently for payment clauses, where we have a relatively high number of false negatives.

Table 7.3 shows the results for the evaluation on the English corpus. Here too, we see a very high precision but a relatively low recall. However, given the small size of the corpus, these results should not be over overvalued.

7.5.2. BERT

With the hyper-parameters described in Section 7.4.2, we evaluated the approach on our test data set. BERT performed very well in the classification of void clauses and achieved an accuracy of 0.90, an F1-score of 0.89, a precision of 0.90, and a recall of 0.90. At first, this might seem like a surprisingly good performance because, from a human perspective, the task of legally assessing clauses is more complex than the task of clause topic classification, in which we achieved similar results (F1 = 0.91). Mathematically speaking, however, the legal assessment is the less complex task: Instead of a multi-label multi-class classification task with 23 possible classes, we are now looking at a binary classification task in which each clause is assigned exactly one value, which can be either 1 or 0. And while the data-set for the clause topic classification was much larger in total (5,020 compared to 1,186 clauses) than the current data-set, the amount of training data that is available per class was smaller for the clause topic classification. A deeper analysis of the results also shows that some clause types, like clauses that define automatic price increases for subscriptions, are literally always void, while other clause types, like clauses that set fixed dunning costs, were virtually always void. Which also “simplifies” the assessment problem.

Unlike the rule-based approach, this binary-classification approach does not take into account the two previous steps of our pipeline, i.e., it neither uses the extracted information nor the results of the topic classification. Therefore, we also refrained from providing results based on individual topics because the classifier is not aware of the taxonomy we are using.

CHAPTER 8

Summarization of Standard Form Contracts

We now have an NLP-pipeline, which can detect, extract, separate, analyze, and assess standard form contracts. This chapter describes an NLG-component that can be appended to this pipeline.

This component serves two purposes: It can summarize the content of standard form contracts in a simplified language. This is particularly useful to consumers because, as we have shown in Section 4.2, two of the three main reasons why consumers do not read T&C are that they are too long and too difficult to understand. With a simplified summarization, we can tackle both of these problems. Additionally, the component can not just be used to summarize the T&C themselves, but it can also be used to summarize the results of the legal assessment of them. Such a summarization can also contain explanatory elements, which help to justify the automated assessment. This can be particularly useful for expert users. Many studies (Ye and Johnson, 1995; Teach and Shortliffe, 1981; Weiner, 1980) have shown that experts are more likely to trust the automatic assessment of an expert system if they explain or justify their results in some way (see also Braun et al. (2018b) and Braun and Matthes (2018) for more on the influence of explanations in expert systems).

8.1. Approach

In Section 3.3, we have presented an extensive overview of the different approaches to text summarization. For the outlined purposes, only abstractive summarizations are applicable. For consumers, extractive summarizations could solve the problem of contracts being too long, but not the problem of contracts being difficult to understand. For experts, extractive summarization can provide some form of additional explanation by providing the clause on which an assessment is based, but they can give no further explanation about what specifically the algorithm found in this clause.

From a technical perspective, the main two approaches are rule-based systems and data-driven (i.e., ML) systems. One of the main problems of data-driven NLG systems is so-called "hallucination". Hallucination describes an error in the text generation in which text is generated that is not backed by the input, whether the input is data, a text, or an image. An example for such an error would be an NLG system that generates from the information that a café is located at King Street the sentence "There is a cheap café in King Street". The price information "cheap" is not backed by the data and would hence be a hallucination. Such hallucinations are usually caused by frequent co-occurrence in the training data. If the word "café" usually co-occurs with "cheap" in the training data, the system will learn that they belong together. Although omitting data is a more frequently occurring problem, hallucination is still a significant problem even in state-of-the-art systems. (Nie et al., 2019; Dušek et al., 2020) Although less frequent, hallucination is usually the more severe type of error, especially in the legal domain. Therefore, we decided to develop a rule-based system.

Since we already have a pipeline that transforms the textual representation of a standard form contract into structured data, by extracting clauses, classifying their topics, extracting information from them, and assessing them legally, we are dealing with a data-to-text problem and not a text-to-text problem. However, this also limits us to using the legal assessment that was produced by the rule-based approach because only the rule-based approach uses the extracted information for the legal assessment, while the ML approach performs a binary classification on the clause text, that is independent of the extracted information, and we, therefore, do not know why a clause was deemed to be void.

8.2. Pipeline

Our NLG component follows the classical pipeline as described by Reiter and Dale (1997, 2000): In the first step, we select which content of the input data should be used for the text generation. In the second step, the structure of the document is planned. In the third step, we choose the words which we want to use and aggregate sentences. In the fourth and final step, we generate the text itself based on the previous stages, which should be correct with regard to syntax, morphology, and orthography. In the following sections, we will explain how each of these steps is implemented in our system.

The output of the different steps of the NLP pipeline serves as input for our NLG module. However, the output of the system is not just determined by this input data, but also by the domain model (i.e., our knowledge-base) and the communicative goal we are trying to achieve (see Figure 8.1), i.e., whether we want to provide an explanation to experts or a summary to consumers.

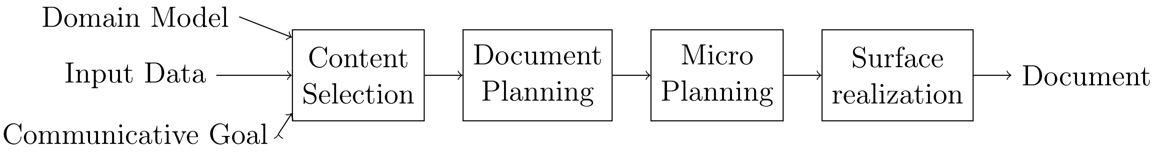


Figure 8.1.: Pipeline for NLG systems; adapted from Reiter and Dale (1997, 2000)

8.3. Content Selection

Our content selection is mainly based on two aspects: clause topics and void clauses. Void clauses are relevant for both consumers and experts and should therefore always be part of the summary. Additionally, as we have explained in Chapter 4, there are certain topics that are usually more interesting for experts or consumers, which should therefore also be part of the summaries. In our system, we will implement the list of such relevant topics dynamically so that each user can define their own priority topics.

For the expert summaries, we always select all void clauses. The length of the summary will, therefore, be determined by the number of void clauses in the document. For the consumer summaries, we let users define how many clauses they want to be contained and, therefore, how long the summary should be. This means that there will be cases in which not all clauses that are generally relevant, by the above-mentioned criteria, can be covered in the summary. For such cases, we need to prioritize the clauses. Clauses that cover one of the priority topics defined by the user and are void are given the highest priority, followed by non-void priority clauses, followed by void non-priority clauses.

The content selection, as well as the next step of document planning, are language independent and therefore identical for German and English.

8.4. Document Planning

For the expert summaries, the document planning is straightforward: each void clause is described as an individual segment because each clause will be reviewed individually by the experts. Each segment consists of one sentence that summarizes the assessment that includes the topic of the clause (e.g., withdrawal or late payment), the fact, based on which the assessment was made (e.g., the withdrawal period or interest rate), the value which was found for this fact (e.g., 7 days or 6 percentage points), and the expected value (e.g., 14 days or 5 percentage points). This sentence is followed by a quote of the original clause text in which the text parts that were used to make the assessment are highlighted.

For the consumer summaries, the document planning is more elaborate. More than for experts, the focus is on shortness, even if that omits more information. The consumer summaries can consist of up to three segments. The first segment summarizes clauses that cover one of the priority topics and are void. For each clause, the facts found and the expected value from the knowledge-base will be reported, e.g., “The contract grants a seven day withdrawal period, by law you are entitled to fourteen days". If none of the priority topic clauses are void, this segment is dropped. The second segment contains a summary of priority topic clauses that were not covered by the first segment (i.e., that are not void). For these clauses, the second part of the summary, starting with "by law", is dropped. The third segment contains clauses that are void but do not cover any priority topics.

8.5. Micro Planning

The micro planning and especially the choice of words is the first language-specific task in the pipeline. The name of the (sub-)topics (like "withdrawal period") can be derived directly from the knowledge-base, then the information that is extracted (like "7 days") can also directly be used in the sentence, as well as the legal requirements (e.g., "14 days"), which can also be taken from the knowledge-base. Based on the label that the extracted information has, we can then decide which verb and subject are needed. e.g., "*The contract grants* a seven day withdrawal period" or "*The shop charges* 3 Euros delivery costs."

8.6. Surface Realization

For the last step in our NLG pipeline, the surface realization, we used the open source library SimpleNLG (Gatt and Reiter, 2009). While surface realization by its very nature is a language-specific task, we will show that thanks to the universal architecture of SimpleNLG, which has been adapted to many languages, we can produce summaries in German and English with relatively little adaption between both languages, based on the surface realizers for both languages. As of 2020, SimpleNLG is available for at least nine languages, including English Gatt and Reiter (2009), German (Bollmann, 2011; Braun et al., 2019b), French (Vaudry and Lapalme, 2013), Italian (Mazzei et al., 2016), Spanish (Soto et al., 2017), Dutch (de Jong and Theune, 2018), Mandarian (Chen et al., 2018), Galician (Cascallar-Fuentes et al., 2018), and Tibetan (Kuanzhuo et al., 2020).

While we could simply use the existing version of SimpleNLG for English (see Section 8.6.1), the only existing German version was based on an outdated version of SimpleNLG, which is based on a very restrictive license, and also came with a very small lexicon, that was by no means sufficient for our purposes. Therefore, we implemented a completely new version of SimpleNLG for German (Braun et al., 2019b), based on the most recent English version of SimpleNLG, and also created the, to the best of our knowledge, biggest lexicon for surface realization in German (Klimt et al., 2020).

8.6.1. English

The original version of SimpleNLG (Gatt and Reiter, 2009) was developed at the University of Aberdeen. The latest versions (4.X) are licensed under the Mozilla Public Licence, while earlier versions have been released under a more restrictive license that prohibits commercial use. SimpleNLG comes with a lexicon that contains around 6,000 lemmata and covers a wide range of grammar, from different tenses to passive voice and even question generation.

8.6.2. German

The original German version of SimpleNLG, which was developed by Bollmann (2011), is not maintained anymore and based on the outdated third version of SimpleNLG, which used a more restrictive license that prohibited commercial use. (Bollmann, 2019) It also comes with a very limited lexicon, consisting of just around 100 lemmata. It also does not automatically recognize and handle separable verbs (see Section 8.6.2.3, for a closer inspection of separable verbs in the German language). The only available open source alternative is a German OpenCCG grammar (Vancoppenolle et al., 2011), which is even more limited with regard to both grammatical coverage and its lexicon. Therefore, we decided to develop SimpleNLG-DE (Braun et al., 2019a), a new German version of SimpleNLG, based on SimpleNLG 4.4.8 and the MPL (see Section 8.6.2.6). In addition, we created a lexicon with more than 100,000 lemmata to be used with SimpleNLG-DE (see Section 8.6.2.5). Both, the new German SimpleNLG version[1] and the lexicon [2], are published under open licenses and already in commercial application by the Bavarian public broadcasting company (Bayerischer Rundfunk). The information contained in this section has been previously published in Braun et al. (2019a) and Klimt et al. (2020).

The German language provides some additional challenges with regard to surface realization, compared to English. Below we outline some of these challenges, which we tackled during the development of the new German version of SimpleNLG.

8.6.2.1. Word Order

The German language is quite liberal when it comes to the order of the words in a sentence and is therefore seen as a “partially free constituent order language” (Vancoppenolle et al., 2011).

The sentence “Ohne Pause in den Hof tragen konnten die Kiste nur zwei kräftige Männer.” (“Only two strong men could carry the box into the yard without a break.”), for example, can be formulated in many different ways, by shuffling the constituents of the sentence:

- “Die Kiste in den Hof tragen konnten, ohne Pause, nur zwei kräftige Männer.”
- “In den Hof tragen konnten die Kiste, ohne Pause, nur zwei kräftige Männer.”
- “Nur zwei kräftige Männer konnten die Kiste ohne Pause in den Hof tragen.”

Such shuffling of the constituents is also called “scrambling” (Eisenberg et al., 2016, p. 881). While in the examples above, the verb remains in the same position, depending on the type of a sentence, its position can also move. A finite verb has to be positioned either in second place, in the first place, or in the last place (Eisenberg et al., 2016, pp. 875-878).

[1] `https://github.com/sebischair/SimpleNLG-DE`, last accessed 2020-01-12

[2] `https://github.com/sebischair/MucLex`, last accessed 2020-01-12

8.6.2.2. Inflection

Another special challenge that makes surface realization for German more complex than English is the number of inflection rules that exist. While the English language only knows “the” as definite article and “a” / “an” as indefinite articles suffice, the German language knows “der”, “die”, and “das” as definite articles and “ein” and “eine” as indefinite articles exist. Additionally, in German, all articles and pronouns must be inflected according to gender, number, person, and the four grammatical case, nominative, genitive, dative, and accusative, which leads to even more article, like “einen”, “einem”, “einer”, and “eines”. (Eisenberg et al., 2016, p. 341)

Nouns, for example, are inflected by gender, grammatical, and number (Eisenberg et al., 2016, pp. 146-228). Adjectives can appear in four different forms: attributive, predicative, adverbial, and nominalized (Eisenberg et al., 2016, pp. 345-372). Attributive adjectives are usually inflected based on the grammatical case, number, and gender of the corresponding noun, e.g.:

- by case: “das große Haus” in accusative, “dem große*n* Haus” in dative (“the big house”)
- by number: “das große Haus” in singular, “die große*n* Häuser” in plural (“the big houses”)
- by gender: “eine große Frau” (“a tall woman”), feminine; “ein große*s* Haus” (“a big house”), neutral

Finally, verb inflection can reflect person, number, tense, voice, and mood (Eisenberg et al., 2016, p. 395). Verbs are grouped into strong and weak verbs in the German language, depending on their inflection pattern in past tense and participle II. Weak verbs build their past tense forms with a syllable introducing t-suffix, e.g., “lachte” (“laughed”), “redete” (“talked”) and their participle II form with “-t”/“-et”, e.g., “gelacht”, “geredet”. The stem vocal of a weak verb does not change. Strong verbs, in contrast, do not build their past tense forms with a suffix, but with an alteration of the stem vocal (so-called “ablaut”), e.g. “rufen - rief” (“to call - called”) or “finden - fand” (“to find - found”). Participle II forms are built with the suffix “-en” and, in some cases, with an “ablaut”: “singen - sang” (“to sing - sang”). Additionally, there are some verbs with a mixed strong and weak conjugation, and other irregularities, like modal verbs, auxiliary verbs, or the verb “wissen” (“to know”). (Eisenberg et al., 2016, pp. 440-466)

8.6.2.3. Separable Verbs

Separable verbs or particle verbs, e.g., “losfahren” (“moving off”), contain a prefix that can be separated from the rest of the verb. The order of the prefix (“los”) and the verb (“fahren”) can be reversed in some cases (Eisenberg et al., 2016, pp. 705-714). The verb “hinausgehen” (“to step out”, “to leave”), for instance, consists of the adverb “hinaus” (“out”), and the verb “gehen” (“to go”). If “hinausgehen” is used in the first person, the present tense is “ich gehe hinaus” (“I step out”), i.e., the prefix is separated from the verb. If it is used in future I tense, however, the prefix stays with the verb “ich werde hinausgehen” (“I will step out”). Prefixes can be prepositional, adverbial, adjective, or substantive particles. The verb “preisgeben” (“to reveal”), for example, contains the substantive particle “Preis” (“price”) in. “Widerspiegeln” (“to reflect”), on the other hand, contains the preposition “wider” (“against”) as a prefix. Separable verbs also exist in other languages, e.g., Dutch (de Jong and Theune, 2018).

8.6.2.4. Compound Words

As part of our NLP-pipeline, we have already discussed the importance of compound nouns for the German language and our specific use-case. However, the German language does not only allow the formation of compound nouns, like "Wunderkind" ("prodigy child"), but also compound adjectives, like "rubinrot" ("ruby red"). Compound words are usually dominated by their last component. Semantically, a "Wunderkind", for example, is a "Kind" ("child") rather than a "Wunder" ("prodigy") and "rubinrot" is a shade of "rot" ("red"), rather than a "rubin" ("ruby"). Grammatically, the last component is also dominant and determines the gender and the inflection type of the whole word. A "Fischerhütte" ("fisherman's hut"), for example, is feminine, although "Fischer" ("fisherman") is masculine because "Hütte" ("hut") is feminine.

Compound words can also be inflected internally, e.g., by a genitive ending, like in "Kapitän*s*mütze" ("captain's head"). Compound words can also not just consist of two words, but multiple, like the infamous "Mindestlohndokumentationspflichtenverordnung" ("minimum wage documentation obligation enactment", Bundesministerium für Arbeit und Soziales (2014)), which combines five words. Generally speaking, compound words are seen as the most important way of building words in the German language (Elsen, 2009).

A similar grammatical phenomenon are the so-called word group lexemes. They are fixed phrases of at least two separately written words, like "Erste Hilfe" ("first aid"), "Europäische Union" ("European Union"), or "Vereinigte Arabische Emirate" ("United Arab Emirates") (Elsen, 2009). From a surface realization perspective, word group lexemes are easier to handle because each part is inflected separately. In dative case, "Vereinigte Arabische Emirate", for example, is inflected "Vereinigte*n* Arabische*n* Emirate*n*".

8.6.2.5. Lexicon

Despite its small lexicon of about 6,000 lemmata, the English version of SimpleNLG covers large parts of the English language because, in comparison to other languages, English has relatively few irregular inflections. If a word is inflected regularly, like "accept" (past simple "accepted" and past participle "accepted"), no lexicon entry is needed to correctly realize the word. However, if a word is inflected irregularly, like "begin" (past simple "began" and past participle "begun"), it will only be realized correctly if the irregular forms are stored in the lexicon. Therefore, a rule-based surface realizer can only be as good as the lexicon it is based on.

In languages with more irregular inflections, a bigger lexicon is needed. The Spanish version of SimpleNLG, for example, comes with a lexicon of 76,000 lemmata (Soto et al., 2017), and the Dutch version with a lexicon of 79,000 lemmata (de Jong and Theune, 2018). The German language is too characterized by various irregular word inflection forms. Adjectives, nouns, and verbs can change their stem vocal in their inflected forms. Moreover, separable prefix verbs are preceded by a prefix, which can be separated from the stem and appear after the verb. Therefore, a comprehensive German lexicon, enriching inflection rules with irregular forms, is needed for German surface realization. A lexicon for the task of NLG, and specifically for surface realization, does not need to contain information about the semantics of a word, unlike, e.g., GermaNet (Hamp and Feldweg, 1997). However, it is not sufficient for such a lexicon to contain only the lemmata, it also has to include the inflections for these lemmata. A noun, for example, can have six inflections: genitive singular, genitive plural, dative singular, dative plural, accusative singular, and accusative plural form. For irregular verbs like "sein" (to be), this number can easily double. Since no well-structured and openly available lexicon of this kind existed for the German language, we created a Lexicon called "MucLex", which is based on Wiktionary. The description of the lexicon was first published in (Klimt et al., 2020).

Wiktionary[3] is a crowd-sourced open lexicon, available in 130 languages. The German language edition contains more than 115,000 lemmata and has 176,431 registered users. Content from Wiktionary is dual-licensed under the Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0)[4] and GNU Free Documentation License (GFDL)[5]. In our work, we opted for the CC BY-SA 3.0 license, under which MucLex is available. While the content of Wiktionary can be downloaded in XML format, the data contained in the XML files is only semi-structured.

The main lemma pages are annotated in the wiki markup language and contain general grammatical information. Listing 8.1 shows the relevant part of semi-structured entry for the verb "sein" ("to be") from Wiktionary, Listing 8.2 shows the same information for the noun "Wortschatz" ("treasury of words") and Listing 8.3 for the adjective "alt" ("old"). The format of the XML files is not structured enough to be used for surface realization. In order to convert the semi-structured information into a well-structured format, we implemented a parser that converts the Wiktionary XML files into a well-structured lexicon that can be used by surface realizers like SimpleNLG.

The parser traverses through the XML files in a top-down approach. XML elements that do not represent lemma pages or inflection tables (e.g., discussion and user pages or disambiguation pages) are skipped. The lemma pages include large information such as word origin, synonyms, and pronunciation, which are not relevant for surface realization. To minimize the file size of the lexicon, we only keep information necessary for the surface realization.

Since lemma pages and inflection tables are separate pages in Wiktionary, they are also part of different XML elements in the dumps, i.e., the information about verbs is distributed within the data. The lemma page or base word entry contains the part of speech, some (but not all) conjugation forms in present and preterit, the participle II form, and some other information. However, information about whether a verb is regular, irregular, or reflexive is not included in the lemma page and can instead be found in verb inflection tables. For each element in the XML dump, the parser first checks whether it contains an inflection table, extracts the relevant information, examines if an entry for the word already was extracted before and if that is the case, adds the information to the extracted entry, instead of generating a new entry. The current version of the lexicon does not include names of persons in order to reduce the size of the lexicon file. The parser is written in Python and published under the Mozilla Public License Version 2.0[6]. The code and the lexicon itself are available from `https://github.com/sebischair/MucLex`.

[3] `https://www.wiktionary.org/`

[4] `https://creativecommons.org/licenses/by-sa/3.0/deed.en`

[5] `https://www.gnu.org/licenses/fdl-1.3.html`

[6] `https://www.mozilla.org/en-US/MPL/2.0/`

```
{{Deutsch Verb Übersicht
|Präsens_ich=bin
|Präsens_du=bist
|Präsens_er, sie, es=ist
|Präteritum_ich=war
|Partizip II=gewesen
|Konjunktiv II_ich=wäre
|Imperativ Singular=sei
|Imperativ Plural=seid
|Hilfsverb=sein
}}
```

Listing 8.1: Semi-structured information for the verb “sein” (to be) in Wiktionary

```
{{Deutsch Substantiv Übersicht
|Genus=m
|Nominativ Singular=Wortschatz
|Nominativ Plural=Wortschätze
|Genitiv Singular=Wortschatzes
|Genitiv Plural=Wortschätze
|Dativ Singular=Wortschatz
|Dativ Singular*=Wortschatze
|Dativ Plural=Wortschätzen
|Akkusativ Singular=Wortschatz
|Akkusativ Plural=Wortschätze
}}
```

Listing 8.2: Semi-structured information for the noun “Wortschatz” (treasury of words) in Wiktionary

```
{{Deutsch Adjektiv Übersicht
|Positiv=alt
|Komparativ=Ülter
|Superlativ=Ültesten
}}
```

Listing 8.3: Semi-structured information for the adjective “alt” (old) in Wiktionary

The old German version of SimpleNLG (Bollmann, 2011) comes with what the author calls a “toy lexicon”. This lexicon consists of only around 100 lemmata and is based on the larger IMSLex from Lezius et al. (2000) which contains more than 50,000 lemmata of which about 11,000 are adjectives, 1,000 adverbs, 22,500 nouns, 300 particles, 10,000 proper nouns, and 6,000 verbs.[7] IMSLex contains information on inflection, word formation, and valence. Like MucLex, it does not contain semantic information. However, the authors suggest that semantic information can be added from GermaNet. Sennrich and Kunz (2014) build a German morphological lexicon for the morphological analyzer SMOR (Schmid et al., 2004) from which is also based on information extracted from Wiktionary. The latest version of the lexicon includes 78,161 lemmata. The lexicon contains the stem, part-of-speech, origin, and SMOR inflection class for each lemma. Tools like SMOR or Morphy (Lezius, 2000) could also be used to create or at least extend lexica, based on their morphological rules.

Multiple German lexica exist for the other popular open source surface realizer OpenCCG.[8] Vancoppenolle et al. (2011) provide a lexicon with approximately 250 lemmata. Hockenmaier

[7] https://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/imslex/, last accessed 2020-08-14

[8] https://github.com/OpenCCG/openccg

(2006) used the TIGER corpus (Brants et al., 2004) to derive a German lexicon for OpenCCG containing more than 2,500 lemmata and more than 46,000 derived word forms. With more than 100,000 lemmata and more than 670,000 word forms, MucLex is, to the best of our knowledge, the biggest open lexicon for the German language of its kind.

In total, it contains 101,509 distinct lemmata, which is slightly less than the total number included in Wiktionary because we excluded, for example, names of people. Table 8.1 shows, how many lemmata from each part-of-speech are included in the lexicon. Overall, the lexicon contains more than 670,000 word forms for the 101,509 lemmata. Due to the relatively large overhead of the XML structure, the file-size of the lexicon is 36 Megabytes. However, the original file of the German Wiktionary, in comparison, takes up more than 1.25 Gigabytes.

PoS	**Lemmata**
Nouns	78,780
Verbs	10,289
Adjectives	11,156
Adverbs	1,127
Total	101,509

Table 8.1.: Amount of lemmata contained in MucLex by part of speech

The XML-format of MucLex is based on the default English lexicon for SimpleNLG. The lexicon contains a `word` entry for each lemma. Depending on the part of speech, the entries include different attributes. Common attributes for all words in the lexicon are the lemma (`base`), a unique identifier (`id`), and the part of speech (`category`).

Entries for nouns (cf. Listing 8.4) include the word's gender (`genus`) and its singular and plural forms for all grammatical cases (nominative, genitive, dative, accusative).

```xml
<word>
    <base>Wortschatz</base>
    <id>47</id>
    <category>noun</category>
    <plural>Wortschätze</plural>
    <genus>m</genus>
    <genitive_sin>
        Wortschatzes
    </genitive_sin>
    <genitive_pl>
        Wortschätze
    </genitive_pl>
    <dative_sin>
        Wortschatz
    </dative_sin>
    <dative_pl>
        Wortschätzen
    </dative_pl>
    <akkusative_sin>
        Wortschatz
    </akkusative_sin>
    <akkusative_pl>
        Wortschätze
    </akkusative_pl>
</word>
```

Listing 8.4: Data format for nouns

Verbs can appear in many different forms in the German language. Including all forms for each tense, person, mood, and voice would inflate the lexicon unnecessarily. Therefore, the lexicon only includes a subset of forms, which can be used to generate all inflections in present, preterite, perfect, pluperfect, and future tenses, for indicative mood, and for active and passive voice. For verbs that change their stem in preterite tense, for example, the preterite stem for all forms can be extracted from just one included preterite form. Listing 8.5 shows the entry for the irregular verb "sein" (to be).

```
<word>
    <base>sein</base>
    <id>35</id>
    <category>verb</category>
    <regular>False</regular>
    <separable>False</separable>
    <reflexive>False</reflexive>
    <plFirstThirdPerPres>
        sind
    </plFirstThirdPerPres>
    <plSecPerPres>seid</plSecPerPres>
    <preterite>war</preterite>
    <participle2>gewesen</participle2>
    <firstPerPres>bin</firstPerPres>
    <secPerPres>bist</secPerPres>
    <thirdPerPres>ist</thirdPerPres>
  </word>
```

Listing 8.5: Data format for verbs

In German, adjectives exist in three forms. The base form, the comparative, and the superlative. The comparative and superlative can be built irregularly and are, therefore, also contained in the lexicon. An example entry is shown in Listing 8.6.

```
</word>
    <base>schnell</base>
    <id>3</id>
    <category>adjective</category>
    <comp>schneller</comp>
    <sup>schnellsten</sup>
```

```
</word>
```

Listing 8.6: Data format for adjectives

Although MucLex is larger than existing lexica, it still has some limitations. Irregular subjunctive and imperative forms are currently not included in the lexicon. Only a limited number of (popular) compound nouns are included. Some nouns commonly used in compound words, for example “Test" (“test"), or “Haus" (“house"), offer large lists of compound words built from them in their Wiktionary entry, e.g., “Crashtest" (“crash test") or “Testfahrer" (“test driver"), however, since these words do not have an individual Wiktionary entry, they are not extracted by the parser and therefore not part of the lexicon. Word entries with the base form and the separated parts used in a compound word could be included in the lexicon in order to increase the coverage. Whether a lexicon should contain compound nouns or whether it should just contain the base nouns and leave the composition to the surface realizer is debatable. There are existing approaches on how to automatically split compound words into their respective parts (e.g., by Baroni et al. (2002); Koehn and Knight (2003); Daiber et al. (2015); Sugisaki and Tuggener (2018), and Weller-Di Marco (2017)). The problem is not trivial and many compound nouns are irregular and would hence still need to be included in the dictionary. Therefore, the best solution might be a mix of lexicon entries for irregular nouns and split approaches for the others.

If an entry in Wiktionary contains multiple equally ranking inflection forms for a word, for example, “des Landes" and “des Lands" for the noun "Land" ("country"), the form listed first is extracted. This leads to limitations for nouns with multiple plural forms that have different meanings. The German word “Bank", for instance, may mean “bench" or “bank" (credit institute). Only the plural form reveals the semantic difference: “benches" in German is “Bänke", but "banks" results in “Banken". The lexicon currently includes only the first-named plural form.

8.6.2.6. SimpleNLG-DE

In this section, we shortly describe which parts of the grammar of the German language are implemented in the first version of SimpleNLG-DE, which parts of the German language are not yet covered, and how we evaluated the first implementation of the library. The code of SimpleNLG-DE is available from `https://github.com/sebischair/SimpleNLG-DE`.

Syntax

The word order in SimpleNLG-DE is handled according to the topological model of the “Duden” (Eisenberg et al., 2016, pp. 874 - 880). The current version supports declarative clauses and questions. Unlike the German SimpleNLG version of (Bollmann, 2011), SimpleNLG-DE can automatically detect separable verbs and adapt the word order if necessary, e.g., “Alice räumt das Auto ein.” / “Alice is loading the car.”. The handling of separable words is similar to the implementation in the Dutch version of SimpleNLG (de Jong and Theune, 2018). Separable verbs are marked as such in the lexicon, and their entries include the prefix. Initiated subordinate clauses that contain a separable verb have to be treated with care. As an example, the complex sentence “Florian geht einkaufen, Alex räumt sein Zimmer auf.” (“Florian goes shopping, Alex cleans his room.”) can be changed to “Florian geht einkaufen, während Alex sein Zimmer aufräumt.” (“Florian goes shopping, while Alex cleans his room.”), with the second sentence added as initiated subordinate clause to the first. In the second sentence, the word order and the verb conjugation are changed. The separable verb is separated in the first clause (“räumt auf“ / “tidies up”), but stays together in the second clause (“aufräumt”). (Agbaria, 2009) For all initiated subordinate clauses, SimpleNLG-DE does not split separable verbs. In addition to declarative clauses, SimpleNLG-DE can generate five different types of questions: yes/no questions and questions about the subject and object of a sentence for both people (“wer” / “who”) and things (“was” / “what”). Beyond main clauses, the library can handle compound sentences connected with “und” (“and”) or comma (“Der Hund bellt und die Katze miaut” / “The dog barks and the cat mews.”), temporal, causal, conditional, consecutive, concessive, modal, comparative, final, and adversative subordinate clauses (“Die Sonne scheint, während es regnet.” / “The sun shines while it is raining.”), appositions (“SAP, eine deutsche Firma, ...” / “SAP, a German company, ...”), and enumerations (“SAP, Bayer und EON” / “SAP, Bayern, and EON”).

Morphology

Morphology in SimpleNLG-DE is based on a combination of hard-coded rules for regular inflections, based on Eisenberg et al. (2016) and Agbaria (2009), and the lexicon described in Section 8.6.2.5. The lexicon does not contain all conjugated forms for all persons in all tenses. However, it contains a big enough set of forms to create all inflected forms with additional rules. If a verb is not in the lexicon, the regular conjugation rules are applied.

The verb conjugation covered by SimpleNLG-DE includes present, past, perfect, and future tense, passive in present and past, modal verbs in present, and the handling of separable verbs. Adjectives are declined according to the case, number, and article. Moreover, the generation of the comparative and superlative of adjectives and adverbs is supported. Nouns can be inflected based on the case and number. Word group lexemes can also be inflected based on case and number. Articles, both definite and indefinite, can be inflected according to the case, number, and gender. Additionally, SimpleNLG-DE is able to automatically detect the contraction of prepositions and inflect adjectives correctly in cases like “in dem großen Haus” which can be contracted to “im großen Haus” (“in the big house”).

Orthography

SimpleNLG-DE features an orthography processor, which handles different tasks, like terminating declarative clauses with “.” and questions with “?”, capitalizing the first character in a sentence, and adding commas. If a phrase is added as a complement to another sentence, and both of them do not add a complementizer, or the complementizer is in a list of conjunctions that requires a comma, the complement is added with a preceding comma. For sentences added with the complementizer “und”, no comma is added. Appositions have a comma added before and after them, no matter if “und” is contained in the apposition or not. Enumerations are built by connecting the first constituents with a comma and separating the last with an "und", for instance, in "A, B und C" ("A, B, and C").[9]

Limitations

Due to the complexity of the German language, its manifold inflected words and rules, and its diverse word order possibilities with a large number of exceptions, SimpleNLG-DE is only able to cover a subset of the German language. In this section, we give an overview of the current limitations and indications on how to extend the library in the future.

Tenses currently not covered by SimpleNLG-DE are future II ("Ich werde es gekauft haben." / "I will have bought it.") and plusquamperfect tense ("Sie hatte Fußball gespielt." / "She had played football."). Passive voice is only supported in present and preterite tenses and modal verbs are only supported for active voice in present tense. Phrases such as "soll verursacht sein" ("shall be caused"), for instance, are not yet covered. Imperative and subjunctive mood verb forms are not implemented, as well as the conjunctive and imperative in general.

Compound words are only supported if they are part of the lexicon and can not be detected automatically. There are existing approaches on how to automatically split compound words into their constituents (e.g., by Baroni et al. (2002), Koehn and Knight (2003), Daiber et al. (2015), Sugisaki and Tuggener (2018), and Weller-Di Marco (2017)), however, the problem is far from trivial and not yet implemented.

For some German verbs, there are several correct ways to conjugate them. The verb "senden" ("to send"), for instance, in the third person past tense, can either be conjugated to "sendete" or to "sandte", while keeping the same meaning. Multiple interchangeable, correct conjugations are currently neither supported by the library nor by the lexicon.

While the meaning is preserved in both forms in the case of "senden", some verbs also change their meaning. The verb "wachsen" in third person present tense in its irregular form is "er wächst" meaning "he is growing", whereas the regular form "er wachst" means "he waxes (sth.)". In the future, the user should have the option to set the desired meaning.

Evaluation

There is no standardized approach for the evaluation of surface realizer because many different aspects have to be considered. In our evaluation, we focused on two facets: First, how robust and correct is the implementation of the grammatical features, and second, how much of everyday language can be covered with the current implementation of SimpleNLG-DE.

The robustness and correctness can be evaluated relatively easy by manually generating test cases for the different grammatical features that have been implemented. The evaluation of the coverage is far more complex. The best, though not perfect, approach is choosing an existing corpus that is believed to be representative of the language to a certain degree. This approach was also chosen by the authors of other versions of SimpleNLG. Bollmann (2011), for example,

[9] The German language does not have an equivalent to the Oxford comma.

used five Wikipedia articles with 152 sentences in total to evaluate the coverage and achieved 75.66%. Since the data used for the evaluation was not published, we can not compare our new version directly with the previous version. Other language versions, including Spanish and Mandarin, used translations of the 144 test sentences from the original SimpleNLG version to conduct evaluations. However, according to the original authors, these sentences were meant to be an “indication of efficiency” test (Gatt and Reiter, 2009) and not an evaluation of the coverage.

In our evaluation, we used more than 3,800 sentences for the evaluation of the correctness of the implementation of grammatical features. These sentences cover, e.g., the inflection of verbs (2,436 sentences), the inflection of adjectives (1,002 sentences), and the inflection of nouns (390 sentences), but also other features like question generation. The sentences were partially extracted from documents from the financial domain and partially written by the author for testing purposes. The realization of the sentences was implemented manually. SimpleNLG-DE generated all test sentences correctly.

In order to get an estimate of how much of the German language is covered, we used the TIGER Corpus (Brants et al., 2004). It contains 50,000 sentences of German newspaper articles taken from the “Frankfurter Rundschau”. Newspaper articles often contain sophisticated phrase structures and can therefore be considered suitable test data for a language realizer. Annotations in the TIGER corpus include semi-automatically generated POS-tags, syntactic structure, and morphological and lemma information. The corpus is available freely for research and evaluation purposes. Since the code for the generation of the sentences has to be written manually, which is very time consuming, we could not use the entire corpus for the testing. Instead, we randomly chose 100 declarative sentences from the corpus (interrogative, imperative, and exclamatory sentences were excluded) and implemented them using SimpleNLG-DE. We used the annotations from the TIGER corpus to semi-automatically create the code for the tests; however, all sentences were manually checked and adapted before they were added to the test set. 84% of all sentences could be generated correctly using the library. Counted as correct are only sentences that are equal to their corresponding sentence in the corpus. The main reasons for wrongly realized sentences include problems with the pluralization of irregular compound nouns which are not part of the lexicon and of verbs in cases where the corresponding noun is a number (e.g., “Im Schnitt waren es seit 1980 jedoch nur 4 208.” / “On average, however, since 1980 it has been only 4 208.”). Since the code for the tests is only compatible with SimpleNLG version 4, we were not able to directly compare the performance of the old against the performance of the new version of SimpleNLG on the TIGER corpus. The tests generated from the TIGER corpus are not published for licensing reasons; however, all other tests are part of the SimpleNLG-DE repository.

8.7. Examples

To show more concretely how we generate summaries, we use the following clause as an example:

> *“**Withdrawal***
> *All products ordered by a consumer can be returned without specifying any reason within 7 days if the contract (using only means of telecommunications in particular letters, catalogues, phone calls, faxes, e-mails as well as radio, television and Media Services) has been established.”*

and its German equivalent:

> *“**Widerrufsrecht***
> *Alle von einem Verbraucher bestellten Artikel können ohne Angabe von Gründen innerhalb von 7 Tagen zurückgegeben werden, wenn der Vertrag unter ausschließlicher Verwendung von Fernkommunikationsmitteln (insbesondere Briefe, Kataloge, Telefonanrufe, Telekopien, E-Mails sowie Rundfunk-, Tele- und Mediendienste) zustande gekommen ist.”*

After successfully passing the clause through our NLP-pipeline, we receive a JSON-representation of the clause as shown in Listing 8.7. It contains the title and text of the clause, as well as four annotations. The first two annotations are made by the rule-based topic classifiers, one by the classifier for topics, and one by the classifier for subtopics.

The third annotation was made by the rule-based information extraction algorithm, which detected an information of type “period” and a value of “7 days”. Finally, the rule-based legal assessment has annotated the clause as void. In comparison to the other annotations, the source label here does not just specify the classifier that made the decision, but the actual rule that was applied, which refers to the name of a legal assessment rule in our knowledge-base (see Listing 7.10).

```
{
    "title": "Withdrawal",
    "text": [...],
    "annotations": [
        {
            "type": "topic",
            "value": "withdrawal",
            "source": "ruleTopicClassifier"
        },
        {
            "type": "topic",
            "value": "withdrawal:period",
            "source": "ruleSubtopicClassifer"
        },
        {
            "type": "period",
            "value": "7 days",
            "source": "ruleIe"
        },
        {
            "type": "void",
            "value": "true",
            "source": "withdrawalPeriodRule"
        }

    ]
}
```

Listing 8.7: JSON representation of a clause after completion of the (rule-based) NLP-pipeline

Based on this JSON representation of the clause and information from the knowledge-base, we generate the following summary for consumers:

> “The contract grants a seven day withdrawal period, by law you are entitled to fourteen days.”

and its German equivalent:

> “Der Vertrag räumt eine Widerrufsfrist von 7 Tagen ein, gesetzlich stehen Ihnen vierzehn Tage zu.”

The code that produces the English summary is shown in Listing 8.8. While quite a lot of code is needed to generate the summary, only very few hard-coded language elements are used. For the first part of the sentence “The contract grants a seven day withdrawal period”, we only need to add “the contract”, “grant”, and “a”. All other elements can be directly derived from the knowledge-base or the JSON representation of the clause. This first part is produced by the `summarize` function in Listing 8.8.

While our example clause contains only one topic and one type of information, a clause could also contain a variety of topics and corresponding information. For each of this information, a separate summary sentence will be generated. The code shown in Listing 8.8 can not only produce summaries for clauses about the withdrawal period but in general for periods (e.g., “The contracts grant a two year warranty period.”). If the clause is not-void with regard to the extracted aspect, the summary finishes after the first sentence. If it is void, the `describeLaw` function is called, which describes the legal requirements based on which the assessment was made.

```
    public SPhraseSpec describeLaw(AssessmentRule assessmentRule){
        SPhraseSpec legalRegulation = this.nlgFactory.createClause();
        legalRegulation.addFrontModifier("by law");

        if (assessmentRule.getInformation.contains("period")) {
            legalRegulation.setSubject("you");
            legalRegulation.setVerb("be");
            legalRegulation.setObject("entitled");
            PPPhraseSpec pp = nlgFactory.createPrepositionPhrase("to", assessmentRule.getComparison().getValue());
            legalRegulation.addComplement(pp);
        }
        else {...}

        return legalRegulation;
    }

    public List<String> summarize(Clause clause) {
        ArrayList<String> summaries = new ArrayList<String>();

        for (Annotation extracedInformation: clause.getExtractedInformation()
    ↪ ) {
            SPhraseSpec summary = this.nlgFactory.createClause();

            if (extracedInformation.getType().equals("period")) {
                summary.setSubject("the contract");
                summary.setVerb("grant");

                NPPhraseSpec nounPhrase = this.nlgFactory.createNounPhrase();
                nounPhrase.setDeterminer("a");
                nounPhrase.setNoun(knowledgeBase.getAssessmentRule(
    ↪ extracedInformation.getType()).getTopic().getName());
                nounPhrase.addPreModifier(clause.getAnnotation("period").
    ↪ getValue());
                summary.setObject(nounPhrase);
            }
            else {...}

            CoordinatedPhraseElement c = nlgFactory.createCoordinatedPhrase()
    ↪ ;
            c.setConjunction(",");
            c.addCoordinate(summary);
            if(clause.isVoid()) {
                c.addCoordinate(describeLaw(knowledgeBase.getAssessmentRule(
    ↪ extracedInformation.getType())));
            }

            summaries.add(realiser.realiseSentence(c));
        }

        return summaries;
    }
```

Listing 8.8: Java code for the generation of summaries for clauses about the withdrawal period in English

One of the advantages of using a surface realizer, rather than, for example, a template-based approach, is that we can change the generated texts without having to change the code by just adapting the information in the knowledge-base. If we change, for example, the name of the topic “withdrawal:period” in the knowledge-base from “withdrawal period” to “withdrawal timeframe”, the summary is automatically adapted accordingly. Even more, the surface realizer does not just change the variable parts of the sentence but also adapts the inflections of the fixed parts. If the withdrawal period would, for example, be eleven days, the summary would be “The contracts grants a*n* eleven day withdrawal period.” instead of “a eleven day”.

A specific additional advantage of SimpleNLG is that we can create the same summary in German, with only little adaptions to the code, as is shown in Listing 8.9. In addition to replacing the words with their English translations, there are only two small changes necessary. In the `describeLaw` method, we can remove the prepositional phrase because the splittable German verb “zustehen” contains the preposition already as prefix. In the `summarize` method, on the other hand, we have to add an additional preposition, which goes before the period itself (“*von* sieben Tag”, “*of* seven days”). These examples also underline the importance of separable verbs again because both sentences use them.

```
public SPhraseSpec describeLaw(AssessmentRule assessmentRule){
    SPhraseSpec legalRegulation = this.nlgFactory.createClause();
    legalRegulation.addFrontModifier("gesetzlich");

    if (assessmentRule.getInformation.contains("period")) {
        legalRegulation.setSubject("IHnen");
        legalRegulation.setVerb("zustehen");
        legalRegulation.setObject(assessmentRule.getComparison().getValue
↪ ());
    }
    else {...}

    return legalRegulation;
}

public List<String> summarize(Clause clause) {
    ArrayList<String> summaries = new ArrayList<String>();

    for (Annotation extracedInformation: clause.getExtractedInformation()
↪ ) {
        SPhraseSpec summary = this.nlgFactory.createClause();

        if (extracedInformation.getType().equals("period")) {
            summary.setSubject("der Vertrag");
            summary.setVerb("einräumen");

            NPPhraseSpec nounPhrase = this.nlgFactory.createNounPhrase();
            nounPhrase.setDeterminer("ein");
```

```
                nounPhrase.setNoun(knowledgeBase.getAssessmentRule(extracedInformation.getType()).getTopic().getName());
                PPPhraseSpec pp = nlgFactory.createPrepositionPhrase("von", clause.getAnnotation("period").getValue());
                nounPhrase.addComplement(pp);
                summary.setObject(nounPhrase);
            }
            else {...}

            CoordinatedPhraseElement c = nlgFactory.createCoordinatedPhrase();
            c.setConjunction(",");
            c.addCoordinate(summary);
            if(clause.isVoid()) {
                c.addCoordinate(describeLaw(knowledgeBase.getAssessmentRule(extracedInformation.getType())));
            }

            summaries.add(realiser.realiseSentence(c));
        }

        return summaries;
    }
```

Listing 8.9: Java code for the generation of summaries for clauses about the withdrawal period in German

The expert summaries have an explanatory element, which consists of the original sentence, enriched with HTML highlighting, in addition to the sentence that consumers also receive. For our example clause, the highlighting looks like this:

All products ordered by a consumer can be **returned** without specifying any reason **within 7 days** if the contract (using only means of telecommunications in particular letters, catalogs, phone calls, faxes, e-mails as well as radio and television) has been established.

The bold parts of the text highlighting the keywords that were relevant for the topic classification. The colored background indicates information that was extracted, and the color itself indicates the legal assessment: If an extracted information was assessed by a rule and the outcome was "void", then the color is red. If the information was assessed by a rule and the outcome was not "void", then the color is green. If the information was not used in any legal assessment, the color is yellow.

8.8. Evaluation

In order to evaluate the automatically generated summaries, we selected five clauses and their respective (German) summaries and presented them to both consumers and consumer advocates.

The consumers only received the pure summary, while the consumer advocates also received the explanatory component. The consumer evaluation was again conducted using Prolific, like described in Section 4.2, with 100 participants. The expert evaluation was conducted with five participants. The numbers are in line with the suggestions from van der Lee et al. (2019), who suggest having at least three participants for expert evaluations and 100 or more for larger-scale studies with non-experts.

We decided to evaluate individual clauses and their summaries, instead of summaries for complete contracts, for two reasons: In order to assess a summary for a complete contract (or better multiple contracts), participants would have to read the complete contracts, which is very demanding. Additionally, when evaluating summaries for whole documents, there are many factors that influence the perception of these summaries, and it would be more difficult to evaluate individual aspects of the summary generation. As a drawback, in this way, we only evaluated the two last stages of our NLG-pipeline, i.e., the micro planning and surface realization, and not the content selection and document planning.

8.8.0.1. Statements

For each pair of clause and summary, we presented the participants with five statements and asked them to rate each statement on a five-point Likert-scale (strongly disagree, disagree, undecided, agree, strongly agree), in order to assess the fluency, readability, accuracy, and utility of the generated summaries. The five statements were[10]:

- **S1** (fluency): The summary is grammatically correct.
- **S2** (readability): The summary is easier to understand than the original text.
- **S3** (accuracy): The content of the summary is backed by the original text.
- **S4** (accuracy): All important information from the original text is part of the summary.
- **S5** (utility): The information in the summary is useful to me.

For accuracy, we provided two statements. The first (S3) rather focuses on precision, i.e., is the information that is in the summary correct, while the second (S4) focuses on recall, i.e., is all the important information present in the original clause also in the summary. In addition, at the end of the questionnaire, we presented consumers and consumer advocates with a free-text field to give additional comments and one additional statement to rate for each group of participants:

- **S6 / consumer**: I am more likely to read a summary of T&C than the actual T&C.
- **S6 / experts**: The explanatory element helps me to understand the reasoning of the algorithm.

[10] The evaluation was conducted in German, the here presented statements are translations of the original statements

8.8.0.2. Summaries

In the following, we will present the five (translated) pairs of clauses and summaries:

- **Clause 1**:
  All products ordered by a consumer can be returned without specifying any reason within seven days if the contract (using only means of telecommunications in particular letters, catalogues, phone calls, faxes, e-mails as well as radio, television and Media Services) has been established.
  Summary: *The contract grants a seven day withdrawal period, by law you are entitled to fourteen days.*
- **Clause 2**:
  The warranty period for new items shall be 24 months. The period shall commence upon transfer of risk.
  Summary: *The contract grants a 24 months warranty period.*
- **Clause 3**:
  If you are in arrears with payment, we reserve the right to charge reminder fees, unless you can prove to us that the costs did not arise for us at all or are considerably lower than the fees claimed. The reminder fees are 0.00 EUR for the 1st reminder and 5.00 EUR for the 2nd reminder.
  Summary: *In case of default, the company charges 5 EUR reminder fees, by law they are eligible to charge 2.50 EUR.*
- **Clause 4**:
  For items in stock, we deliver in 1–2 working days from receipt of the order and clearance of payment.
  Summary: *The delivery period is 1-2 days.*
- **Clause 5**:
  You may revoke your contractual declaration within two weeks without providing reasons in written form (e.g., fax, email). The period commences with the receipt of the goods at the earliest. To ensure the period of revocation the timely sending of the revocation or the object suffices
  Summary: *The contract grants a two weeks withdrawal period.*

8.8.0.3. Results

The results of the evaluation for Clause 1 are shown in Figure 8.2. In order to make the responses from the two groups better comparable, the figure shows percentages rather than the absolute number of responses. The responses to the first statement show that consumers and experts alike unanimously agree that the summary is grammatically correct. Both groups also agreed in the same way that the summary is was easier to understand than the original clause.

This consensus between consumers and experts ended when we provided statements about the accuracy of the summary. The consumers overwhelmingly agreed that the content of the summary is backed by the original text. Since the survey was completely anonymous, we, unfortunately, did not have the possibility to follow-up with the participants. However, based on the responses to the other clauses, we assume that the small number of participants that choose neutral (21%), mainly did so, because the addition "by law you are entitled to fourteen days" can not be derived from the text. In hindsight, this was a design failure of the survey because, if the statement was to be taken literally, the "correct" answer here would be to disagree with the statement because of this addition. With experts, we could follow up. For them, more than this addition, the problem was that the summary suggests the withdrawal period would be universal, while the original clause specifies that it only applies to contracts that have been established using distance communication services.

Consumers unanimously found the summary to be helpful for them, while only two experts (40%) found them helpful. For these two experts, the explanatory element was the reason why they found the summary helpful.

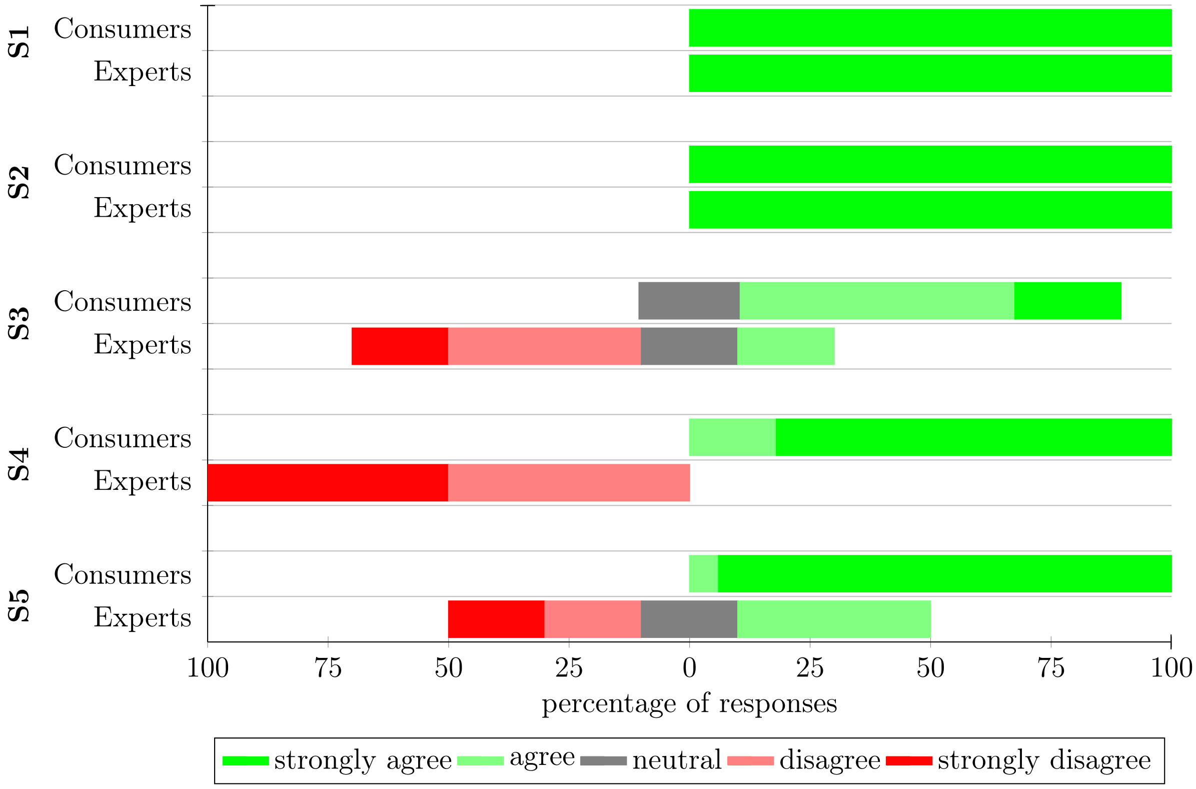


Figure 8.2.: Likert scale results for the different statements for the summary of Clause 1.

For Clause 2 (see Figure 8.3), both groups again agreed that the summary is grammatically correct and easier to understand than the original text. In comparison to Clause 1, more consumers agreed that the summary is backed by the original text. Our assumption is that this is the case because the summary did not contain the second part, describing the legal situation. Experts, however, disagreed even more than before, especially, as we found out in discussion after the survey, because the information that the warranty period applies to *new* products was removed. This also shows that our idea of testing separately for precision and recall of the summary did not fully work out because the differentiation, whether leaving some part out, leads to a lower precision or recall (or both), is one that is not easy to make. While most consumers still found the summary helpful, not a single expert said the summary would have been helpful for them, mostly because the information in the summary alone would not have been sufficient to make a legal assessment.

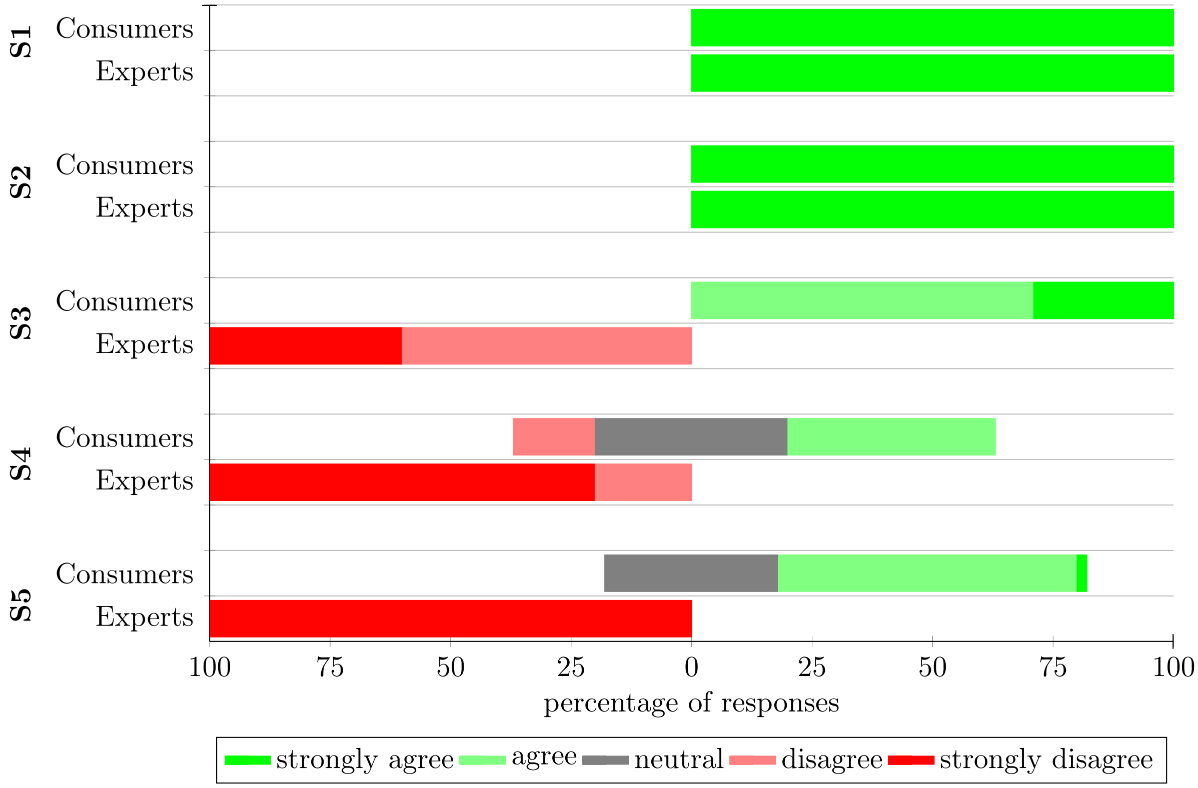


Figure 8.3.: Likert scale results for the different statements for the summary of Clause 2.

Interestingly, for Clause 3, the view were turned (see Figure 8.4). From an objective standpoint, the summary “ignores” that the first payment reminder is free, and only the second reminder is charged with 5 Euros. Based on the responses, for consumers, this is an important piece of information, which, from their perspective, is missing in the summary. For the legal assessment, however, this information is irrelevant. The amount that can be charged for a payment reminder is independent of how many reminders have already been sent. Therefore, from the perspective of our experts, the summary contains all the information they need. This is a very nice example that the question of whether a summary is “complete” and useful depends very much on the reader and the task the reader hopes to achieve by reading the summary.

The summary of Clause 4 shows good results in both groups, as shown in Figure 8.5. Although both groups agree that the summary is accurate, consumers find the summary more useful than experts, which is not surprising, given that delivery times are a priority for consumers, but not for consumer advocates, as we explained in Chapter 4.

Finally, the results for the summary of Clause 5 are shown in Figure 8.6. For Clause 5, we see

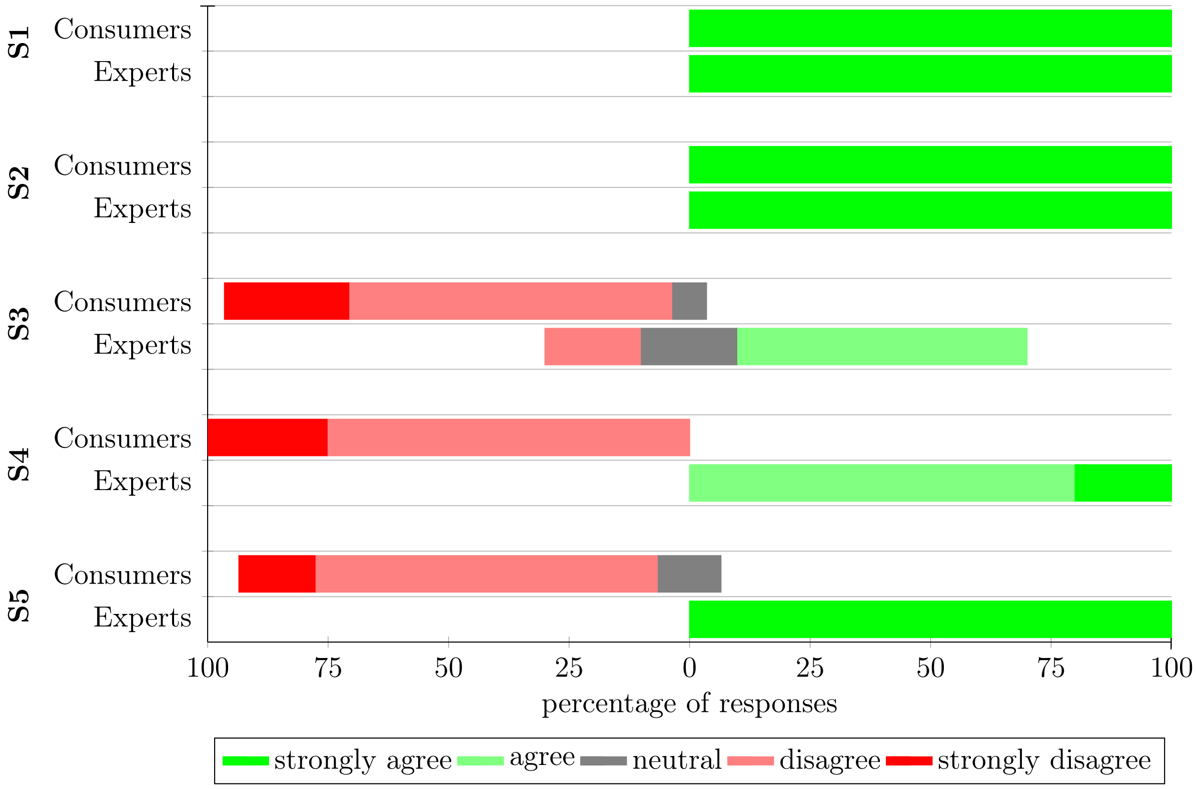


Figure 8.4.: Likert scale results for the different statements for the summary of Clause 3.

similar results to Clause 1 and 2, i.e., consumers find the summary to be helpful, while experts think important information is missing and therefore do not find the summary to be helpful.

In conclusion, the summary has shown that our automatically generated summaries are fluent and easily readable. With one exception, consumers also found them to be an accurate representation of the original clauses. Experts, however, found them to be not accurate, mainly because to leave out information that is important to them. Given these results, it is not surprising that consumers, overall, also found the summaries useful, while experts, overall, did not find them useful.

For our final statements S6, 41% of consumers were neutral to the statement "I am more likely to read a summary of T&C than the actual T&C", 18% agreed, and 41% strongly agreed. For experts, 60% agreed with the statement "he explanatory element helps me to understand the reasoning of the algorithm" and 40% strongly agreed.

For the question of whether automated summaries can improve the way in which consumers and consumer advocates work with standard form contracts (RQ8), the answer has to be split. For consumers, the answer is yes. Even if the summaries sometimes leave out information that might be legally relevant, consumers feel well informed by them. And if they indeed make it more likely that consumers inform themselves at all about the T&C, which statement S6 suggests, then this is a positive result.

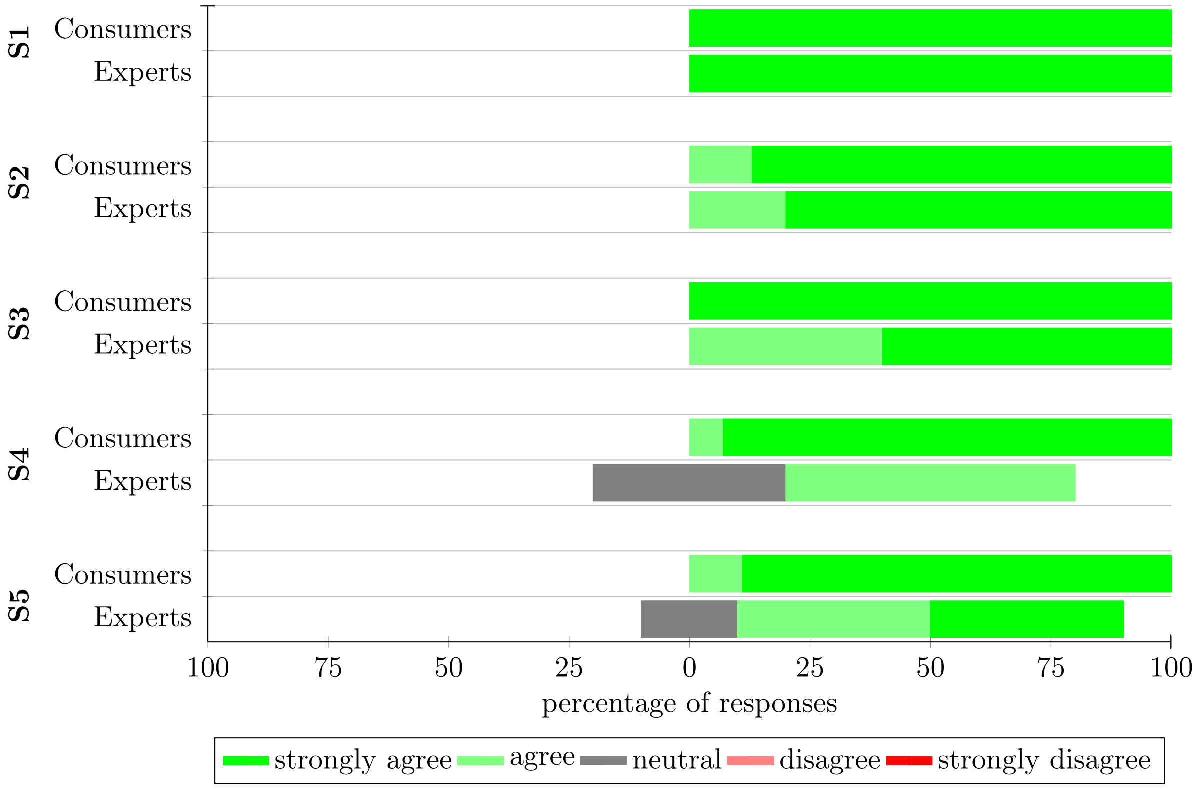


Figure 8.5.: Likert scale results for the different statements for the summary of Clause 4.

For consumer advocates, however, the summaries have shown to be not useful because to leave out too much relevant information. While, in the survey, the explanatory element has been shown to be appreciated by the experts, its utility in practical application is questionable. While at least the first part of the consumer summary can also be created from the results of ML classifiers, the explanatory element completely relies on the rule-based classifiers, which have shown to perform considerably worse than the ML classifiers. In summary, we have to conclude that, for consumer advocates, automated summaries provide little benefit.

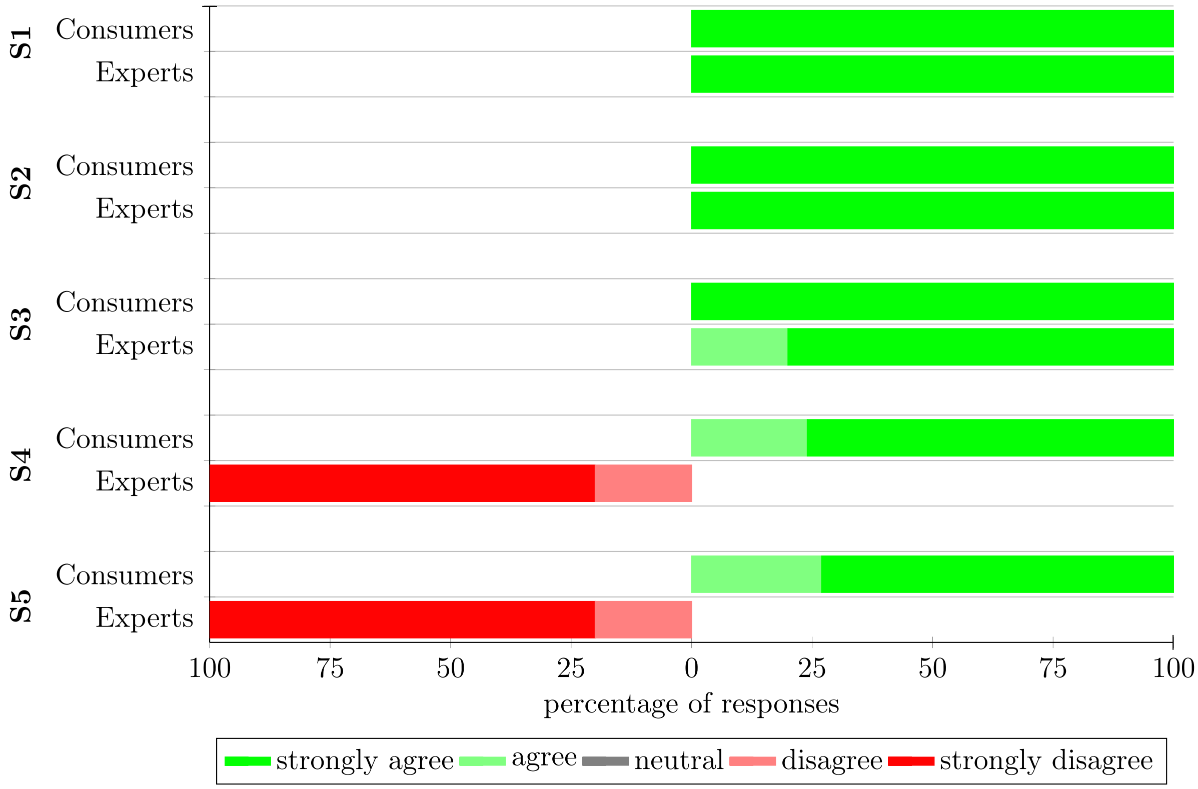


Figure 8.6.: Likert scale results for the different statements for the summary of Clause 5.

CHAPTER 9

Tool-Supported Legal Assessment of Standard Form Contracts

In this chapter, we describe the prototypical implementation of a web-based client application for the AI-assisted legal analysis of standard form contracts, based on the technology presented in this thesis. While the focus of this thesis is NLP and NLG technology, we believe that the only way of assessing the possible impact these technologies can have on the practical work of consumer advocates is by evaluating them as part of a usable prototype.

9.1. Architecture

The different parts of the pipeline we developed have very different technical requirements: The first steps are lightweight web scraping and pre-processing tasks, implemented in Python. They are followed by computationally relatively cheap rule-based analyzes implemented in Java and computationally (especially GPU) heavy classifications with deep neural networks. At the end of the pipeline, we can add an NLG component, which is implemented in Java and computationally relatively cheap again but depends on big lexica, which have to be loaded into the RAM. Therefore, it was a natural choice to separate these different steps by their functionality and hardware requirements into separate (micro-)services.

Micro-service architectures consist of independently deployed and running processes, which communicate through lightweight, technology-independent interfaces, like HTTP APIs. They usually contain a central component, which also forms an independent process and orchestrates the individual services. (Namiot and Sneps-Sneppe, 2014) In our prototype, this role is taken on by the web-based client application, which is an Angular 9 application that runs on a Node.js server.

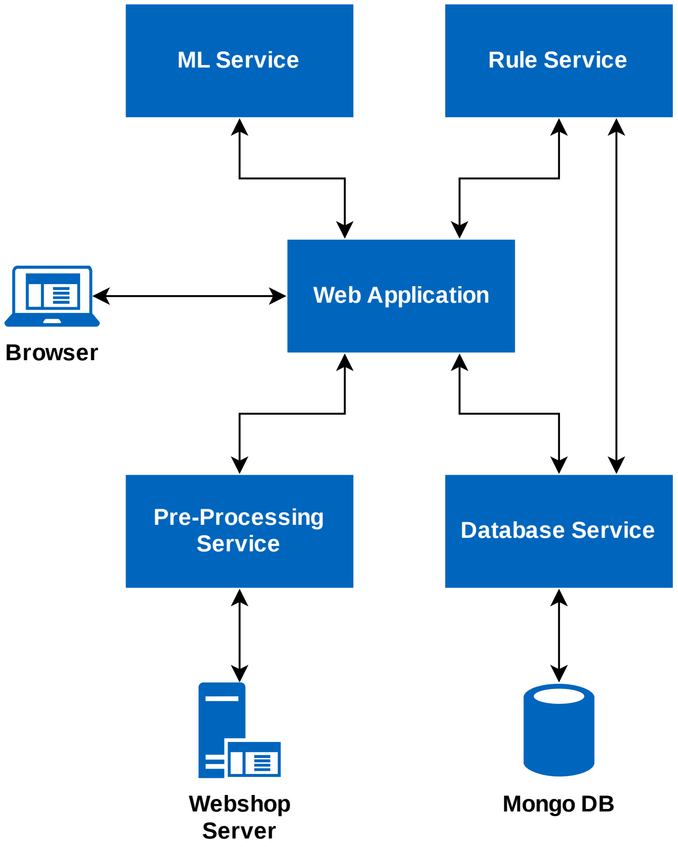


Figure 9.1.: Architecture of the research prototype

Figure 9.1 visualizes the architecture of the prototype. In addition to the aforementioned three services and the web application, the architecture also contains a database service with a connected MongoDB. The web application will use the database to save user accounts and configurations, but also the texts and annotations of contracts and the knowledge-base. The rule service needs direct access to the database service because the rules it applies are stored in the knowledge-base, i.e., in the database.

In the following sections, we will briefly describe the internal architecture and implementation of each of these components and how their APIs are designed.

9.1.1. Pre-Processing Service

For the implementation of the pre-processing service, and also for the implementation of the ML Service, both of which are implemented in Python, we used the connexion framework[1], which was developed by Zalando. The framework automatically generates HTTP APIs in Python

[1] `https://github.com/zalando/connexion`, last accessed 2020-09-21

based on OpenAPI[2] specifications. In this way, we do not only save implementation time but also make sure that the API documentation is always synced with the implementation.

The pre-processing service offers two core functionalities, which are accessible through two different routes:

- GET `/find/{url}`: Takes the URL of a website as input and returns the link to the T&C page, if one could be found. If no T&C page could be found, a response with status 222 is returned. If the input URL could not be found, a response with status 404 is returned.
- GET `/extract/{url}`: Takes the URL to a T&C page as input and returns the extracted contract text as result. If the query parameter `structured` is set to false, the content will be returned as plain text. If the parameter is set to true (or not set at all), the content is returned structured, in the format described in Section 6.2.3.2 (see Listing 6.3 for the JSON schema).

Since we already described the implementation of both functionalities in Chapter 6, Section 6.1 and 6.2, we will not repeat the details of the implementation at this point.

9.1.2. Rule Service

The rule service is implemented in Java, using the Restlet framework (Louvel et al., 2012) to provide the HTTP API. The rule service combines three functionalities, each of which is exposed by an individual API route:

- POST `/classify`: Classifies and returns the topics and subtopics that the clause covers, which is transmitted in the body of the request.
- POST `/extract`: Extracts and returns the information from the clause that is transmitted in the body of the request, together with the topics and subtopics of the clause.
- POST `/assess`: Legally assesses the clause that is sent in the body of the request as array of strings, together with the topics and subtopics of the clause, and returns whether the clause is void or not.

In Restlet, each route is attached to a so-called handler. Each handler provides a function `handle`, which is invoked every time the attached route is called. The function is called with two parameters, one representing the incoming request and one representing the response that will be sent out after the handle method finished. For each of the routes, we created a separate class which inherits from the `Handler`-class (see Figure 9.2).

Each of the classes contains a so-called Analyzer. These classes contain the business logic for each route, perform the analysis of the clauses, and return a JSON object that will be sent back in response to the request. While for the topic classification only the clause text is needed as input, the `ExtractionAnalyzer` and `AssessmentAnalyzer` additionally need the clause topics as input (see Figure 9.3).

[2] Formerly known as Swagger.

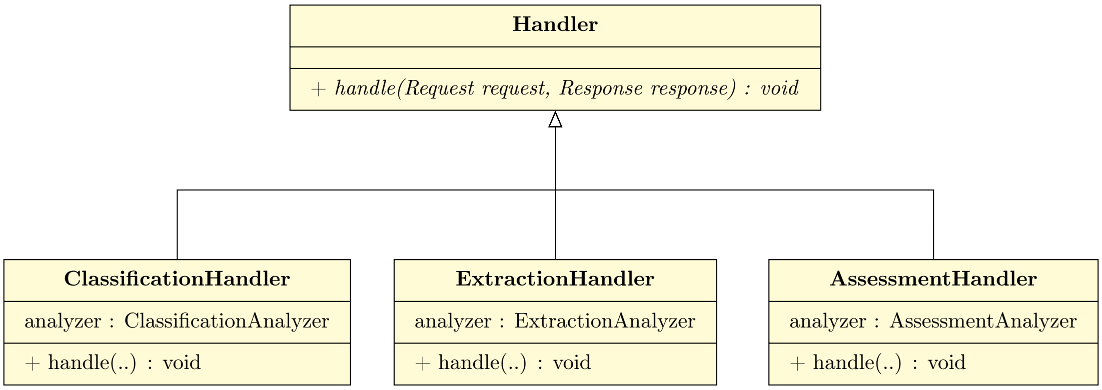

Handler
+ handle(Request request, Response response) : void
ClassificationHandler
analyzer : ClassificationAnalyzer
+ handle(..) : void
ExtractionHandler
analyzer : ExtractionAnalyzer
+ handle(..) : void
AssessmentHandler
analyzer : AssessmentAnalyzer
+ handle(..) : void

Figure 9.2.: Class diagram for Handler classes ("..'' mark omission of redundant parameters for space reasons)

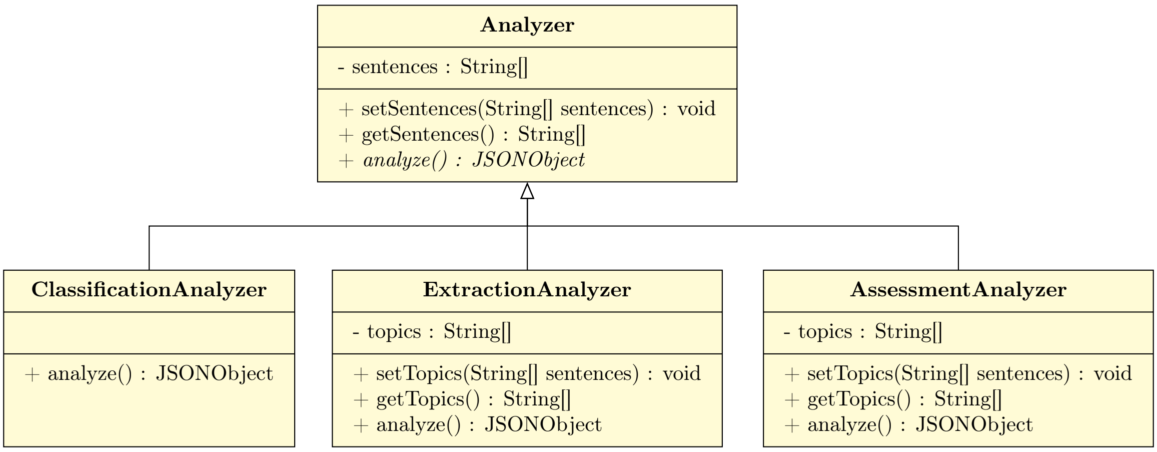


Figure 9.3.: Class diagram for Analyzer classes

We will again refrain from re-describing the internal implementation of the `analyze` functions, as their inner workings have already been described in Chapter 6. All analyze methods return JSON objects which contain a field `text`, in which the input clause is returned as array of strings. The `analyze` method in the `ClassificationAnalayzer` additionally returns an array of strings with the name `topics`, which contains the identified topics and subtopics. The `analyze` function in the `ExtractionAnalayzer` returns an array of JSON objects with the name `annotations`. The objects within the array have the same format as we used in the information extraction corpus described in Section 6.4.1. Finally, the `analyze` method in the `AssessmentAnalyzer` returns a boolean field `void`, which indicates whether the clause was assessed to be void. Additionally, all functions also return a field `source` in their response, which can be used to indicate by which classifier an assessment was made, e.g., a rule-based classifier. This allows us to easily plug new classification micro-services into our architecture, which can be distinguished by the web application through their source-tag.

9.1.3. ML Service

The ML service is implemented in Python and uses connexion to provide the API. The API has the same routes as the API of the rule service and provides the same responses. However, the responses are generated using the ML approaches, which we evaluated. For the topic classification, we only implemented the best performing ML classifier, i.e., the BERT classifier.

9.1.4. Database Service

The database service provides a middleware for accessing the MongoDB database. It is implemented in Node.js using the Mongoose library[3]. In addition to providing a REST API for accessing the MongoDB, it is also responsible for authentication and data validation by making sure that only valid data will be persisted to the database. The authentication is based on JSON Web Tokens. When a user logs in with correct credentials, they receive a token, which they send with each request to the database service to authenticate themselves. In this research prototype, we did not implement object-level access control, i.e., each registered user can access, modify, and delete all data, except for user objects, where each user can only modify and delete their own user object.

We persist four different types of information to the database:

- Users,
- Contracts,
- Knowledge-Base, and
- Settings.

In the following sections, we will shortly introduce each of these types individually and describe how the schemata that are used to validate and persist the data and which API routes are used to make the data accessible.

9.1.4.1. Users

For each user, we store the email, name, and password in our database (see Figure 9.4). The email is used as unique identifier for the object (∼primary key). The password is stored as salted hash using the bcrypt library (Sriramya and Karthika, 2015).

The REST API exposes four routes to interact with user objects:

- POST `/register`: Creates a new user object.
- POST `/login`: Returns a JSON Web Token if the submitted credentials are correct.
- GET `/user`: Returns all user objects (without password field).

[3] `https://mongoosejs.com/`, last accessed 2020-09-21

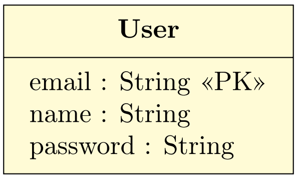


Figure 9.4.: Database schema for User object

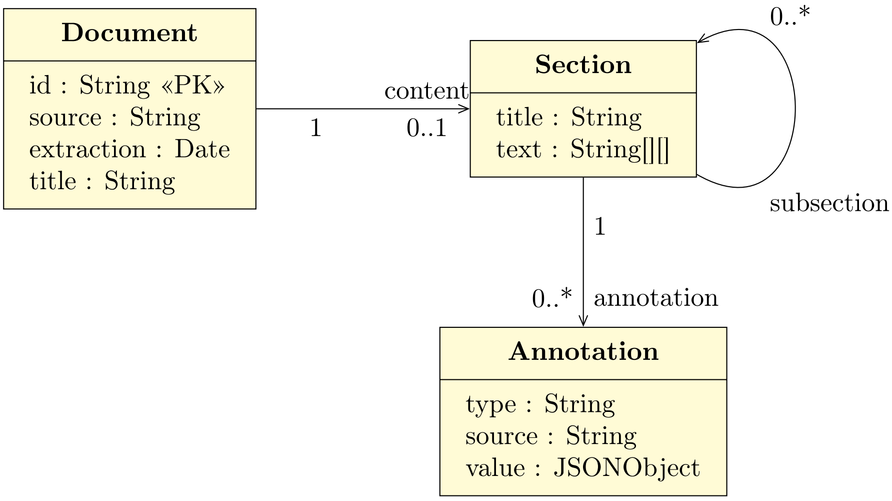


Figure 9.5.: Database schema for Document objects

- GET `/user/{email}`: Returns the user object with the given email address (without password field).
- PUT `/user/{email}`: Updates the user object with the given email address (can only be performed on one's own user object).
- DELETE `/user/{email}`: Deletes the user object with the given email address (can only be performed on one's own user object).

9.1.4.2. Contracts

The format in which contracts are persisted to the database is based on the format for the structured representation of contracts, which we introduced in Section 6.2.3.2. A field `id` was added in order to make direct access to single documents easier and distinct and annotations were added to sections. An Annotation object can hold all the annotations that are returned by the rule and ML Service (i.e., topic annotations and legal assessments), but can also be used to store user-generated annotations and even comments (see Figure 9.5).

The database service provides five routes to interact with contracts (i.e., Document objects):

- GET `/contract`: Returns a list of all contracts.

- POST `/contract`: Creates a new contract.
- GET `/contract/{id}`: Returns the contract with the given id.
- PUT `/contract/{id}`: Updates the contract with the given id.
- DELETE `/contract/{id}`: Deletes the contract with the given id.

Additionally, we added a route that returns statistics on the whole contract corpus currently stored in the database (GET `/stats`), like the total number of contracts, the number of contracts that have already been legally assessed, and the number of contracts which contain illegal clauses. We use this information in the dashboard of the prototype (see Section 9.2). While we could also calculate all these numbers in the client by simply downloading all contracts, for performance reasons, we decided to do this on the server.

9.1.4.3. Knowledge-Base

In order to persist the Knowledge-Base, we use the data schema which we described in Section 7.3: We use the topics as the main access point, which contain information about the subtopics, the information that can be extracted, and the rules that can be applied to the topic. Figure 7.2 in Chapter 7 shows a graphical representation of the schema.

The REST API provides five routes to interact with the knowledge-base:

- GET `/knowledge`: Returns all topic objects.
- POST `/knowledge`: Creates a new topic object.
- GET `/knowledge/{identifier}`: Returns the topic object with the given identifier.
- PUT `/knowledge/{identifier}`: Updates the topic object with the given identifier.
- DELETE `/knowledge/{identifier}`: Deletes the topic object with the given identifier.

9.1.4.4. Settings

Finally, the database service also provides a storage point for the settings of the web application. For this, we use a key-value store, as shown in Figure 9.6.

The respective routes provided by the Database Service are:

- GET `/settings`: Returns all settings objects.
- POST `/settings`: Creates a new settings object.
- GET `/settings/{key}`: Returns the settings object with the given key.
- PUT `/settings/{key}`: Updates the settings object with the given key.
- DELETE `/settings/{key}`: Deletes the settings object with the given key.

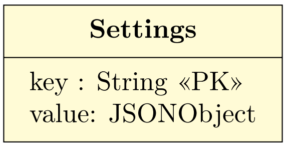


Figure 9.6.: Database schema for Settings objects

In our research prototype, the only settings we are actively using are the base URLs to all services, which are stored as settings, and a language setting for the user interface, which can be provided in either German or English. Notably, the routes to the rule and ML service are missing a language parameter to switch between English and German. The initial import of the different language models is the most time-consuming task for both the ML service and the rule service (where the language model for the dependency parser has to be loaded). Switching between both languages would be very time-consuming. In practice, we believe it is very unlikely that users want to switch between languages (regularly). Therefore, we decided that the services can be started with a command-line parameter that sets the language, which is then fixed during runtime.

9.1.5. Web Application

Before we take a look at the UI of the web application in the next section, we first introduce its general architecture. The app is implemented with Angular 9 and consists of five main views: a dashboard which serves as landing page, a page for the administration of the settings, a page for administrating and editing the knowledge-base, a page for the database of all contracts, and the main view in which contracts are analyzed. Each of these views is implemented as an Angular component in addition to a central app component, as shown in Figure 9.7. In addition, there is a component for the navigation bar, and alerts that can be displayed to provide user feedback, e.g., when a contract was imported successfully. Some components also have sub-components to render data. Additionally, the app has a central routing module and six services, which contain centralized functionalities that are injected into all components, like user management or database connection.

In order to illustrate the inner workings of the app and how it orchestrates the different micro-services, we take a closer look at two use cases: importing a new contract into the system and adding a manual annotation to a clause.

9.1.5.1. Import

We assume the user is already logged in and opened the database component view. In order to import a new contract from a URL into the system, the user opens the import modal dialog by clicking on the “Import” button, which starts the sequence shown in Figure 9.8. From the import dialog, the URL provided by the user is sent to the document service. The document service first identifies the URL of the T&C page from the given URL. For this, the AI service is used, which communicates with the rule, ML, and pre-processing service.

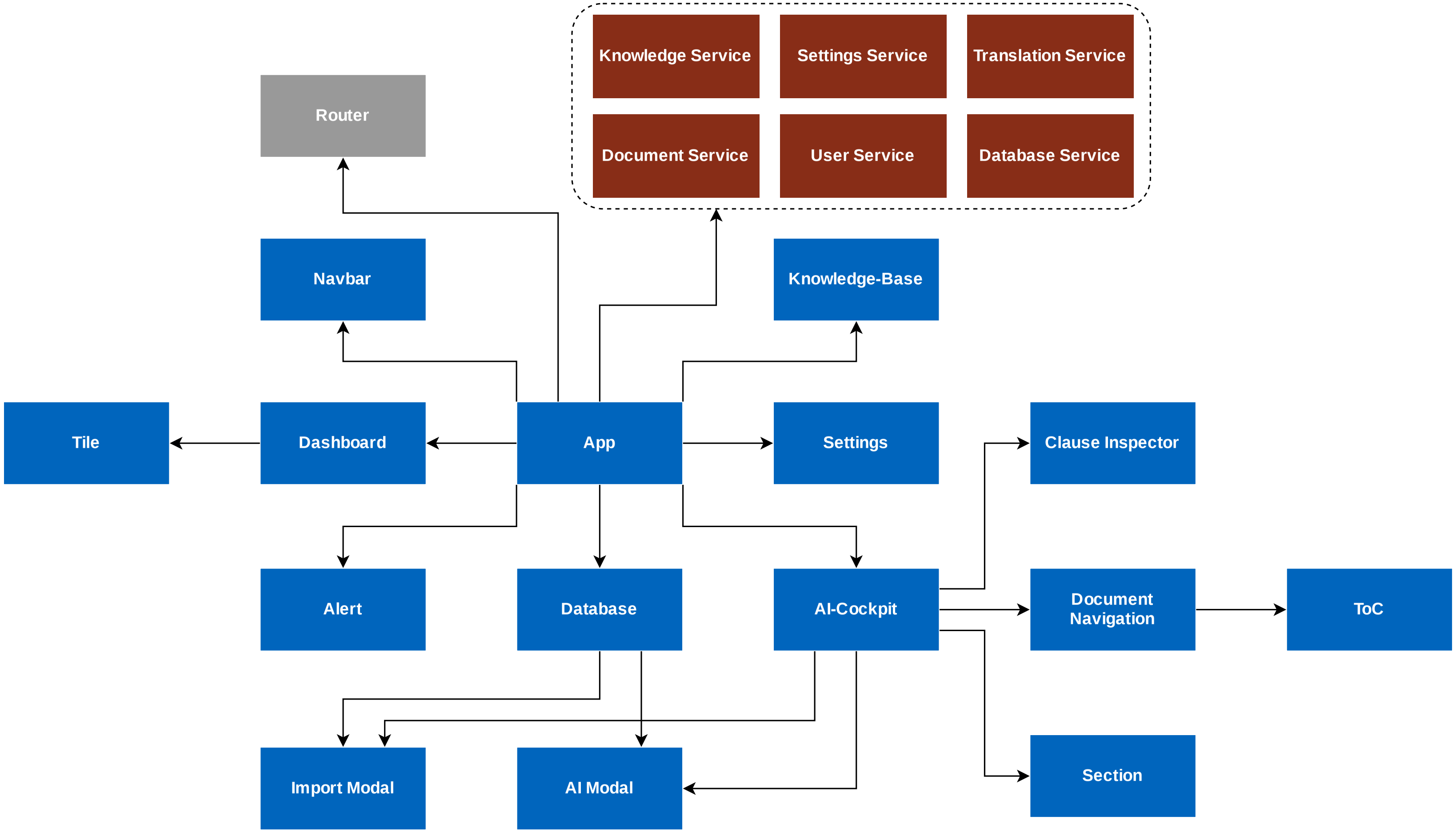


Figure 9.7.: Building blocks of the web application (blue = component, gray = model, red = service)

Once the AI service has retrieved the URL from the pre-processing service, it calls the server again with this URL in order to extract the content from the T&C page. The result sent by the pre-processing service is the contract in structured JSON format. This contract is sent back to the document service and from there to the import modal. The import modal now calls the `annotate` function of the document service, which first calls functions from the AI service for the topic classification, the information extraction, and the legal assessment. In the sequence in Figure 9.8, the AI service calls only the rule server to perform this classification, but it could also call the ML server in the same way.

The document service then adds all the annotations to the contract object and passes it back to the import modal. In the last step, the contract is stored in the database, using the database service of the web application and the API of the database server that persists the object to the MongoDB.

9.1.5.2. Annotate

For the use case of submitting a manual annotation, we assume that the user is already logged in, in the AI-Cockpit, and has selected a clause. Once a clause is selected, it is opened in the clause inspector, which is a sub-component of the AI-Cockpit.

By clicking the "Add +" button in the clause inspector, the `addAnnotation` method is triggered, which adds an empty form to the UI, in which the user can enter the type and value of the annotation (e.g., "topics" and "withdrawal"). Once the information has been entered, it can be persisted by clicking the "Save" button. The button invokes the `submitAnnotation` method and adds the annotation to the local copy of the clause, which is currently loaded in the clause inspector and thereby also to the local copy of the contract.

Since the local copy of the contract object is already updated, the only necessary method call is to the document service, which passes the contract on to the database service, which in turn calls the API of the database server. Since we do not create a new contract but update an existing contract, unlike in Figure 9.8, this time, we send a PUT request to the database server.

9.2. User Interface

We designed the UI of the web application to suit the needs of consumer advocates who are experts in their domain but not necessarily in the application of software tools. In order to meet their needs best, we started the design process with very low-fidelity wireframe models (see Figure 9.10) and iterated based on them in short cycles until we reached a final design.

Figure 9.11 shows the final design of the dashboard in our prototype. It provides access to the four main views of the application:

- **AI-Cockpit:** This is the main view in which experts interact with a contract. They can review automatically generated annotations, use automated and manual annotations to navigate through the document, and add manual annotations.

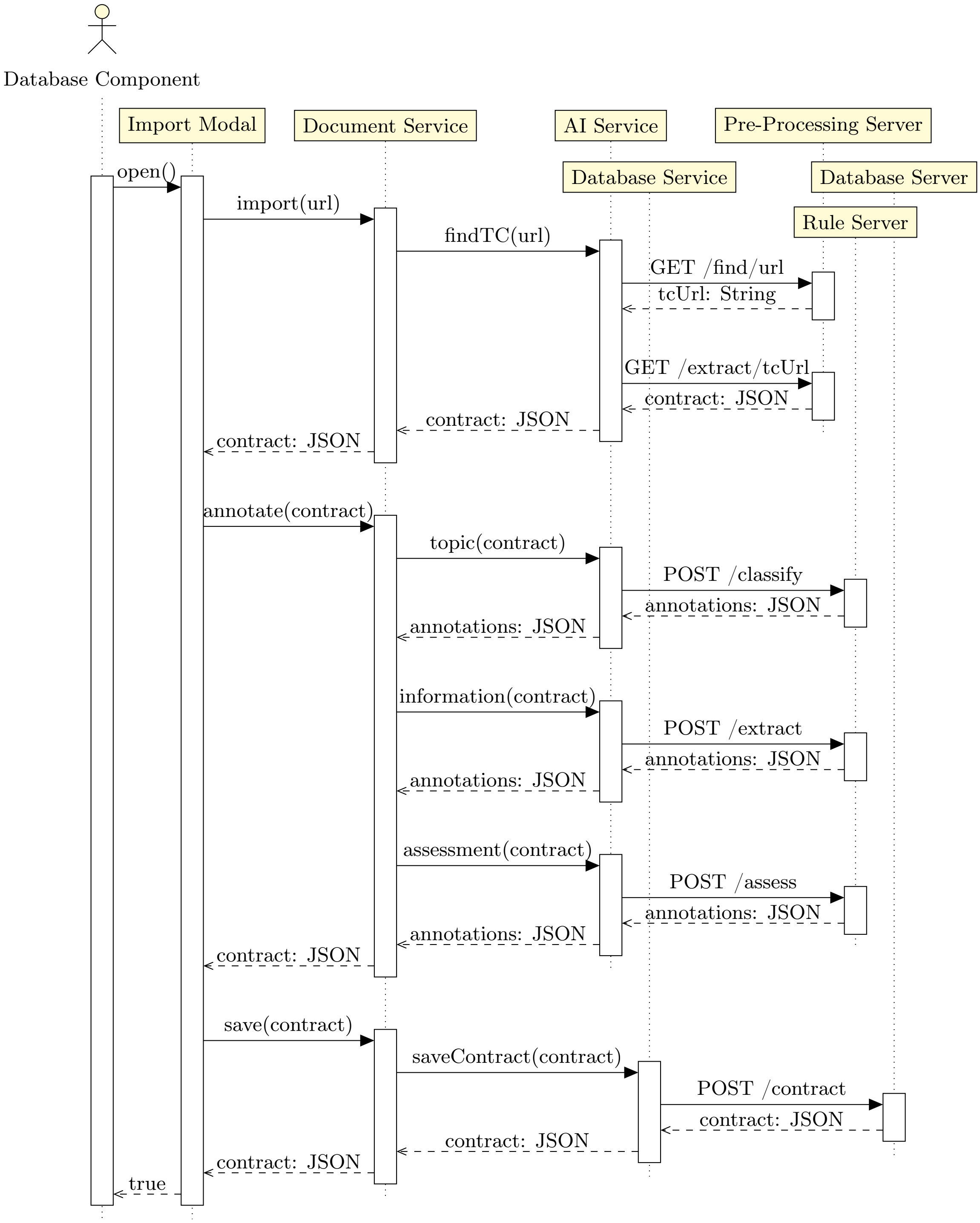


Figure 9.8.: Sequence diagram for importing a new contract into the prototype

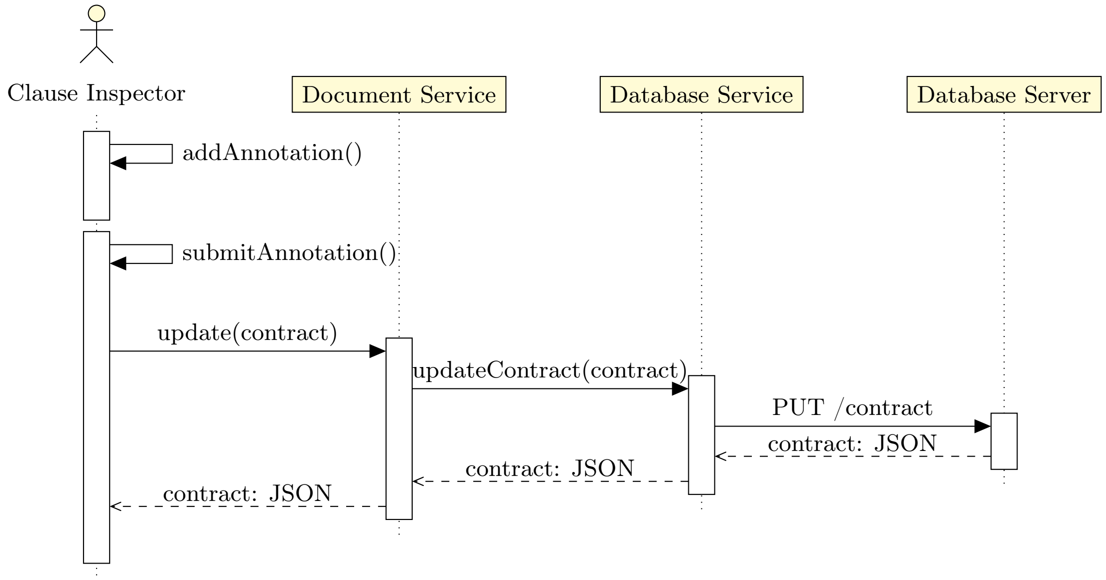


Figure 9.9.: Sequence diagram for adding a manual annotation to a clause

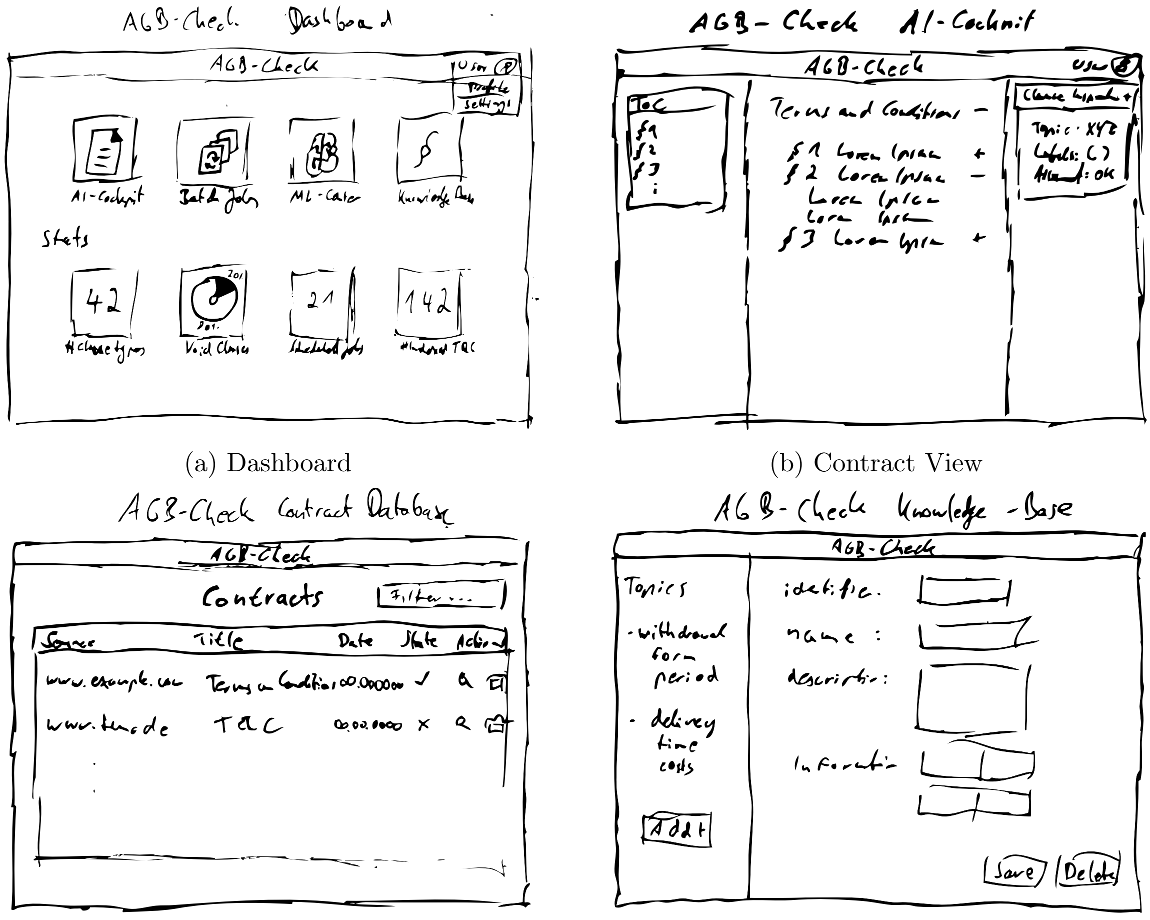
- **Contract Database:** This view provides a list of all the contracts in the system and their assessment status. From here, new contracts can be imported into the system, and their automated annotation can be triggered.
- **Knowledge-Base:** This view provides the users with means to alter the knowledge-base in order to change the rule-based analysis.
- **Settings:** Basic settings for the web application can be changed in this view.

In addition to providing links to these four views, the dashboard also contains a section with statistics, which gives a high-level overview of the corpus of contracts that are currently imported into the system by providing the total number of contracts, the number of contracts that have already been analyzed (i.e., completely legally assessed), and the number of contracts that contain void clauses.

In this section, we want to focus on the two views which are the main interaction points for the consumer advocates, the contract database (Section 9.2.1) and the AI-Cockpit (Section 9.2.2).

9.2.1. Contract Database

The UI of the Contract Database is shown in Figure 9.12. The main element is a table that provides a list of all contracts which are currently imported into the system. The table can be filtered with a search field at the bottom and ordered by the source of the contract (i.e., the URL from which it was downloaded), its title, the date when it was imported into the system, and its status. The status is indicated by a colored square and can take three different states:



(a) Dashboard

(b) Contract View

(c) Contract Database

(d) Knowledge-Base

Figure 9.10.: Wireframe models of the UI

- Green: All clauses have been legally assessed, and none are void.
- Red: At least one clause has been assessed as void.
- Gray: Not all clauses have been assessed yet, but none is void so far.

There are three actions available per contract: The contract can be opened in the AI-Cockpit (see Section 9.2.2), can be (re-)assessed by the AI or deleted from the database.

A click on the “Import” button opens the import modal shown in Figure 9.13. Currently, only the import of web pages is supported. In the future, we would also like to support the import of PDF and other document formats. The links that are entered into the text box can be either direct links to T&C pages or general links, in which case the pre-processing service will first try to identify the T&C pages. The legal analysis can be performed directly during the import or later through the action button in the table.

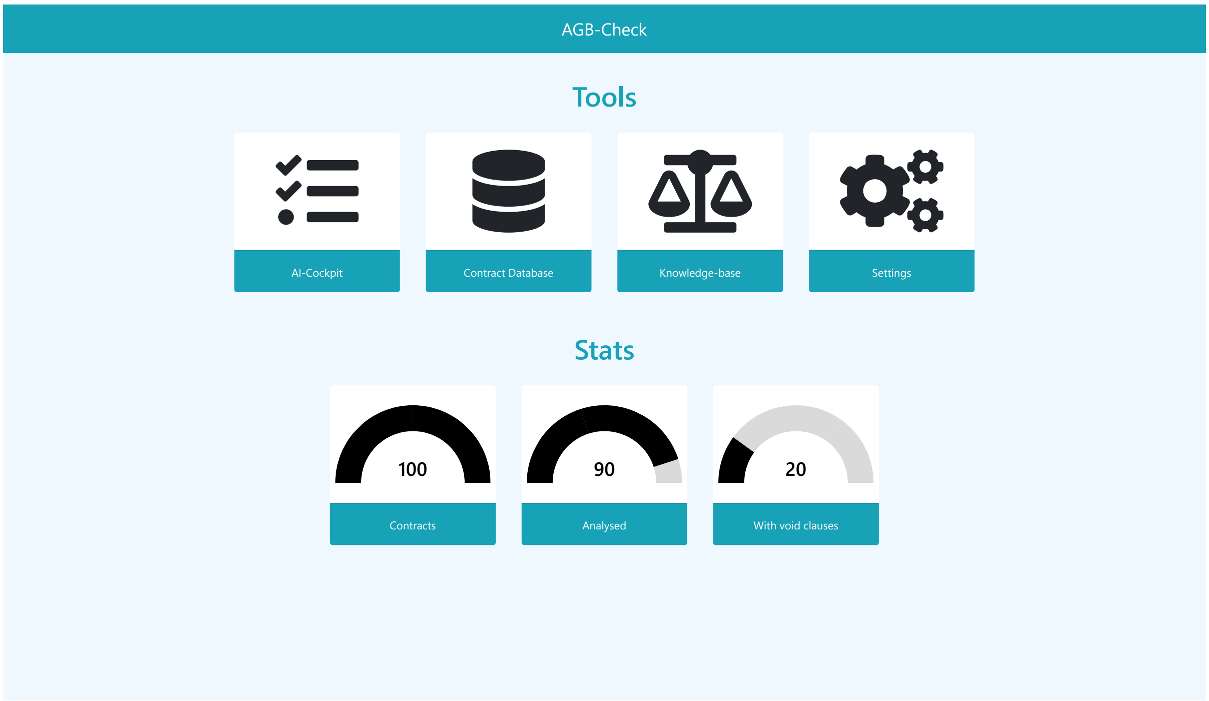
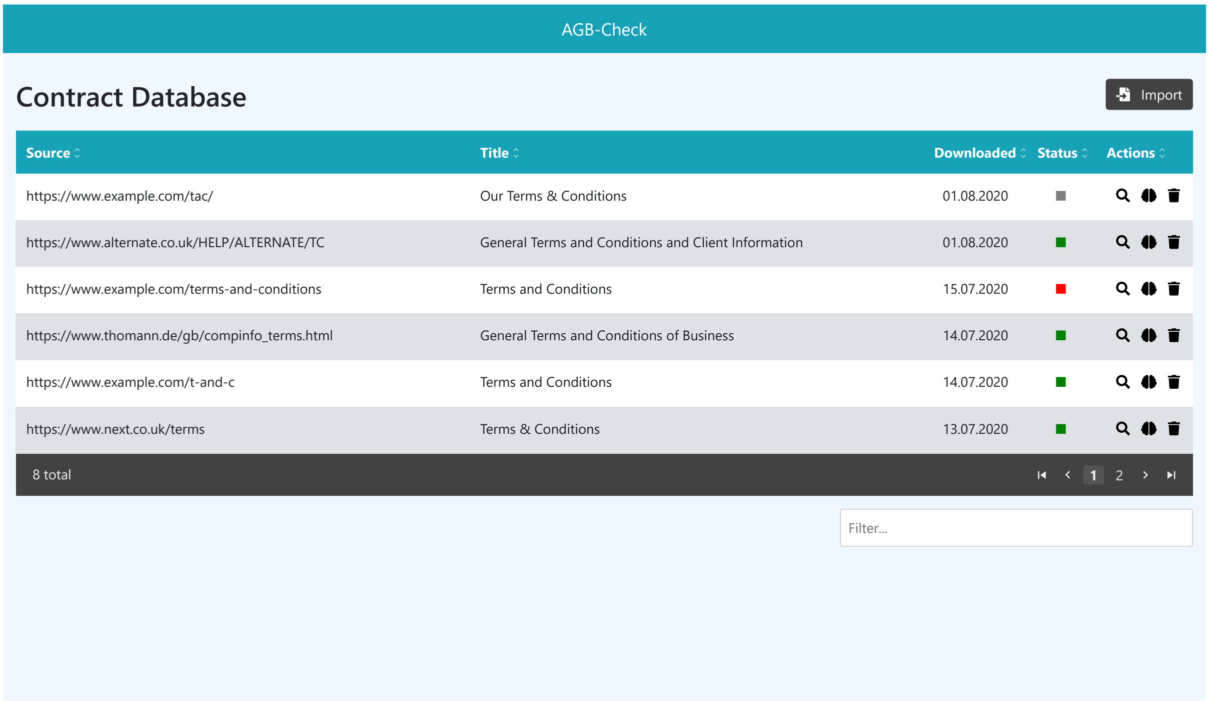


Figure 9.11.: Screenshot of the dashboard

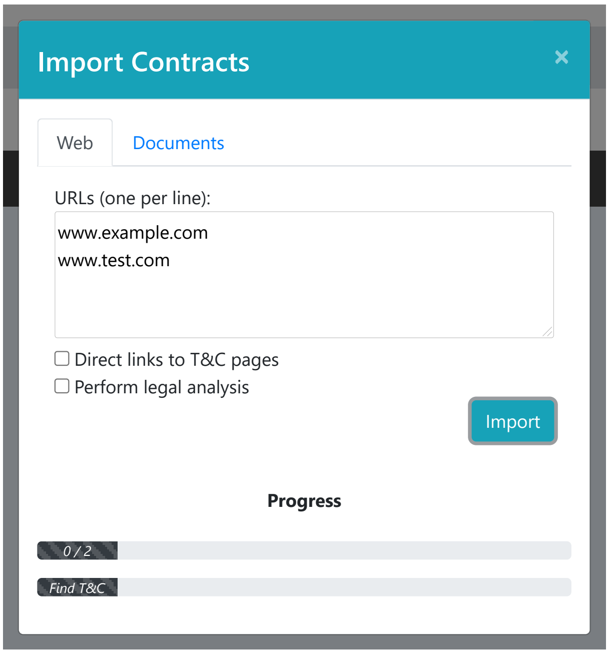


Contract Database

Source	Title	Downloaded	Status	Actions
https://www.example.com/tac/	Our Terms & Conditions	01.08.2020	◼	🔍 🧠 🗑
https://www.alternate.co.uk/HELP/ALTERNATE/TC	General Terms and Conditions and Client Information	01.08.2020	🟩	🔍 🧠 🗑
https://www.example.com/terms-and-conditions	Terms and Conditions	15.07.2020	🟥	🔍 🧠 🗑
https://www.thomann.de/gb/compinfo_terms.html	General Terms and Conditions of Business	14.07.2020	🟩	🔍 🧠 🗑
https://www.example.com/t-and-c	Terms and Conditions	14.07.2020	🟩	🔍 🧠 🗑
https://www.next.co.uk/terms	Terms & Conditions	13.07.2020	🟩	🔍 🧠 🗑

8 total

Figure 9.12.: Screenshot of the contract database



Import Contracts

URLs (one per line):

www.example.com
www.test.com

☐ Direct links to T&C pages
☐ Perform legal analysis

Figure 9.13.: Screenshot of the import modal

9.2.2. AI-Cockpit

Figure 9.14 shows the AI-Cockpit, i.e., the view that is presented to the user when they open a contract. The view consists of three components: The contract text is shown in the middle, on the left side, the document navigation is shown, and on the right side, the clause inspector. The contract text is shown as a structured tree, which can be collapsed and expanded based on the hierarchy of the headings.

At the top of the document navigation, all clauses can be expanded or collapsed at once. Below that, the table of contents is shown, which is automatically derived from the document structure. The main view automatically scrolls to the position of the clause when it is selected in the table of contents. If it is currently collapsed, it will also be automatically expanded. Below the table of contents, a list of topics is shown. This list is derived from the topic annotations in the document, whether they have been created automatically or manually. When a topic is selected (indicated by the square symbol in front of it), all collapsed instances of the topic are expanded, and all clauses containing the topic are graphically highlighted (like clause 2.1 and 2.2 in Figure 9.14). Finally, the document navigation contains a table with general statistics on the currently opened contract, including the number of clauses that have been annotated as void, the total number of clauses in the contract, and the number of words in the contract.

If a clause is selected in the text view (indicated by bold font), the details of the clause are shown in the clause inspector on the right side of the screen. The Clause inspector shows the existing annotations and allows users to add new annotations, which can either be a legal assessment, a topic label, or a general comment. Clauses that have been annotated as void will be highlighted in the text view by red bars.

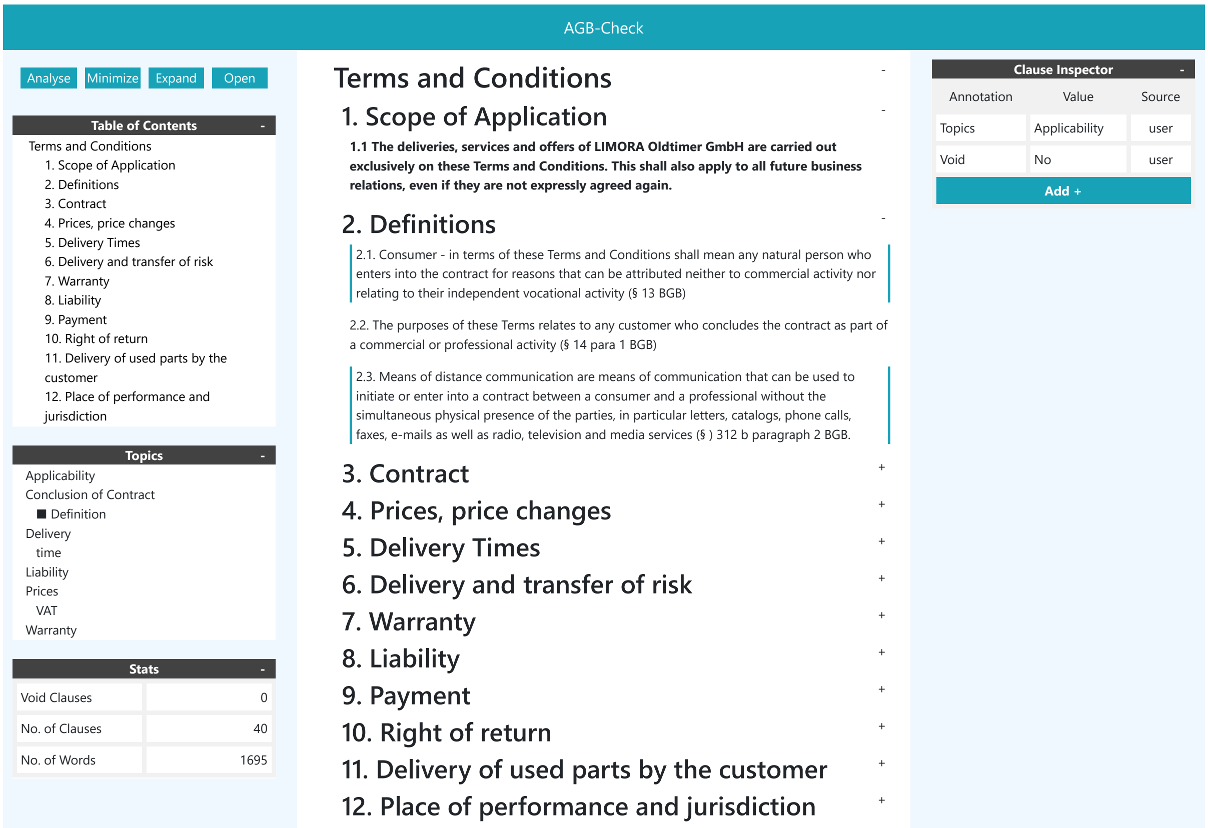


Table of Contents

Terms and Conditions
1. Scope of Application
2. Definitions
3. Contract
4. Prices, price changes
5. Delivery Times
6. Delivery and transfer of risk
7. Warranty
8. Liability
9. Payment
10. Right of return
11. Delivery of used parts by the customer
12. Place of performance and jurisdiction

Topics

Applicability
Conclusion of Contract
■ Definition
Delivery
time
Liability
Prices
VAT
Warranty

Stats

Void Clauses	0
No. of Clauses	40
No. of Words	1695

Terms and Conditions

1. Scope of Application

1.1 The deliveries, services and offers of LIMORA Oldtimer GmbH are carried out exclusively on these Terms and Conditions. This shall also apply to all future business relations, even if they are not expressly agreed again.

2. Definitions

2.1. Consumer - in terms of these Terms and Conditions shall mean any natural person who enters into the contract for reasons that can be attributed neither to commercial activity nor relating to their independent vocational activity (§ 13 BGB)

2.2. The purposes of these Terms relates to any customer who concludes the contract as part of a commercial or professional activity (§ 14 para 1 BGB)

2.3. Means of distance communication are means of communication that can be used to initiate or enter into a contract between a consumer and a professional without the simultaneous physical presence of the parties, in particular letters, catalogs, phone calls, faxes, e-mails as well as radio, television and media services (§) 312 b paragraph 2 BGB.

3. Contract
4. Prices, price changes
5. Delivery Times
6. Delivery and transfer of risk
7. Warranty
8. Liability
9. Payment
10. Right of return
11. Delivery of used parts by the customer
12. Place of performance and jurisdiction

Clause Inspector

Annotation	Value	Source
Topics	Applicability	user
Void	No	user

Figure 9.14.: Screenshot of the final AI-Cockpit design

CHAPTER 10

Evaluation

Throughout this thesis, we have conducted numerous evaluations. For each step of our NLP-pipeline, we have evaluated and compared multiple approaches. We have evaluated two approaches for the legal assessment of clauses, and evaluated the generated summaries of standard form contracts. So far, all evaluations, except for the evaluation of the summaries, have been intrinsic evaluations of individual components. Intrinsic evaluation criteria relate to the main objective of a system or component (Jones and Galliers, 1996, p. 19), in our case the main objectives were the automatic detection, segmentation, topic classification, legal assessment, and summary generation. While these are important measurements and indications for how well the system works in theory, they do not necessarily reflect how useful or successful the system will be in practice. However, especially for the action research approach we are following, this practical utility is an important measurement. Therefore, we conducted an extrinsic evaluation with the prototype described in Chapter 9. In an extrinsic evaluation, the functionality of a system within a given task is evaluated (Jones and Galliers, 1996, p. 19). We developed three different tasks, which are based on the daily work of consumer advocates and presented them to our participants. Each task had to be solved twice, once with support from the system and once without. The setup of the experiment is described in detail in Section 10.2.

The good results we achieved in the intrinsic evaluation are a necessary prerequisite for a successful extrinsic evaluation but do not guarantee success. If the system is unable to correctly identify clause topics, it is very unlikely that it will be useful in solving tasks related to the analysis of standard form contracts. However, just because the system is able to correctly do so does not automatically mean that it will also be helpful in solving the tasks consumer advocates face in their daily work.

We have made this observation in one of our previous projects, where we developed a system to support first-level customer support via email. We developed a system that classifies the topic of incoming requests and extracts important information from them. In the intrinsic evaluation, we found that the system is successful in performing those tasks. (Legenc, 2018) In the next step, we used this information to automatically generate reply templates for customer support workers. We implemented the functionality, the topic classification, information extraction, and reply template generation into a prototype and conducted an evaluation with customer support workers in which they replied to incoming requests using the prototype in two different conditions, one with the functionality activated and one without the functionality. Although we found that the classification, extraction, and template generation worked well, we did not see any positive effects on task completion. The features neither improved the quality of the responses nor did they reduce the time which was needed to send the responses. (Weißl, 2018)

In order to make sure that the functionalities we developed as part of our prototype can also have an impact on the practical work of consumer advocates, we conducted this additional, extrinsic, task-based evaluation.

10.1. Hypothesis

This evaluation is an important part of the test of our main research hypothesis that AI can help finding, analyzing, and assessing standard form contracts, as well as an important step towards answering our eighth research question whether AI-based automation and support tools can improve the ways consumer advocates work with standard form contracts.

There are mainly two ways in which we can directly improve the work of consumer advocates related to standard form contracts: We can either help them to be faster, or we can help them to produce better results by reducing the number of mistakes they make. We believe that it is very unlikely that our system will be able to improve the results a consumer advocate produces, given sufficient time. We do, however, believe that our system can help consumer advocates to work faster. We also believe that indirectly, this will increase the quality of the results because, given more time, they will be able to check aspects of contracts they were not able to assess before due to time constraints, hence resulting in a more thorough analysis.

Our main hypothesis for this experiment was, therefore, that the tool will help the participants to perform tasks related to the legal assessment of standard contracts faster, without changing the quality of the results. More concretely:

- **Hypothesis 1:** The tool will help the participants to find relevant clauses faster.
- **Hypothesis 2:** The tool will help the participants to legally assess (individual) clauses faster.
- **Hypothesis 3:** The tool will help the participants to legally assess complete standard form contracts faster.

10.2. Experimental Design

In this section, we will describe how we designed our experiment in order to test these three hypotheses. At the heart of our experiment is a task-based evaluation with an augmented version of the prototype described in Chapter 9, which logs timestamps for each interaction with the UI and thereby enables a quantitative analysis of the impact of the system. In addition, we also performed a qualitative analysis by interviewing the participants after they completed all tasks.

10.2.1. Tasks

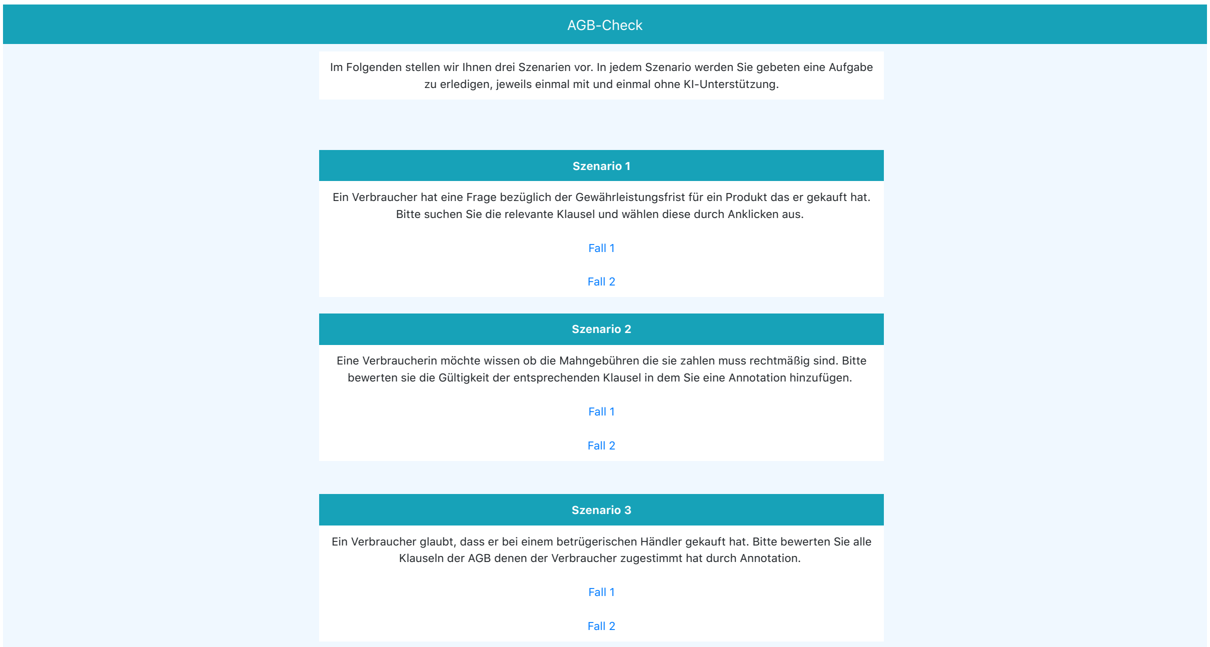
In order to test our three hypotheses we developed, we designed three different tasks, which derive naturally from the hypotheses: find a clause of a specific topic, assess a specific clause legally, assess all clauses of a contract legally. In order to make the tasks feel more natural to the participants, we wrapped them in three scenarios:

- **Scenario 1:** A consumer has a question about the warranty period for a product they bought. Please find the relevant clause in the contract and select it by clicking on it.
- **Scenario 2:** A consumer wants to know whether the reminder fees they have to pay for being late with their payment are legal. Please assess the relevant clause by adding an annotation.
- **Scenario 3:** A consumer believes they bought from a malicious dealer, please assess the T&C they agreed to legally, by adding an annotation to every clause in the contract.

For each scenario, we compiled two different contracts by mixing clauses from our existing corpus, which already has been legally assessed. Both compiled contracts are designed to have the same number of clauses and approximately the same length. Each participant had to conduct every scenario twice, once with activated AI support (i.e., automatic topic annotation, legal assessment, and all support features of the web application) and once without any support (i.e., no automatic annotation and the Document Navigator in the AI-Cockpit was deactivated). In this way, we made sure that the only changing variable between the two executions of the scenario was the absence or presence of the support features.

The order (i.e., whether they first had to conduct the scenario with or without support features) was randomized, and so was the assignment of the two contracts to the two conditions. In each instance, we measured the time from opening the contract to completing the specified action.

In all three scenarios, the automatically generated topic classifications were correct. For Scenario 2, in both contracts, the clause about late fees was void and no other clause in the contract was void, which was also correctly annotated by the system. In Scenario 3, both contracts contained three void clauses, however, in both cases, only two were identified as void by the system, i.e., both contracts contained one false negative.



AGB-Check

Im Folgenden stellen wir Ihnen drei Szenarien vor. In jedem Szenario werden Sie gebeten eine Aufgabe zu erledigen, jeweils einmal mit und einmal ohne KI-Unterstützung.

Szenario 1

Ein Verbraucher hat eine Frage bezüglich der Gewährleistungsfrist für ein Produkt das er gekauft hat. Bitte suchen Sie die relevante Klausel und wählen diese durch Anklicken aus.

Fall 1

Fall 2

Szenario 2

Eine Verbraucherin möchte wissen ob die Mahngebühren die sie zahlen muss rechtmäßig sind. Bitte bewerten sie die Gültigkeit der entsprechenden Klausel in dem Sie eine Annotation hinzufügen.

Fall 1

Fall 2

Szenario 3

Ein Verbraucher glaubt, dass er bei einem betrügerischen Händler gekauft hat. Bitte bewerten Sie alle Klauseln der AGB denen der Verbraucher zugestimmt hat durch Annotation.

Fall 1

Fall 2

Figure 10.1.: Task presentation for participants

10.2.2. Interview Questions

After finishing all scenarios, we asked the participants the following six questions:

1. Do you think that the system is easy to use?
2. Do you think the features offered by the current prototype can help you in the analysis of standard form contracts?
3. If so, do you think they can help you to be faster or find more void clauses or both?
4. Do you think wrong annotations from the system could decrease the quality of your work?
5. Which feature was most useful to you?
6. Are there any other features which you would like to see in the prototype?

10.2.3. Execution

The evaluation was conducted in one-on-one online sessions, using the open-source web conference tool BigBlueButton, with five of the seven experts introduced in Chapter 5, all of which are fully qualified lawyers. We started each session with a presentation of the prototype and its functionalities through screen sharing, which took about ten minutes. Then, participants were given a link to the prototype and asked to share their screen. For around ten minutes, they were asked to familiarize themselves with the system and ask any questions they might have. After this, they were presented with the scenarios (see Figure 10.1) and were asked to work through

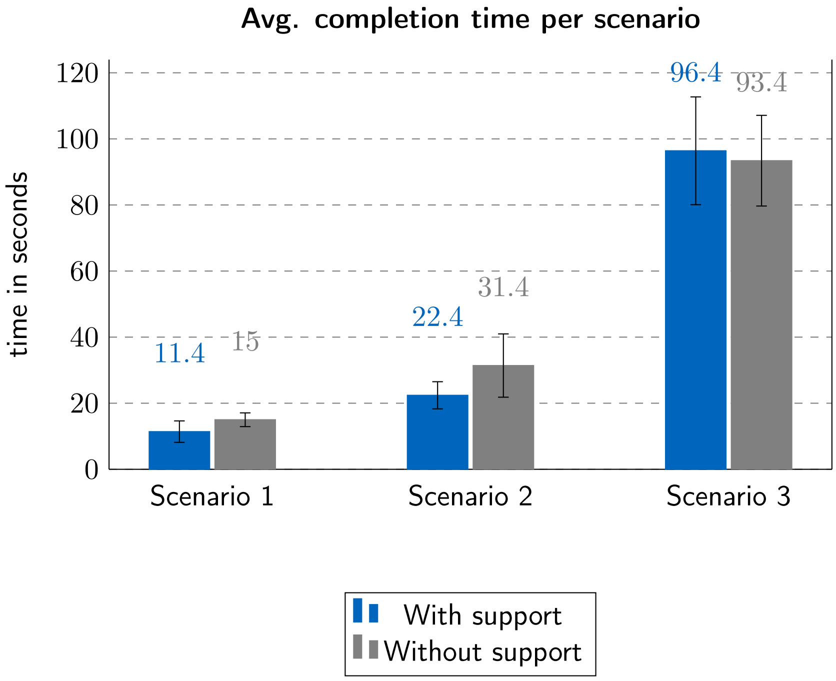


Figure 10.2.: Average completion time in seconds per scenario

them, one after another. After the last scenario was completed, the session was concluded with the interview.

10.3. Results

In the next sections, we will first discuss the qualitative results of the evaluation, which are based on the completion of the three scenarios, then discuss the quantitative results based on the responses to the interview questions, and finally, we draw a conclusion based on the results.

10.3.1. Quantitative

Figure 10.2 shows the average completion time per task depending on whether or not the support features have been activated. For Scenario 1 and 2, the average completion time was reduced by 25% and 29% when the support features were activated. For Scenario 3, the completion time was slightly increased when the support features were active. In all three scenarios, we found that all participants made use of support features in the Document Navigator (i.e., either the paragraph or topic navigation) when they were offered.

For Scenarios 1 and 2, the participants knew (or could at least reasonably assume) that there will be only one relevant clause. Therefore, when they used the support features to find that clause and were successful, they knew that they did not have to look further. In these scenarios,

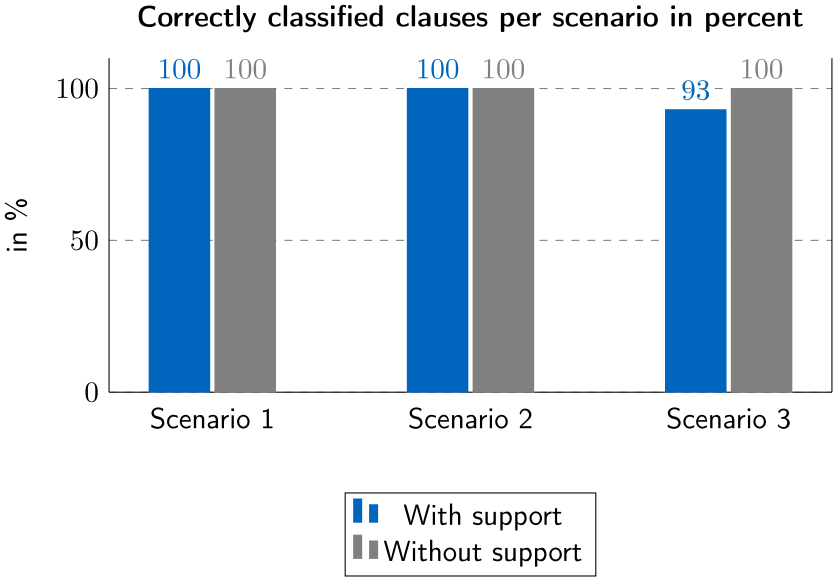


Figure 10.3.: Correctly classified clauses per scenario in percent

the output of the system could be directly validated by the participants. In Scenario 3, however, they did not know whether the results provided by the system were exhaustive. We found that the participants were faster to annotate the two clauses as void that were also identified as void by the system when the support features were activated. However, for the clause that was wrongly not identified as void by the system, it took them longer to find and assess it when the support features were activated.

In Scenario 1 and 2, all participants chose the right clauses and made the right annotations, independent of whether the support features were activated or not (see Figure 10.3). In Scenario 3, one participant did not classify the void clause, which was overlooked by the system, correctly, despite the fact that the participant already clicked on the clause and investigated it closely. While this is only a single instance, it does emphasize the danger that users might rely too much on the system and its annotations, especially in cases where they can not quickly verify all annotations made by the system.

10.3.2. Qualitative

Figure 10.4 shows a categorization of the responses to our interview questions. For the first question ("Do you think that the system is easy to use?"), four participants said they did find the system easy to use; one participant said that they were not sure yet and would need to use the system longer, before they can make an assessment.

For the second question, all five participants agreed that the system and its current features could help them in the analysis of standard form contracts, which supports the results of the quantitative analysis. Since all participants responded with yes to the second question, we also asked them the third question ("If so, do you think they can help you to be faster or find more void clauses or both?"). Three participants replied that they believe the system can help them to assess clauses faster; two replied that it would help them to be faster and find more clauses. We asked the three participants who replied only with faster whether they think by being faster, they will have more time and therefore also be able to find more void clauses. Two participants agreed, but they did not reply with both because they consider this only a side effect of the system. One participant disagreed, saying that they always take the time that is necessary for a thorough analysis and being faster would not change that.

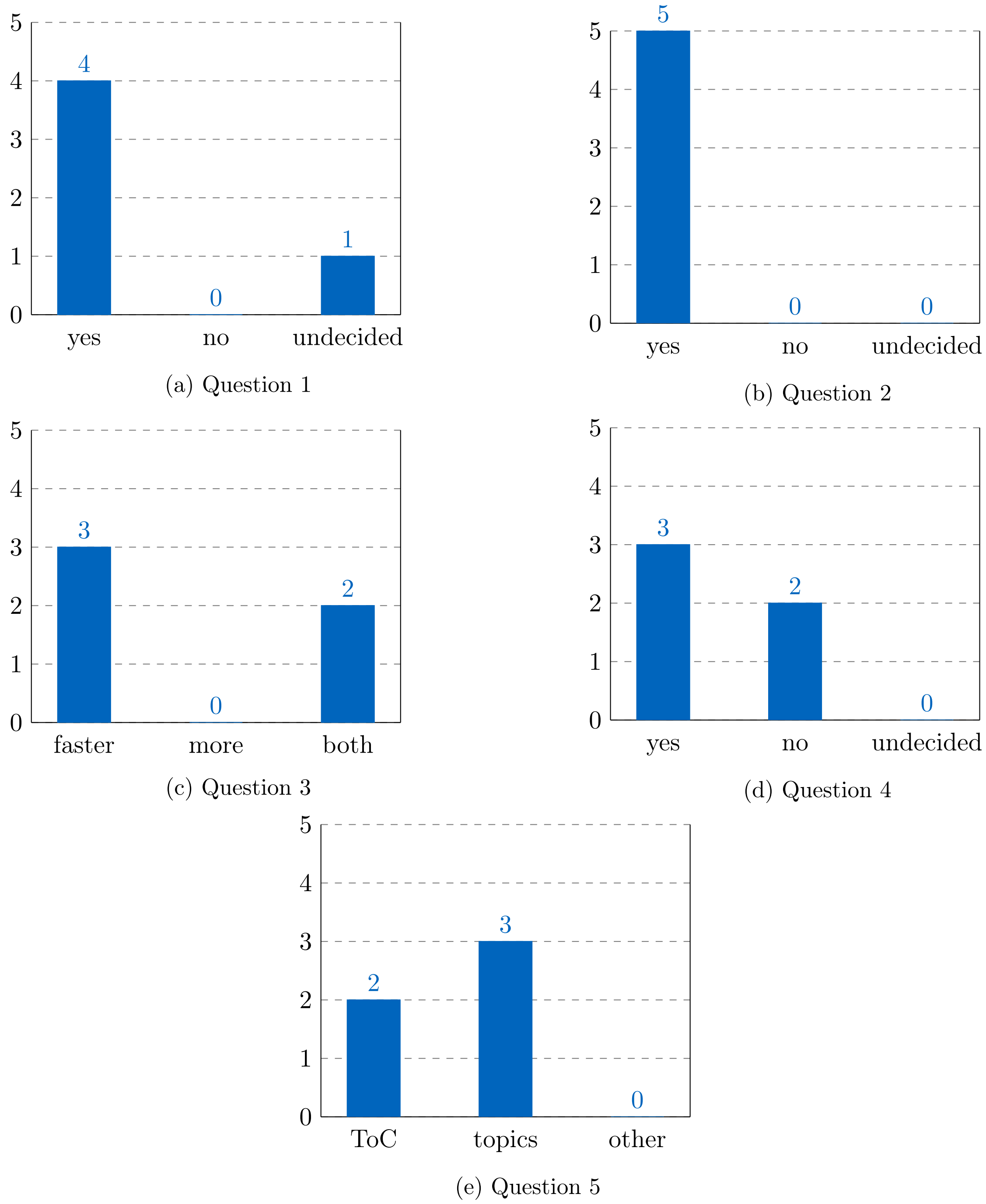
When asked directly ("Do you think wrong annotations from the system could decrease the quality of your work?"), three out of five participants were worried that wrong annotations from the system could decrease the quality of their assessment. One participant added that they would be worried that the introduction of such a tool would increase expectations about how many consumer inquiries they can process and hence put pressure on them that could lead to mistakes. Another participant was worried that the fact that the system always provides a definite answer (void or valid) could put them under pressure in cases where they do not want to give a yes or no response in situations that are legally not completely clear.

When asked about their favorite feature, all participants referred to navigation features. Two participants replied that the automatically generated table of contents was most useful to them and three replied that the automatically generated topic list was most useful to them. Notably, none of the participants mentioned the automated legal assessment.

Finally, when we asked about additional features participants would like to see in the prototype, one participant stated that they would like to have a third option for the legal assessment of clauses, to mark clauses they believe need further investigation, but are not sure about whether they are void or not.

10.3.3. Conclusion

Based on the results of the quantitative and qualitative evaluation, Hypothesis 1 could be confirmed. The tool did help participants to find relevant clauses faster, and participants also perceived that the tool is helpful in regard to this. Hypothesis 2 can also be confirmed because participants were indeed faster in assessing individual clauses while maintaining the same quality. However, the improvement purely came from them being able to find the relevant clause faster. Once they found the clause, the assessment process itself was not faster than before. Based on the results, Hypothesis 3 has to be rejected. The tool did not help participants to assess complete standard form contracts faster. However, it is unclear how long-term usage of the system might influence this. At some point, users might start to trust the system and hence become faster, but also more likely to adopt wrong annotations from the system. Therefore, the long-term effects of using such a system are an interesting research point for future work.



(a) Question 1

(b) Question 2

(c) Question 3

(d) Question 4

(e) Question 5

Figure 10.4.: Categorized number of responses to the interview questions

CHAPTER 11

Conclusion

We started this thesis with the research hypothesis that "AI can help consumers and consumer advocates in finding, analyzing, and assessing standard form contracts.". The evaluations we reported in Chapter 8 (for consumers) and Chapter 10 (for consumer advocates) present evidence to support this hypothesis. In this final chapter, we will revisit the research questions we presented in Chapter 1 and briefly summarize the answers we found in the course of this thesis. We will critically reflect on the limitations of our work and the insights we gained during our research, with regard to consumer protection and interdisciplinary collaboration, before we conclude with an outlook on possible (and already existing) follow-up work.

11.1. Summary

In Chapter 1, we motivated the research topic of this thesis by highlighting the importance of standard form contracts to our economy and the problems consumers and consumer advocates are facing. We also introduced T&C from online shops as the focus of our work and motivated the choice. We formulated the main research hypothesis for this thesis and derived eight research questions from it. Subsequently, we introduced the research methodology and described the contributions and the outline of this thesis.

In Chapter 2, we introduced the different types of statistical classifiers that are relevant to this thesis, as well as basic NLP-concepts that are relevant. We also offered a brief overview of national and European regulations on the drafting of standard form contracts, as well as on the standing of consumer protection agencies.

In order to position our work in comparison to existing work, we presented an overview of the related work in Chapter 3, in which we not just analyzed scientific results but also existing commercial products. From the field of science, we found the CLAUDETTE project from Lippi et al. (2019c) to be most closely related to our work, yet different in important aspects. We also found that, while recent scientific works tend to focus on ML technology, many of the commercial tools are still based on simpler techniques.

Chapter 4 elaborates on the general relevance of standard form contracts and more specifically from the perspective of consumers and consumer advocates. It presents a taxonomy for the classification of clause topics in standard form contracts, which we derived from different sources. Based on this taxonomy, we report which clauses of standard form contracts are most important to consumers and consumer advocates. Finally, we assessed some of the potential unwanted consequences a system like the one we designed could have when applied in the real world and explained how we are trying to mitigate these possible negative consequences.

We analyzed the existing processes in two German consumer protection agencies, with regard to the processing of standard form contracts, and formalized them in a process model which we presented in Chapter 5. Additionally, we gathered requirements for a tool to support the legal assessment of standard form contracts.

In Chapter 6, the largest chapter of this thesis, we introduced the individual building blocks of our NLP-pipeline. For every stage of the pipeline, we identified the most suitable technologies for German and English. For each of the tasks that are tackled by our pipeline, we introduced corpora in both languages, which were annotated as part of this thesis. For the automatic detection of T&C, we used a corpus with almost 16,000 pages from online shops, of which one-third were T&C pages. We found a rule-based analysis of URLs to be most effective in predicting whether a page is a T&C page or not. For the automated extraction and segmentation of T&C, we compared three existing libraries, of which we found Trafilatura to work best. To hierarchically order the extracted T&C, we developed a rule-based approach that uses the structure, enumeration, and appearance of headings to determine the position of a clause in the hierarchy of the contract.

For the classification of topics and subtopics of clauses, we compared a rule-based approach with logistic regression, decision trees, different neural networks, and a transformer model. We found the BERT classifier to be most effective with an F1-score of 0.91 in German and 0.85 in English for clause topics, and 0.86 and 0.73 for subtopics. We found that the computationally much cheaper logistic regression classifier, which used Tf-idf vectors as input, performed second best, outperforming different types of neural networks, which used word embeddings as input.

For the final step of our NLP-pipeline, the extraction of information from clauses, we compared the ML library MITIE with a rule-based approach that analyzes dependency tree representations of the clauses. We found that the rule-based approach performed significantly better with an F1-score of 0.88 in English and 0.81 in German, compared to 0.66 and 0.60, achieved by MITIE.

In Chapter 7, we defined relevant legal provisions for the automated legal assessment of clauses and introduced a corpus with more than 1,000 clauses that have been legally assessed by consumer advocates. We also defined the data format for a knowledge-base, which can not only hold the information necessary to legally assess clauses but also the rules necessary for the topic classification and information extraction. Finally, we compared a rule-based legal assessment, based on the information extracted in the previous step, with a binary classification approach using BERT. The rule-based legal assessment achieved an F1-score of 0.75. BERT significantly outperformed it with an F1-score of 0.89. Additionally, while the BERT approach covered all types of clauses, the rule-based approach could only be applied to about 15% of all clauses in the corpus.

The advantage of the rule-based approach, however, is that the results are based on easily interpretable rules, which we could use for the summarization of standard form contracts, described in Chapter 8. We presented an approach for the summarization that uses the extracted information and the rule-based analysis to generate two different kinds of summarizations, one targetted at consumers, containing only a short and simplified summary, and one targeted at consumer advocates, also containing the original text with additional explanatory elements. Both summaries can be generated in German and English. In German, we use a surface realizer that was implemented as part of this thesis. We evaluated both types of summaries with the respective target audience and found that consumers felt, on average, well informed after reading the summaries. Consumer advocates, on the other hand, felt that the summaries simplify too much and leave out information that is crucial to them. However, they did like the explanatory elements.

Finally, in Chapter 9, we presented the prototypical implementation of our results as part of a web application that is designed to support consumer advocates in the analysis and legal assessment of standard form contracts. The prototype uses a micro-service architecture, in which the individual services are orchestrated by the web application, which was implemented using Angular 9. An evaluation of the tool, which we presented in Chapter 10, showed that the tool did successfully support the experts in finding individual clauses faster and, therefore, also in assessing them faster. For the assessment of complete contracts, however, we did not find that the tool reduced the necessary time. All participants reported that they did like the tool and believe that it could help to improve their daily work. Participants particularly liked the navigation features, which allow them to jump directly to certain clauses or clause topics. These features are enabled by the automatic clause topic classification and the automated hierarchical segmentation of the contracts. The automated legal assessment was perceived as less helpful because participants still spent the same time analyzing the clauses to double-check whether the automatic assessment was correct.

To conclude this summary, we revisit the research questions that guided our work and recapitulate on the answers we found:

> ***Research Question RQ1:*** What are existing technical approaches to the semantic analysis, legal assessment, and summarization of legal documents?

In our literature research, we found that a wide variety of NLP technologies is applied to the semantic analysis, legal assessment, and summarization of legal documents, from rule-based approaches to transformer language models. In our analysis of commercial tools, we found that simpler technologies, as simple as keyword searches, dominate, at least partially, because they do not need large corpora to be trained on.

> ***Research Question RQ2:*** Which clauses in standard form contracts are most relevant from a consumer and consumer advocate perspective?

We conducted a survey among 100 consumers and interviewed seven consumer advocates to find out which clauses are most relevant. We found that for consumers, the most relevant clauses are about their right to withdraw from contracts, data protection and privacy, payment, and warranty. We found that the view of the consumer advocates is mostly defined by the cases consumers bring to them. Therefore, for them, all clauses related to monetary obligations, like fees for certain payment methods or default charges, are most relevant, but also withdrawal clauses, especially with respect to long-term contractual obligations, like mobile phone contracts or insurances. At least partially, this can be explained with the fact that the counseling offered by the consumer advocates is not free of charge and consumers are therefore more likely to bring forward cases that are connected to monetary obligations. The different perceptions of the relevance of data protection and privacy clauses might be caused by the fact that in Germany, each state provides a designated commissioner for data protection, with which consumers can file complaints free of charge.

> ***Research Question RQ3:*** How do consumer advocates work with standard form contracts?

We have analyzed the processes related to the handling of standard form contracts in two German consumer protection agencies and formalized them. We found that they mostly work reactive, i.e., they get active once a consumer contacts them with a specific complaint or question. They then analyze the contract of the consumer with regard to their inquiry. They usually do not have the time to additionally check other aspects of the contract, which are unrelated to the inquiry.

In addition to providing counseling to the consumer, the analysis can also lead to a warning or a cease-and-desist order that is sent to the company which drafted the contract. In case of sending a cease-and-desist order, all German consumer protection agencies follow a common workflow in which they first notify each other in order to avoid overlapping actions. One of the problems we identified is that the consumer advocates lack the resources to effectively monitor adherence to signed cease-and-desist orders.

Less frequently, consumer advocates conduct concentrated campaigns in which they proactively check specific industries or types of clauses by acquiring larger amounts of contracts and checking them systematically. The aim of such campaigns is usually not to notify or send cease-and-desist orders to individual companies but to create public awareness for a certain problem. While all the consumer advocates we interviewed agreed that such campaigns are a useful tool, they noted that, under normal circumstances, they do not have the resources to conduct such campaigns.

> ***Research Question RQ4:*** How should software be designed to support consumer advocates in their efforts to protect consumers from void clauses in standard form contracts?

Based on the above-described process analysis, we identified three tasks at which a software would be most helpful for consumer advocates. Either by helping them to analyze consumer contracts faster, or by helping them to analyze a large number of contracts for a specific clause, or by helping them to monitor the adherence to cease-and-desist orders. In addition to an easy to use UI, it is also important for the legal experts that the UI provides them access to the full texts at all times so that they can verify any automatic assessment that is made by the system.

> ***Research Question RQ5:*** Which methods can be used to semantically analyze standard form contracts in German and English?

In extensive evaluations, we have found that, depending on the nature of the analysis, different methods provide the best results. For the automatic detection of T&C in online shops, we found that a rule-based analysis of the page URL is most effective (F1 = 0.95 for German and F1 = 0.89 for English). For the content extraction and clause segmentation, we found that an analysis of shallow text features (i.e., a structural rather than a semantic analysis) provides the best results. For the clause topic classification, we found transformer models (specifically BERT) to perform best (F1 = 0.91 for German and 0.85 for English). However, we also found that the computationally much cheaper and mathematically simpler logistic regression classifier, when applied to Tf-idf input vectors, performs remarkably well with an F1-score of 0.87 for German and 0.80 for English.

For the extraction of structured information from clauses, we found hand-crafted rules, which are applied to the dependency tree representation of the input, to perform best with an F1-score of 0.81 for German and 0.88 for English. Finally, for the legal assessment of clauses, we found transformer models to perform best with an F1-score of 0.89 for German. For English, we did not have enough data to evaluate transformer models on this task.

> ***Research Question RQ6:*** How can the body of law governing the drafting of standard form contracts in Germany and the European Union be formalized and represented in order to enable the automated legal assessment of standard form contracts?

Most of the proposed and existing formats for modeling knowledge in the legal domain are designed for the introspective analysis of laws and regulations, rather than trying to make them accessible for practical application in software. Therefore, we decided to implement a light-weight JSON format, which is focused purely on the application, at the cost of simplifying legal regulations to what seemed permissible and practical to our experts for the analyses we wanted to conduct.

> ***Research Question RQ7:*** Which methods can be used to automatically summarize German and English standard form contracts?

Neural methods to text generation, and more specifically summarization, were ruled out from the beginning because of the danger of "hallucination", i.e., facts that appear in the summary but are not backed by the original text. Because the overwhelming majority of clauses in T&C is in accordance with legal regulations, the possibility would be high that the most important parts, i.e., void clauses, would be summarized wrongly. We, therefore, used a rule-based approach, which takes the results of the information extraction and the rule-based legal assessment as input to generate abstractive summarizations.

For consumers, we generated summarizations that mainly focus on simplifying and condensing the information which is most relevant to them. For consumer advocates, the summarization focuses on summarizing void clauses and additionally provides an explanatory component on how the assessment was made. We evaluated both types of summarization with the respective target audiences and found that consumers do like the generated summaries and feel well informed after reading them. Consumer advocates, however, did not like the summaries, mainly because they thought they simplify matters too much and leave out crucial information. They did, however, like the explanatory component that was provided.

> ***Research Question RQ8:*** Can AI-based automation and support tools improve the way consumers and consumer advocates work with standard form contracts?

Based on RQ 7, we can conclude that automatic text summarization can improve the way consumers work with standard form contracts. They do feel informed after reading them, and they self-reported to be more likely to read summaries than actual T&C. For consumer advocates, this is not the case because they believe the summaries leave out important information and would therefore not rely on them. However, with the prototype we implemented, consumer advocates were able to reduce the time they need to find and assess individual clauses in T&C. Therefore, with regard to RQ8, but also with regard to the overall hypothesis of this thesis, we can conclude that AI technology, from simple rules to transformers, can indeed help consumers and consumer advocates in the different steps that are necessary to read and understand standard form contracts, if the right technology is applied for the right target audience.

11.2. Critical Reflection

Throughout this thesis, we highlighted limitations directly, wherever they appeared. Here, we want to briefly summarize them and add some critical reflections of the work on a higher level.

11.2.1. Limitations of the Consumer Survey

The results of the consumer survey (Section 4.2) are purely based on a self-assessment by the participants. It is therefore not unlikely that some of the responses, e.g., about how often participants read T&C, are overstated, whether purposefully or not. However, if participants are even less likely to read T&C, it makes our case only stronger. In other instances, e.g., when we asked about whether participants could imagine using a tool that summarizes relevant aspects of T&C for them, we believe that, in reality, far fewer people would actually use such a tool.

11.2.2. Limitations of the Requirement Analysis

The analysis of the consumer requirements (Section 5.3.2) is based on interviews with a very heterogeneous group, which only consisted of students. Therefore, it is likely that the results are not representative for consumers at large. Since we already knew, before conducting the interviews, that we would focus on consumer advocates in our tool development, we focused our resources on gathering their requirements and accepted this limitation.

11.2.3. Limitations of the NLP-Pipeline

The weakest point in the NLP-pipeline is the automatic extraction and segmentation of T&C. Even the best performing state-of-the-art approach only extracted 73% of German documents and 82% of English documents correctly. This number does not yet include errors in the hierarchy of the extraction. This is especially problematic since it is one of the earliest steps in the pipeline. If a clause is not even extracted, it can never be assessed correctly. Even if “just” the hierarchy is not correct, we might not have the corresponding heading for a clause, which, as we have shown, leads to a lower performance for the topic classification. In order to significantly increase the extraction performance, we believe it will be necessary to develop techniques which are tailored to standard form contracts, which usually provide much more and smaller structures (i.e., more subheadings and shorter paragraphs), than, e.g., news articles, for which a lot of the state-of-the-art tools are optimized.

When we started this thesis, we hoped that, especially due to new multilingual ML models, we would be able to leverage benefits from working with data in two languages. However, it became clear quickly that, even for high-level tasks like the automatic detection of T&C, multilingual training data decreases, rather than increases, performance for most tasks.

Despite being governed by the same EU regulations in many cases, at least at the time we collected our data, T&C under UK jurisdiction still structurally look very different to T&C under German jurisdiction. On average, they have fewer clauses and shorter sentences. And even when their content refers to regulations that originate from the same EU legislation, e.g., withdrawal rights, due to the fact the national laws implementing these regulations are named differently and might differ in details, it is difficult to benefit from multilingual data sets. These aspects come on top of the fact that we are dealing with different languages and their own specificities, like the extensive usage of compound nouns in German contracts.

Due to the amount of time required to build corpora, and due to the fact that we only collaborated with German consumer advocates, our English corpora are relatively small. Our initial hope was that we could utilize this fact to show that multilingual data sets can help to improve performance in cases with low resources in one language. However, since that turned out not to be the case, for most of our evaluations, we could only conclude that most approaches performed well on English data, despite the data scarcity. This might be an indication that, given an equally sized data set, the results on English texts might outperform the results in German. However, since we do not have the necessary data to test that, this remains an assumption.

11.2.4. Limitations of the Generated Summaries

In our evaluation, consumers were happy with the generated summaries and felt well informed after reading them. However, after reading summaries that contain the same information, consumer advocates felt that the summaries were simplifying matters too much.

Either consumers did not care about these simplifications, or they were not even aware of the fact that the summary was missing out on some information. There are two ways of looking at these results. One can argue that the consumers would not lose any information that is relevant to them because either they also did not notice it in the original version or they did not care. And even if some information is missing, given that most consumers usually do not read T&C at all, one could still argue that having even shallow information might still be better than having no information. However, one could also argue that the summaries potentially provide wrong information to the consumer, which might give them a false sense of security.

Consumer advocates did not find the automated summaries to be accurate. The part they did like, the explanatory elements, is based on the rule-based analysis of clauses, because only based on the rule-based analysis we can say which words were relevant for the clause topic classification and which information was used for the legal assessment. However, in both cases, the legal assessment and the topic classification, the rule-based approaches were significantly outperformed by ML approaches. Therefore, we have to conclude that the generated summaries, even with the explanatory component, have little to no utility for the experts.

11.2.5. Limitations of the Developed Prototype

The prototype was developed to facilitate the evaluation of the developed NLP-pipeline in a realistic and task-based setting. While we believe that the prototype was very well suited for that, and the feedback from the participants of the evaluation was very positive, as a tool in itself, the prototype has many limitations.

Table 11.1 presents an overview of the implementation status of the system requirements presented in Section 5.4 in the prototype. Five system requirements were fully implemented (marked with an ✓), five were partially implemented (O), and three were not implemented (X). All system requirements with a high research priority were at least partially implemented.

ID	**requirement**	**user prio**	**research prio**	**status**
SysRq 1	The system should be able to handle input from different sources, like websites, PDF documents and printed documents.	high	low	X
SysRq 2	The system should be able to automatically separate a contract into its clauses.	high	high	✓
SysRq 3	The system should be able to automatically detect the topic of individual clauses.	high	high	✓
SysRq 4	The system should be able to automatically assess whether a clause is potentially void or not.	high	high	✓
SysRq 5	Users should be able to define their own evaluation criteria within the system and train new ML models.	high	medium	O
SysRq 6	The system should be able to justify its decisions by providing the text passage which led to a certain decision.	high	medium	O

Table 11.1.: Implementation status of system requirements

ID	requirement	user prio	research prio	status
SysRq 7	The system should be able to automatically create reports which contain the text of the clause, its topic, its source, and why it was classified void.	medium	medium	O
SysRq 8	The user should be able to define a list of websites that should be checked periodically and for which specific clause they should be checked.	high	medium	X
SysRq 9	The system should be able to notify a user, e.g., via email, in case a periodical check finds a void clause.	low	low	X
SysRq 10	The system should be able to automatically identify documents and websites which represent T&C.	high	high	✓
SysRq 11	The system should be able to automatically identify T&C that contain a certain (void) clause.	high	high	O
SysRq 12	The system should be able to assess T&C as a whole, as to whether or not they contain any potentially void clauses or undesired clauses.	high	medium	✓
SysRq 13	The system should be able to shortly summarize the content of the T&C in natural language.	medium	high	O

Table 11.1.: Implementation status of system requirements

The system can currently not handle input from sources other than online shops (SysRq 1) because, in the area of our focus, T&C from online shops, other document sources are mostly irrelevant. While users can edit the rule-based evaluation criteria in a rudimentary editor on the prototype, they can not directly train ML models from within the system (SysRq 5).

As we have shown in Chapter 8, the system can, in general, provide justification for its decisions through text highlighting (SysRq 6). However, as we have discussed in the previous section, it can only do so for decisions that were made based on rules. Since rule-based approaches were outperformed significantly by ML in the relevant tasks (i.e., topic classification and legal assessment), in practice, the system will not be able to explain most of its decisions. Therefore, we decided not to include a visual representation of the explanatory component in the prototype.

Similarly, while we are in general able to generate summaries for experts about individual clauses (SysRq 13) and whole contracts (SysRq 7), as presented in Chapter 8, we have shown that experts do not like the generated summaries and therefore decided not to add them to the UI of the prototype.

Since the periodical checking of websites (SysRq 8, SysRq 9) was not part of our evaluation and can only be evaluated in a long-term study (see Section 11.4.4 on future plans for a long-term study), it was not implemented in the prototype. The system does have the necessary assessment capacities. However, a corresponding UI would have to be implemented and a scheduler, which automatically performs assessments at given points in time.

Lastly, while the system is able to identify void clauses of a certain topic (SysRq 11), there is currently no UI functionality implemented to perform such an analysis simultaneously on a larger corpus.

11.2.6. Limitations of the Tool Evaluation

Although the participants used the tool for the first time during the evaluation, they were closely involved in the development of the tool and also the design of the UI. "Average" users, less familiar with the system, might have benefited less from the support features because they might have needed more time to adjust to the system.

We also have to assume that, in the setting of the experiment, participants might be more cautious than in everyday situations. Therefore, the fact that only one participant missed the void clause in Scenario 3, which was falsely not annotated by the system, might be caused by this extra caution. A long-term study would be necessary to closer investigate such effects (see Section 11.4.4).

11.2.7. Generalisation to Non-T&C Contracts

We have shown that the technologies we applied can also be used on other types of standard form contracts, e.g., from the domain of banking. However, the concrete models and most notably the taxonomy, on which most of our rule-based approaches depend, cannot be directly applied to all other types of contracts. While some clauses, like severability and withdrawal clauses, for example, are more universal, there are also clauses that are specific to a certain kind or at least a group of standard form contracts. It would, for example, still require a considerable amount of work before our results could be applied to rental agreements.

11.2.8. Influence on Consumer Protection

Although we have shown that different AI technologies can indeed be used to support consumers and the work of consumer advocates and hence further consumer protection, there is a limit to what technology can achieve. One could even argue that, in a situation where neither consumers nor consumer advocates can keep up with the number of legal texts we face on a daily basis, instead of supporting the status quo by mitigating the underlying problems with technology, one should focus on the root causes. Realistically speaking, it is unlikely that the underlying problem, the increasing complexity and length of standard form contracts, will be solved on a legislative level any time soon. Therefore, we do think pragmatic solutions, like the one we presented in this thesis, are useful. Nevertheless, we do not want to miss the opportunity to suggest other actions, which, based on our research, we believe to be helpful in tackling the underlying problems.

With our consumer survey (see Section 4.2), we have confirmed what we know from both anecdotal and scientific evidence: consumers largely do not read T&C they agree to, mostly because they are simply too long. What we did not know before is that consumers are also very hesitant to seek legal help, even if they lose money in a fraudulent transaction. We also found that a, even for our experts, surprisingly large amount of T&C contains void clauses. While the area is already comparably strictly regulated, with many consumer-friendly regulations, there is a gap when it comes to the enforcement of these regulations. If void clauses are challenged, they are mainly challenged by competitors, which, most likely, do not represent consumer interests.

Our findings suggest that consumer protection measures that are based on adding additional information obligations to T&C are not only not helpful, because the new information will not be read by the majority of the consumers anyway, but also might actively hurt consumer protection, by increasing the length of T&C and therefore making it even more unlikely that consumers will read them. A relatively recent example of such an information obligation became effective in 2016 based on Regulation No 524/2013 of the European Parliament and Council of the European Union (2013), according to which commercial websites have to inform consumers about the ODR platform of the EU. According to the survey by Föhlisch and Groß (2018), 10% of all cease-and-desists orders that were received by the participants were targeted at missing or wrong information about the ODR platform. According to the consumer advocates we spoke to, these cease-and-desist orders most likely almost exclusively originated from competitors because, from a consumer protection perspective, they do not deem this information obligation relevant enough to take legal steps. For other information obligations, e.g., the mandatory provision of a modal withdrawal form within the cancellation policy, it is also questionable whether they actually provide an added value for consumers or just increase the length of texts that they would have to read and provide targets for cease-and-desist orders from competitors.

Based on our findings, we believe that effective consumer protection measures should rather try to reduce the size of T&C than increase it in order to increase the likelihood that consumers will read them. More importantly, they should focus on enforcing regulations before consumers have negative experiences, e.g., by providing organizations like the consumer protection agencies with the means to actively monitor marketplaces and ensure that T&C adhere to regulations. Our research has shown that AI can be an effective mean to support consumer advocates with this task.

11.3. Lessons Learned from the Interdisciplinary Collaboration

Interdisciplinary research involving practitioners and researchers from computer science and law is becoming more popular in recent years, not least thanks to a growing number of programs from different funding bodies. We found many of the well-known “pearls of wisdom”, e.g., about the importance of finding a common language, to be true, but we also gained some less expected insights.

We found none of the popular prejudices about the technology aversion of legal experts to be true, on the contrary, all involved parties, independent from their tech-savviness, were very open and interested. Admittedly, there was a selection-bias involved, because only experts with a certain interest in technology would join such a project. Additionally, we communicated very clearly from the beginning that we want to build a *support* tool, which supports expert users, but also leaves them with the final decisions.

Experts were very much aware of the fact that any AI system, especially if it is data-driven, will most likely never achieve “perfect” results and they did not found that to be problematic. However, when it came to their own annotations, the experts were less relaxed. When we asked them to annotate clauses as “void”, there were many concerns that the final decision whether a clause is void can only be made by a court. Therefore, we agreed on labeling clauses as “potentially void” instead of “void”. Even after we explained to the experts that they should label each clause as potentially void, that they would like the system to present to them for a manual assessment, some were still very reluctant to annotate a clause as “potentially void”.

From a technical perspective, we wanted a binary labeling because we treated the problem as a binary classification problem (void / valid or present to expert / do not present). In reflection, we underestimated the understandable aversion to make annotations the experts themselves were not sure about, caused by their professional standards. It would have been the appropriate solution to offer a third annotation label and convert it in the training data to (potentially) void.

11.4. Outlook

As it is usually the case with larger research projects, during the course of this thesis, at least as many new questions opened up, as questions have been answered. In conclusion of this thesis, we want to highlight the questions that arose, which we believe to be most interesting and relevant. While some of the questions are yet to be addressed, others are currently already under investigation, as part of the “AGB-Check” project, which is funded by the German Federal Ministry of Justice and Consumer Protection and builds on the results presented in this thesis.

11.4.1. Continuous Learning

The assumption that often comes with ML systems is that their performance increases over time. However, this is only true if user input is used to re-train the involved models. The prototype in its current form already supports user annotations, which can correct (or confirm) automatic annotations. However, there is currently no automated pipeline in place to re-train the models based on these inputs. While, in theory, this is an easy task since the data is available in a structured format comparable to the original training data and could therefore be used easily to train a new model, in practice, especially for transformer models, this training process is very resource-intensive, and the system which is used to run the model is not necessarily suitable to re-train it. Additionally, if enough new data is fed into the system, it might also become necessary to re-adjust the hyper-parameters that are used for the training of the model.

11.4.2. Collaboration

A related problem, that is one of the focuses of the “AGB-Check” project, arises from the challenges of collaborative work. During the annotation of the corpora, we already found that there was a relatively low inter-annotator agreement between the experts when it came to the assessment of whether a clause is potentially void or not. In order to train an ML model, however, we need a single annotation, which can be used as training data.

Since consumer protection in Germany, in many aspects, is a state matter, and even more so on the European level, we are likely to not just encounter inter-annotator disagreement but also diverging views and rules between states and organizations. How such disagreements can be handled in a shared platform (e.g., by offering a process to consolidate annotations or providing different ML models) is an interesting field of research.

11.4.3. Presentation of Conflicting Annotations

Not just human annotators can produce conflicting assessments, but algorithms too. The prototype already supports displaying annotations from different sources, which are indicated by different labels. However, there is little doubt that this is not the optimal way to present conflicting annotations. The information that a given annotation was produced by a rule-based classifier or a logistic regression classifier will provide little guidance to most users about which label they should rely on. There are different measurements based on which we can make assumptions about which annotation is more likely going to be correct, e.g., based on the results of our evaluation or based on confidence scores that many classifiers can provide. Investigating which of these measurements is most appropriate, but also how to present them most effectively (e.g., by ranking the results, reporting the confidence scores or simply presenting the annotation which is most likely correct) is another interesting field of research that will be investigated as part of the “AGB-Check” project.

11.4.4. Long-Term Effects

The evaluation of the prototype has already shown that there is a danger that false negatives from the system could lead to potentially void clauses being overlooked. However, the possibilities to study such effects in short task-based evaluations is very limited. Due to the artificial setting, some participants might be more critical with the automated assessment than they would be in a more natural setting, others might be less critical because they might feel like they have to be fast.

It is also likely that such behavior patterns change during long-term usage of the system. If the system is perceived to be very often correct by a user, they might, over time, put more and more trust into the system and therefore put less effort into checking the results, which might become problematic. If, on the other hand, the system is perceived to be incorrect very often, users might not trust the judgment of the system at all anymore, which would also mean that they would not benefit from using it anymore. Finding the sweet spot between the two extremes, in which users trust the system but not blindly, is far from trivial and might even need some unconventional interventions, like purposefully presenting wrong annotations to the user in order to raise their awareness (Braun et al., 2020).

However, in order to find out if such problems occur at all, it would be necessary to observe user behavior in the natural setting of daily work. Conducting such a longitudinal study does not only need a lot of time, but it also poses new requirements regarding data and privacy rules, and most importantly, a system that is ready for production use is needed to conduct such an evaluation.

Bibliography

Evelyn Agbaria. *PONS die deutsche Rechtschreibung*, volume 1. PONS GmbH, Stuttgart, 2009.

Fouad Nasser A Al Omran and Christoph Treude. Choosing an nlp library for analyzing software documentation: a systematic literature review and a series of experiments. In *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, pages 187–197. IEEE, 2017.

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. Predicting judicial decisions of the european court of human rights: a natural language processing perspective. *PeerJ Computer Science*, 2:e93, 2016. ISSN 2376-5992. doi: 10.7717/peerj-cs.93. URL `https://doi.org/10.7717/peerj-cs.93`.

Nawal Aljedani, Reem Alotaibi, and Mounira Taileb. Hmatc: Hierarchical multi-label arabic text classification model using machine learning. *Egyptian Informatics Journal*, 2020.

Waleed Ammar, Shomir Wilson, Norman Sadeh, and Noah A Smith. Automatic categorization of privacy policies: A pilot study. *School of Computer Science, Language Technology Institute, Technical Report CMU-LTI-12-019*, 2012.

Amtsgericht Brandenburg an der Havel. 31 c 190/06. `https://openjur.de/u/684362.html`, 2007. Accessed 2020-10-08.

Association of Corporate Counsel. Global Legal Department Benchmarking Report. Technical report, 2019.

David E Avison, Francis Lau, Michael D Myers, and Peter Axel Nielsen. Action research. *Communications of the ACM*, 42(1):94–97, 1999.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Yannis Bakos, Florencia Marotta-Wurgler, and David R Trossen. Does anyone read the fine print? consumer attention to standard-form contracts. *The Journal of Legal Studies*, 43(1): 1–35, 2014.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W05-0909`.

Adrien Barbaresi. Generic web content extraction with open-source software. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 267–268, Erlangen, Germany, 2019. German Society for Computational Linguistics & Language Technology.

Adrien Barbaresi. Evaluating scraping and text extraction tools for python. `http://adrien.barbaresi.eu/blog/evaluating-text-extraction-python.html`, 2020. Accessed 2020-05-21.

Marco Baroni, Johannes Matiasek, and Harald Trost. Predicting the components of german nominal compounds. In *ECAI 2002: 15th European Conference on Artificial Intelligence*, pages 470–474, 2002.

Marco Baroni, Francis Chantree, Adam Kilgarriff, and Serge Sharoff. Cleaneval: a competition for cleaning web pages. In *Lrec*, 2008.

Roberto Bartolini, Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, and Claudia Soria. Automatic classification and analysis of provisions in italian legal texts: A case study. In Robert Meersman, Zahir Tari, and Angelo Corsaro, editors, *On the Move to Meaningful Internet Systems 2004: OTM 2004 Workshops*, pages 593–604, Berlin, Heidelberg, 2004a. Springer Berlin Heidelberg. ISBN 978-3-540-30470-8.

Roberto Bartolini, Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, and Claudia Soria. Semantic mark-up of Italian legal texts through NLP-based techniques. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004b. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2004/pdf/709.pdf`.

Jürgen Basedow. § 305b vorrang der individualabrede. In Franz Jürgen Säcker, Roland Rixecker, Hartmut Oetker, and Bettina Limperg, editors, *Münchener Kommentar zum Bürgerlichen Gesetzbuch*. C. H. Beck Verlag, Munich, 8 edition, 2019.

Michael Beißwenger, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. EmpiriST 2015: A shared task on the automatic linguistic annotation of computer-mediated communication and web corpora. In *Proceedings of the 10th Web as Corpus Workshop*, pages 44–56, Berlin, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2606. URL `https://www.aclweb.org/anthology/W16-2606`.

Valentina Bellomaria, Giuseppe Castellucci, Andrea Favalli, and Raniero Romagnoli. Almawave-slu: A new dataset for SLU in italian. In Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019. URL `http://ceur-ws.org/Vol-2481/paper5.pdf`.

Anja Belz and Ehud Reiter. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, April 2006. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E06-1040.

BGBl. Konsumentenschutzgesetz (kschg). Bundesgesetzblatt für die Republik Österreich, Teil 1, Nummer 51, 2018.

C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni, and C. Soria. Automatic semantics extraction in law documents. In *Proceedings of the 10th International Conference on Artificial Intelligence and Law*, ICAIL '05, page 133–140, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595930817. doi: 10.1145/1165485.1165506. URL https://doi.org/10.1145/1165485.1165506.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstrations*, pages 33–36, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL https://www.aclweb.org/anthology/C10-3009.

Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*, 2017.

Alexander Boer, Rinke Hoekstra, Radboud Winkels, T Van Engers, and F Willaert. Metalex: Legislation in xml. *Legal Knowledge and Information Systems (Jurix 2002)*, pages 1–10, 2002.

Marcel Bollmann. Adapting simplenlg to german. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 133–138, 2011.

Marcel Bollmann. Simplenlg for german. https://marcel.bollmann.me/software/simplenlg.html, 2019. Last accessed 2019-09-10.

Igor A Bolshakov. Crosslexica, the universe of links between russian words. *Busyness Informatica*, (3), 2013.

Igor A. Bolshakov and Alexander Gelbukh. Text segmentation into paragraphs based on local text cohesion. In Václav Matoušek, Pavel Mautner, Roman Mouček, and Karel Taušer, editors, *Text, Speech and Dialogue*, pages 158–166. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001. ISBN 978-3-540-44805-1.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. Tiger: Linguistic interpretation of a german corpus. *Research on language and computation*, 2(4):597–620, 2004.

Daniel Braun and Florian Matthes. Generating explanations for algorithmic decisions of usage-based insurances using natural language generation. In Matthias Tichy, Eric Bodden, Marco Kuhrmann, Stefan Wagner, and Jan-Philipp Steghöfer, editors, *Software Engineering und Software Management 2018*, pages 219–220, Bonn, 2018. Gesellschaft für Informatik.

Daniel Braun and Florian Matthes. Automatic detection of terms and conditions in german and english online shops. In *Proceedings of the 16th International Conference on Web Information Systems and Technologies - Volume 1: WEBIST,*, pages 233–237. INSTICC, SciTePress, 2020. ISBN 978-989-758-478-7. doi: 10.5220/0010154302330237.

Daniel Braun, Adrian Hernandez-Mendez, Florian Matthes, and Manfred Langen. Evaluating natural language understanding services for conversational question answering systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 174–185, Saarbrücken, Germany, 2017a. Association for Computational Linguistics. doi: 10.18653/v1/W17-5522. URL `http://www.aclweb.org/anthology/W17-3622`.

Daniel Braun, Elena Scepankova, Patrick Holl, and Florian Matthes. Satos: Assessing and summarising terms of services from german webshops. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 223–227, Santiago de Compostela, Spain, 2017b. Association for Computational Linguistics. doi: 10.18653/v1/W17-3534. URL `http://aclweb.org/anthology/W17-3528`.

Daniel Braun, Anne Faber, Adrian Hernandez-Mendez, and Florian Matthes. Automatic relation extraction for building smart city ecosystems using dependency parsing. In *Proceedings of the 2nd Workshop on Natural Language for Artificial Intelligence (NL4AI 2018)*, 2018a. URL `http://ceur-ws.org/Vol-2244/paper_03.pdf`.

Daniel Braun, Ehud Reiter, and Advaith Siddharthan. Saferdrive: An nlg-based behaviour change support system for drivers. *Natural Language Engineering*, 24(4):551–588, 2018b. doi: 10.1017/S1351324918000050.

Daniel Braun, Elena Scepankova, Patrick Holl, and Florian Matthes. Customer-centered legaltech: Automated analysis of standard form contracts. In *Tagungsband Internationales Rechtsinformatik Symposium (IRIS) 2018*, pages 627–634. Editions Weblaw, 2018c. URL `http://jusletter-it.weblaw.ch/issues/2018/IRIS/customer-centered-le_86c0d312a1.html`.

Daniel Braun, Kira Klimt, Daniela Schneider, and Florian Matthes. Simplenlg-de: Adapting simplenlg 4 to german. In *Proceedings of the 12th International Conference on Natural Language Generation*, Tokio, Japan, 2019a. Association for Computational Linguistics.

Daniel Braun, Kira Klimt, Daniela Schneider, and Florian Matthes. Simplenlg-de: Adapting simplenlg 4 to german. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 415–420, Tokyo, Japan, 2019b. URL `https://www.aclweb.org/anthology/W19-8651`.

Daniel Braun, Elena Scepankova, Patrick Holl, and Florian Matthes. The potential of customer-centered legaltech. *Datenschutz und Datensicherheit - DuD*, 43(12):760–766, 2019c. ISSN 1862-2607. doi: 10.1007/s11623-019-1202-7. URL `https://doi.org/10.1007/s11623-019-1202-7`.

Daniel Braun, Elena Scepankova, Patrick Holl, and Florian Matthes. Consumer protection in the digital era: The potential of customer-centered legaltech. In Klaus David, Kurt Geihs, Martin Lange, and Gerd Stumme, editors, *INFORMATIK 2019: 50 Jahre*

Gesellschaft für Informatik – Informatik für Gesellschaft, pages 407–420, Bonn, 2019d. Gesellschaft für Informatik e.V. ISBN 978-3-88579-688-6. doi: 10.18420/inf2019_58. URL `https://doi.org/10.18420/inf2019_58`.

Daniel Braun, Manoj Bhat, Andreas Biesdorf, and Florian Matthes. Would you lie to me bot? supporting decision-making processes with deceiving virtual agents. *Procedia Computer Science*, 177:587 – 592, 2020. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2020.10.083. URL `http://www.sciencedirect.com/science/article/pii/S1877050920323541`. The 11th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2020) / The 10th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2020) / Affiliated Workshops.

Joost Breuker. Modelling artificial legal reasoning. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 66–78. Springer, 1993.

Bundesgerichtshof. Viii zr 95/18. `http://juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=en&nr=97735&pos=0&anz=1`, 2019. Accessed 2020-10-08.

Bundeskriminalamt. Links zu den Onlinewachen bzw. zu den Kontaktdaten der Landespolizeien. `https://www.bka.de/DE/KontaktAufnehmen/Onlinewachen/onlinewachen_node.html`, 2020. Last accessed 2020-08-16.

Bundesministerium für Arbeit und Soziales. Verordnung zu den dokumentationspflichten nach den §§ 16 und 17 des mindestlohngesetzes in bezug auf bestimmte arbeitnehmergruppen (mindestlohndokumentationspflichten-verordnung – milodokv). *Bundesanzeiger*, 2014. URL `https://www.bundesanzeiger.de/pub/publication/cRO3xjBGOFyhUKBDHDO`.

c. 43. Consumer protection act 1987. Bundesgesetzblatt für die Republik Österreich, Teil 1, Nummer 51, 1987.

Giuseppe Carenini and Jackie Chi Kit Cheung. Extractive vs. nlg-based abstractive summarization of evaluative text: The effect of corpus controversiality. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 33–41. Association for Computational Linguistics, 2008.

Andrea Cascallar-Fuentes, Alejandro Ramos-Soto, and Alberto Bugarın. Adapting simplenlg to galician language. *INLG 2018*, page 67, 2018.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. Extracting contract elements. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, page 19–28, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348911. doi: 10.1145/3086512.3086515. URL `https://doi.org/10.1145/3086512.3086515`.

Branden Chan, Stefan Schweter, and Timo Möller. German's next language model. *arXiv preprint arXiv:2010.10906*, 2020.

Udit Chandna and Manjunath Ramachandra Iyer. System for semi-automated chatbots query classification training corpus generation. In *International Conference on Advanced Computing and Communications (ADCOM 2018)*, Bengaluru, India, 2018. Advanced Computing and Communications Society (ACCS).

Danqi Chen and Christopher Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1082. URL https://www.aclweb.org/anthology/D14-1082.

Guanyi Chen, Kees van Deemter, and Chenghua Lin. Simplenlg-zh: a linguistic realisation engine for mandarin. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 57–66, 2018.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, and Hui Jiang. Distraction-based neural networks for document summarization. *arXiv preprint arXiv:1610.08462*, 2016.

Jianpeng Cheng and Mirella Lapata. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1046. URL https://www.aclweb.org/anthology/P16-1046.

Emmanuel Chieze, Atefeh Farzindar, and Guy Lapalme. Automatic summarization and information extraction from canadian immigration decisions. In *Proceedings of the Semantic Processing of Legal Texts Workshop*, pages 51–57. LREC 2008, LREC 2008, may 2008.

Michele Chinosi and Alberto Trombetta. Bpmn: An introduction to the standard. *Computer Standards & Interfaces*, 34(1):124 – 134, 2012. ISSN 0920-5489. doi: https://doi.org/10.1016/j.csi.2011.06.002. URL http://www.sciencedirect.com/science/article/pii/S0920548911000766.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://www.aclweb.org/anthology/D14-1179.

Noam Chomsky. *Syntactic structures.* Moutin Publishers, 14 edition, 1985.

Sumit Chopra, Michael Auli, and Alexander M. Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1012. URL https://www.aclweb.org/anthology/N16-1012.

Arman Cohan and Nazli Goharian. Revisiting summarization evaluation for scientific articles. *arXiv preprint arXiv:1604.00400*, 2016.

Giuseppe Contissa, Koen Docter, Francesca Lagioia, Marco Lippi, Hans-W Micklitz, Przemysław Pałka, Giovanni Sartor, and Paolo Torroni. Claudette meets gdpr: Automating the evaluation of privacy policies using artificial intelligence. Technical report, European Consumer Organisation (BEUC), 2018a.

Giuseppe Contissa, Francesca Lagioia, Marco Lippi, Hans-Wolfgang Micklitz, Przemyslaw Palka, Giovanni Sartor, and Paolo Torroni. Towards consumer-empowering artificial intelligence. In *International Joint Conference on Artificial Intelligence*, pages 5150–5157, 2018b.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces, 2018.

Council of the European Union. Council directive 93/13/eec of 5 april 1993 on unfair terms in consumer contracts. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31993L0013`, 1993. Accessed 2020-03-30.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Joachim Daiber, Lautaro Quiroz, Roger Wechsler, and Stella Frank. Splitting compounds by semantic analogy. In *Proceedings of the 1st Deep Machine Translation Workshop*, pages 20–28, Praha, Czechia, 2015. ÚFAL MFF UK. URL `https://www.aclweb.org/anthology/W15-5703`.

C. De Boom, S. Van Canneyt, S. Bohez, T. Demeester, and B. Dhoedt. Learning semantic similarity for very short texts. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1229–1234, 2015.

Ruud de Jong and Mariët Theune. Going dutch: Creating simplenlg-nl. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 73–78, 2018.

Emile de Maat and Radboud Winkels. Categorisation of norms. In *Proceedings of the 2007 Conference on Legal Knowledge and Information Systems: JURIX 2007: The Twentieth Annual Conference*, page 79–88, NLD, 2007. IOS Press. ISBN 9781586038106.

Emile de Maat and Radboud Winkels. *Automated Classification of Norms in Sources of Law*, pages 170–191. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-12837-0. doi: 10.1007/978-3-642-12837-0_10. URL `https://doi.org/10.1007/978-3-642-12837-0_10`.

Marie-Catherine de Marneffe and Christopher D. Manning. Stanford typed dependencies manual. Technical report, Stanford University, 2008.

Der Spiegel. Verbraucherzentrale bremen meldet insolvenz an. `https://www.spiegel.de/wirtschaft/service/verbraucherzentrale-bremen-meldet-insolvenz-an-a-1254380.html`, 2019. Accessed 2020-05-07.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://www.aclweb.org/anthology/N19-1423`.

Jeremy Dick, Elizabeth Hull, and Ken Jackson. *Introduction*, pages 1–32. Springer International Publishing, Cham, 2017a. ISBN 978-3-319-61073-3. doi: 10.1007/978-3-319-61073-3_1. URL `https://doi.org/10.1007/978-3-319-61073-3_1`.

Jeremy Dick, Elizabeth Hull, and Ken Jackson. *Requirements Engineering in the Problem Domain*, pages 113–134. Springer International Publishing, Cham, 2017b. ISBN 978-3-319-61073-3. doi: 10.1007/978-3-319-61073-3_5. URL `https://doi.org/10.1007/978-3-319-61073-3_5`.

Jeremy Dick, Elizabeth Hull, and Ken Jackson. *Requirements Engineering in the Solution Domain*, pages 135–158. Springer International Publishing, Cham, 2017c. ISBN 978-3-319-61073-3. doi: 10.1007/978-3-319-61073-3_6. URL `https://doi.org/10.1007/978-3-319-61073-3_6`.

Janelle M Diller. Private standardization in public international lawmaking'(2012). *Michigan Journal of International Law*, 33:481, 2012.

Paulo Sergio Medeiros dos Santos and Guilherme Horta Travassos. Action research use in software engineering: An initial survey. In *2009 3rd International Symposium on Empirical Software Engineering and Measurement*, pages 414–417. IEEE, 2009.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech & Language*, 59:123 – 156, 2020. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2019.06.009. URL `http://www.sciencedirect.com/science/article/pii/S0885230819300919`.

Peter Eisenberg, Jörg Peters, Peter Gallmann, Cathrine Fabricius-Hansen, Damaris Nübling, Irmhild Barz, Thomas A Fritz, Reinhard Fiehler, and Mathilde Henning. *Duden - Die Grammatik. Unentbehrlich für richtiges Deutsch.* Dudenverlag, Mannheim, 2016.

Hilke Elsen. Komplexe komposita und verwandtes. *Germanistische Mitteilungen: Zeitschrift für Deutsche Sprache, Literatur und Kultur*, (69):57–71, 2009.

European Court of Justice. Judgment of the court case c-329/02 p, 2014. `http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1416170084502&uri=CELEX:32014R0269`.

European Parliament and Council of the European Union. Regulation (eu) no 524/2013 of the european parliament and of the council of 21 may 2013 on online dispute resolution for consumer disputes and amending regulation (ec) no 2006/2004 and directive 2009/22/ec (regulation on consumer odr). `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32013R0524`, 2013. Accessed 2020-07-10.

Ekkehard Felder and Friedemann Vogel. *Handbuch Sprache im Recht*, volume 12. Walter de Gruyter GmbH & Co KG, 2017.

Michael Fingerhut. *7. Teil: Allgemeine Geschäftsbedingungen*. Carl Heymanns Verlag, 12th edition edition, 2009a. ISBN 978-3-452-28858-5.

Michael Fingerhut. *Vertrags- und Formularbuch*. Carl Heymanns Verlag, 12th edition edition, 2009b. ISBN 978-3-452-28858-5.

Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707, Dec 2018. ISSN 1572-8641. doi: 10.1007/s11023-018-9482-5. URL `https://doi.org/10.1007/s11023-018-9482-5`.

Carsten Föhlisch and Christian Groß. *Online-Handel: Wegweiser durch die rechtlichen Rahmenbedingungen des E-Commerce unter Berücksichtigung des neuen Verbraucherrechts*. DIHK Verlag, Meckenheim, 2 edition, 2018.

Carsten Föhlisch. Abmahnumfrage 2019. Technical report, Trusted Shops, 2019.

Blaise Gadanecz. The syndicated loan market: structure, development and implications. *BIS Quarterly Review*, pages 75 – 89, 2004.

Albert Gatt and Ehud Reiter. SimpleNLG: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 90–93, Athens, Greece, March 2009. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W09-0613`.

Sean M. Gerrish and David M. Blei. Predicting legislative roll calls from text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 489–496, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Shantanu Godbole and Sunita Sarawagi. Discriminative methods for multi-labeled classification. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 22–30. Springer, 2004.

David Goldblatt and Tyler O'Neil. How a bill becomes a law-predicting votes from legislation text, 2012.

Thomas F. Gordon. An overview of the legal knowledge interchange format. In Witold Abramowicz, Robert Tolksdorf, and Krzysztof Węcel, editors, *Business Information Systems Workshops*, pages 240–242, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15402-7.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*, 2016.

Gillian K Hadfield. Weighing the value of vagueness: An economic perspective on precision in the law. *California Law Review*, 82:541, 1994.

Jaap Hage. A theory of legal reasoning and a logic to match. *Artificial Intelligence and Law*, 4 (3-4):199–273, 1996.

Birgit Hamp and Helmut Feldweg. Germanet-a lexical-semantic net for german. *Automatic information extraction and building of lexical semantic resources for NLP applications*, 1997.

Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 531–548, 2018.

Marti A. Hearst. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997. URL `https://www.aclweb.org/anthology/J97-1003`.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701, 2015.

Robert A Hillman. On-line consumer standard-form contracting practices: A survey and discussion of legal implications. 2005.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Julia Hockenmaier. Creating a CCGbank and a wide-coverage CCG lexicon for German. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 505–512, Sydney, Australia, July 2006. Association for Computational Linguistics. doi: 10.3115/1220175.1220239. URL `https://www.aclweb.org/anthology/P06-1064`.

Günter Hörmann. Die verbraucherzentralen. In *Verbraucherwissenschaften*, pages 517–523. Springer, 2017.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA, June 2006. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/N06-2015`.

Baotian Hu, Qingcai Chen, and Fangze Zhu. LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1229. URL `https://www.aclweb.org/anthology/D15-1229`.

Rik Huijzer. Automatically responding to customers. Master's thesis, Eindhoven University of Technology, 2019.

Margareta Hult and Sven-Åke Lennung. Towards a definition of action research: A note and bibliography. *Journal of Management Studies*, 17(2):241–250, 1980. doi: 10.1111/j.1467-6486.1980.tb00087.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6486.1980.tb00087.x`.

IHK Frankfurt am Main. Allgemeine Geschäftsbedingungen (AGB). https://www.frankfurt-main.ihk.de/recht/themen/vertragsrecht/agb/index.html, 2020. Last accessed 2020-07-10.

IHK Munich and Upper Bavaria. Allgemeine Geschäftsbedingungen für einen Webshop. https://www.ihk-muenchen.de/ihk/documents/Recht-Steuern/Vertragsrecht/AGB-Webshop_2020.docx, 2020. Last accessed 2020-07-09.

Institut für Handelsforschung. Online monitor 2019, 2019.

IWW Institut. Vertraglich höhere verzugszinsen vereinbaren? *FMP Forderungsmanagement professionell*, (2):37, 2011.

Andrew JI Jones and Marek Sergot. Deontic logic in the representation of law: Towards a methodology. *Artificial Intelligence and Law*, 1(1):45–64, 1992.

Karen Sparck Jones and Julia Rose Galliers, editors. *Evaluating Natural Language Processing Systems, An Analysis and Review*, volume 1083 of *Lecture Notes in Computer Science*. Springer, 1996. ISBN 3-540-61309-9. doi: 10.1007/BFb0027470. URL https://doi.org/10.1007/BFb0027470.

Audun Jøsang and Viggo A Bondi. Legal reasoning with subjective logic. *Artificial Intelligence and Law*, 8(4):289–315, 2000.

S. Joshi, P. Shah, and A. K. Pandey. Location identification, extraction and disambiguation using machine learning in legal contracts. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pages 1–5, 2018.

Arshdeep Kaur and Bojan Bozic. Convolutional neural network-based automatic prediction of judgments of the european court of human rights. In *AICS*, pages 458–469, 2019.

Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

Tibor Kiss and Jan Strunk. Unsupervised multilingual sentence boundary detection. *Computational linguistics*, 32(4):485–525, 2006.

Bernd Klemm, Hendrik Kornbichler, Hans-Peter Löw, Ingrid Ohmann-Sauer, Eckard Schwarz, Thomas Ubber, Andreas Ege, Ann-Christine Hamisch, Heiko Langer, Anja Lingscheid, et al. *Beck'sches Formularbuch Arbeitsrecht*. C. H. Beck Verlag, Munich, 2014. ISBN 978-3-406-62565-7.

Kira Klimt, Daniel Braun, Daniela Schneider, and Florian Matthes. Muclex: A german lexicon for surface realisation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4655–4659, Marseille, France, May 2020. European Language Resources Association. URL https://www.aclweb.org/anthology/2020.lrec-1.572.

Oleksandra Klymenko, Daniel Braun, and Florian Matthes. Automatic text summarization: A state-of-the-art review. In *Proceedings of the 22nd International Conference on Enterprise Information Systems - Volume 1: ICEIS,*, pages 648–655. INSTICC, SciTePress, 2020. ISBN 978-989-758-423-7. doi: 10.5220/0009723306480655.

Philipp Koehn and Kevin Knight. Empirical methods for compound splitting. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 187–193, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. ISBN 1-333-56789-0. doi: 10.3115/1067807.1067833. URL https://doi.org/10.3115/1067807.1067833.

Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 441–450, 2010.

Elizabeth Koshy, Valsa Koshy, and Heather Waterman. *Action research in healthcare*. Sage, 2010.

Tomaž Kovačič. *Evaluating web content extraction algorithms*. PhD thesis, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, 2012.

Zewang Kuanzhuo, Li Lin, and Zhao Weina. SimpleNLG-TI: Adapting SimpleNLG to Tibetan. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 86–90, Dublin, Ireland, December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.inlg-1.12.

Catherine Lai, Mireia Farrús, and Johanna D Moore. Automatic paragraph segmentation with lexical and prosodic features. *Interspeech 2016; 2016 Sep 08-12; San Francisco (CA).[place unknown]: ISCA; 2016. p. 1034-8.*, 2016.

Guiraude Lame. Using nlp techniques to identify legal ontology components: Concepts and relations. *Artificial Intelligence and Law*, 12(4):379–396, 2004.

Landgericht Dortmund. 10 o 19/19, 2020.

Landgericht Koblenz. 4 hko 165/05. https://www.juris.de/jportal/prev/KORE550962006, 2019. Accessed 2020-03-30.

Landgericht Köln. 31 o 164/18, 2018.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1131. URL https://www.aclweb.org/anthology/D19-1131.

Reed C. Lawlor. What computers can do: Analysis and prediction of judicial decisions. *American Bar Association Journal*, 49(4):337–344, 1963. ISSN 00027596, 21627975. URL http://www.jstor.org/stable/25722338.

JeeHee Lee, June-Seong Yi, and JeongWook Son. Development of automatic-extraction model of poisonous clauses in international construction contracts using rule-based nlp. *Journal of Computing in Civil Engineering*, 33(3):04019003, 2019. doi: 10.1061/(ASCE)CP.1943-5487.

0000807. URL https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29CP.1943-5487.0000807.

Michael Legenc. Using natural language processing and machine learning to assist first-level customer support for contract management. Master's thesis, Technical University of Munich, 2018.

Elena Leitner, Georg Rehm, and Julián Moreno-Schneider. A dataset of german legal documents for named entity recognition. *arXiv*, pages arXiv–2003, 2020.

Gaël Lejeune and Lichao Zhu. A new proposal for evaluating web page cleaning tools. *Computación y Sistemas*, 22(4), 2018.

Kurt Lewin. Action research and minority problems. *Journal of social issues*, 2(4):34–46, 1946.

Wolfgang Lezius. Morphy-german morphology, part-of-speech tagging and applications. In *Proceedings of the 9th EURALEX International Congress*, pages 619–623. University of Stuttgart Stuttgart, 2000.

Wolfgang Lezius, Stefanie Dipper, and Arne Fitschen. Imslex representing morphological and syntactic information in a relational database. In *Proceedings of the 9th EURALEX International Congress*, pages 133–139. Citeseer, 2000.

Daniel Liebig. Änderungen an Bürgerliches Gesetzbuch (BGB). In Bundesrecht - tagaktuell. https://www.buzer.de/gesetz/6597/l.htm, 2020. Last accessed 2020-07-15.

Ruta Liepina, Giuseppe Contissa, Kasper Drazewski, Francesca Lagioia, Marco Lippi, Hans-Wolfgang Micklitz, Przemysław Pałka, Giovanni Sartor, and Paolo Torroni. Gdpr privacy policies in claudette: Challenges of omission, context and multilingualism. In *Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Texts co-located with the 17th International Conference on Artificial Intelligence and Law (ICAIL 2019)*, 2019. URL http://ceur-ws.org/Vol-2385/paper9.pdf.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

Marco Lippi, Przemyslaw Palka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Yannis Panagis, Giovanni Sartor, and Paolo Torroni. Automated detection of unfair clauses in online consumer contracts. In *JURIX*, pages 145–154, 2017.

Marco Lippi, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Przemysław Pałka, Giovanni Sartor, and Paolo Torroni. Consumer protection requires artificial intelligence. *Nature Machine Intelligence*, 1(4):168–169, Apr 2019a. ISSN 2522-5839. doi: 10.1038/s42256-019-0042-3. URL https://doi.org/10.1038/s42256-019-0042-3.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139, Jun 2019b. ISSN 1572-8382. doi: 10.1007/s10506-019-09243-2. URL https://doi.org/10.1007/s10506-019-09243-2.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139, 2019c.

Feifan Liu and Yang Liu. Correlation between ROUGE and human evaluation of extractive meeting summaries. In *Proceedings of ACL-08: HLT, Short Papers*, pages 201–204, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P08-2051`.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. Benchmarking natural language understanding services for building conversational agents. 4 2019. URL `https://iwsds2019.unikore.it/`. 10th International Workshop on Spoken Dialogue Systems Technology 2019, IWSDS 2019 ; Conference date: 24-04-2019 Through 26-04-2019.

Elena Lloret and Manuel Palomar. Text summarisation in progress: a literature review. *Artificial Intelligence Review*, 37(1):1–41, 2012.

Jerome Louvel, Thierry Templier, and Thierry Boileau. *Restlet in Action: Developing RESTful Web APIs in Java*. Manning Publications Co., USA, 2012. ISBN 193518234X.

Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Ross A Malaga. Worst practices in search engine optimization. *Communications of the ACM*, 51(12):147–150, 2008.

Tomohiro Manabe and Keishi Tajima. Extracting logical hierarchical structure of html documents based on headings. *Proc. VLDB Endow.*, 8(12):1606–1617, August 2015. ISSN 2150-8097. doi: 10.14778/2824032.2824058. URL `https://doi.org/10.14778/2824032.2824058`.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-5010. URL `https://www.aclweb.org/anthology/P14-5010`.

Steve Mansfield-Devine. The malware arms race. *Computer Fraud & Security*, 2018(2):15–20, 2018.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. URL `https://www.aclweb.org/anthology/J93-2004`.

Andrei Marmor. The pragmatics of legal language. *Ratio Juris*, 21(4):423–452, 2008.

Florencia Marotta-Wurgler and Robert Taylor. Set in stone: Change and innovation in consumer standard-form contracts. *NYUL Rev.*, 88:240, 2013.

Andrew D. Martin, Kevin M. Quinn, Theodore W. Ruger, and Pauline T. Kim. Competing approaches to predicting supreme court decision making. *Perspectives on Politics*, 2(4):761–767, 2004. doi: 10.1017/S1537592704040502.

Alessandro Mazzei, Cristina Battaglino, and Cristina Bosco. Simplenlg-it: adapting simplenlg to italian. In *Proceedings of the 9th International Natural Language Generation conference*, pages 184–192, 2016.

Aleecia M McDonald and Lorrie Faith Cranor. The cost of reading privacy policies. *Isjlp*, 4:543, 2008.

Jean McNiff. *Teaching as learning: An action research approach.* Taylor & Francis US, 1993.

Masha Medvedeva, Michel Vols, and Martijn Wieling. Using machine learning to predict decisions of the european court of human rights. *Artificial Intelligence and Law*, 28(2):237–266, 2020.

Parth Mehta. From extractive to abstractive summarization: A journey. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 100–106, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-3015. URL https://www.aclweb.org/anthology/P16-3015.

Merriam-Webster. a. In Merriam-Webster.com dictionary. https://www.merriam-webster.com/dictionary/a, 2020a. Last accessed 2020-07-15.

Merriam-Webster. an. In Merriam-Webster.com dictionary. https://www.merriam-webster.com/dictionary/an, 2020b. Last accessed 2020-07-15.

Merriam-Webster. Deliver. In Merriam-Webster.com dictionary. https://www.merriam-webster.com/dictionary/deliver, 2020c. Last accessed 2020-07-20.

Merriam-Webster. Paragraph. In Merriam-Webster.com dictionary. https://www.merriam-webster.com/dictionary/paragraph, 2020d. Last accessed 2020-07-15.

Yishu Miao and Phil Blunsom. Language as a latent variable: Discrete generative models for sentence compression. *arXiv preprint arXiv:1609.07317*, 2016.

Hans-W Micklitz, Przemysław Pałka, and Yannis Panagis. The empire strikes back: digital control of unfair terms of online services. *Journal of consumer policy*, 40(3):367–388, 2017.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013. URL http://arxiv.org/abs/1301.3781.

Marie-Francine Moens and Roxana Angheluta. Concept extraction from legal cases: The use of a statistic of coincidence. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law*, ICAIL '03, page 142–146, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581137478. doi: 10.1145/1047788.1047823. URL https://doi.org/10.1145/1047788.1047823.

Carlos Molina-Jimenez, Santosh Shrivastava, Ellis Solaiman, and John Warne. Run-time monitoring and enforcement of electronic contracts. *Electronic Commerce Research and Applications*, 3(2):108–125, 2004.

David Nadeau. *Semi-supervised named entity recognition: learning to recognize 100 entity types with little supervision.* PhD thesis, University of Ottawa, 2007.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1028. URL `https://www.aclweb.org/anthology/K16-1028`.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*, pages 3075–3081, 2017.

Dmitry Namiot and Manfred Sneps-Sneppe. On micro-services architecture. *International Journal of Open Information Technologies*, 2(9):24–27, 2014.

Shashi Narayan, Nikos Papasarantopoulos, Shay B Cohen, and Mirella Lapata. Neural extractive summarization with side information. *arXiv preprint arXiv:1704.04530*, 2017.

Pablo E Navarro and Jorge L Rodríguez. *Deontic logic and legal systems.* Cambridge University Press, 2014.

John J. Nay. Predicting and understanding law-making with word vectors and an ensemble model. *PLOS ONE*, 12(5):1–14, 05 2017. doi: 10.1371/journal.pone.0176999. URL `https://doi.org/10.1371/journal.pone.0176999`.

Ani Nenkova, Kathleen McKeown, et al. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233, 2011.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1256. URL `https://www.aclweb.org/anthology/P19-1256`.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 380–384, 2013.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175, 2013.

Jonathan A Obar and Anne Oeldorf-Hirsch. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1):128–147, 2020.

Oberlandesgericht Köln. 6 u 184/19. http://www.justiz.nrw.de/nrwe/olgs/koeln/j2020/6_U_184_19_Urteil_20200219.html, 2019. Accessed 2020-03-30.

Oberlandesgericht München. 29 u 634/11. https://openjur.de/u/438877.html, 2011. Accessed 2020-10-08.

Oberlandesgericht München. 29 u 4666/18. https://www.gesetze-bayern.de/Content/Document/Y-300-Z-BECKRS-B-2019-N-23796, 2019. Accessed 2020-03-30.

National Institute of Standards and Technology. Duc 2002. https://www-nlpir.nist.gov/projects/duc/data/2002_data.html, 2002. Last accessed 2019-12-15.

Malte Ostendorff, Till Blume, and Saskia Ostendorff. Towards an open platform for legal information. *arXiv preprint arXiv:2005.13342*, 2020.

Zina O'Leary. *The Essential Guide to Doing Your Research Project.* SAGE Publications Ltd, London, UK, 2017.

Stefan Palan and Christian Schitter. Prolific.ac - a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22 – 27, 2018. ISSN 2214-6350. doi: https://doi.org/10.1016/j.jbef.2017.12.004. URL http://www.sciencedirect.com/science/article/pii/S2214635017300989.

Monica Palmirani, Guido Governatori, Antonino Rotolo, Said Tabet, Harold Boley, and Adrian Paschke. Legalruleml: Xml-based rules and norms. In Frank Olken, Monica Palmirani, and Davide Sottara, editors, *Rule-Based Modeling and Computing on the Semantic Web*, pages 298–312, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-24908-2.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

Andrea M. Partikel. *Formularbuch für Sportverträge.* C. H. Beck Verlag, Munich, 2015. ISBN 978-3-406-66563-9.

Daraksha Parveen, Hans-Martin Ramsl, and Michael Strube. Topical coherence for graph-based extractive summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1949–1954, 2015.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12: 2825–2830, 2011.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Kai Petersen, Claes Wohlin, and Dejan Baca. The waterfall model in large-scale development. In *International Conference on Product-Focused Software Process Improvement*, pages 386–400. Springer, 2009.

Felipe Pezoa, Juan L Reutter, Fernando Suarez, Martín Ugarte, and Domagoj Vrgoč. Foundations of json schema. In *Proceedings of the 25th International Conference on World Wide Web*, pages 263–273, 2016.

Victoria C Plaut and Robert P Bartlett III. Blind consent? a social psychological investigation of non-readership of click-through agreements. *Law and human behavior*, 36(4):293, 2012.

Jan Pomikálek. *Removing boilerplate and duplicate content from web corpora.* PhD thesis, Masarykova univerzita, Fakulta informatiky, 2011.

Presse- und Informationsamt der Bundesregierung. Verbraucherverbände. `https://www.bundesregierung.de/breg-de/themen/tipps-fuer-verbraucher/verbraucherverbaende-399350`, 2019. Last accessed 2020-12-12.

Thomas Proisl and Peter Uhrig. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin, 2016. Association for Computational Linguistics (ACL). URL `http://aclweb.org/anthology/W16-2607`.

Ofer Raban. The fallacy of legal certainty: Why vague legal standards may be better for capitalism and liberalism. *BU Pub. Int. LJ*, 19:175, 2009.

Anis Nadiah Che Abdul Rahman, Imran Ho Abdullah, Intan Safinaz Zainuddin, and Azhar Jaludin. The comparisons of ocr tools: A conversion case in the malaysian hansard corpus development. *MALAYSIAN JOURNAL OF COMPUTING*, 4(2):335–348, 2019. ISSN 2600-8238. doi: 10.24191/mjoc.v4i2.5626. URL `http://103.8.145.246/index.php/mjoc/article/view/5626`.

GJ Rath, A Resnick, and TR Savage. The formation of abstracts by the selection of sentences. part i. sentence selection by men and machines. *Journal of the Association for Information Science and Technology*, 12(2):139–141, 1961.

Ehud Reiter. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401, 2018. doi: 10.1162/coli_a_00322. URL `https://doi.org/10.1162/coli_a_00322`.

Ehud Reiter and Robert Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87, 1997. doi: 10.1017/S1351324997001502.

Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems.* Cambridge University Press, 2000.

Kervy Rivas Rojas, Gina Bustamante, Marco A Sobrevilla Cabezudo, and Arturo Oncevay. Efficient strategies for hierarchical text classification: External knowledge and auxiliary tasks. *arXiv preprint arXiv:2005.02473*, 2020.

LL Royakkers. *Extending deontic logic for the formalisation of legal rules*, volume 36. Springer Science & Business Media, 2013.

Theodore W. Ruger, Pauline T. Kim, Andrew D. Martin, and Kevin M. Quinn. The supreme court forecasting project: Legal and political science approaches to predicting supreme court decisionmaking. *Columbia Law Review*, 104(4):1150–1210, 2004. ISSN 00101958. URL `http://www.jstor.org/stable/4099370`.

Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1044. URL `https://www.aclweb.org/anthology/D15-1044`.

Norman Sadeh, Alessandro Acquisti, Travis D Breaux, Lorrie Faith Cranor, Aleecia M McDonald, Joel R Reidenberg, Noah A Smith, Fei Liu, N Cameron Russell, Florian Schaub, et al. The usable privacy policy project. In *Technical report, Technical Report, CMU-ISR-13-119*. Carnegie Mellon University, 2013.

Horacio Saggion and Thierry Poibeau. Automatic text summarization: Past, present and future. In *Multi-source, multilingual information extraction and summarization*, pages 3–21. Springer, 2013.

George Sanchez. Sentence boundary detection in legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 31–38, 2019.

J Savelka and Kevin D Ashley. Using conditional random fields to detect different functional types of content in decisions of united states courts with example application to sentence boundary detection. In *Workshop on Automated Semantic Analysis of Information in Legal Texts*, 2017.

Martin Schirmbacher. Allgemeine geschäftsbedingungen (online-shop). `https://www.bevh.org/fileadmin/content/01_leistungen/rechtshilfen/muster-agb/muster-agb-internetshop-2018.pdf`, 2018. Last accessed 2020-07-10.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2004/pdf/468.pdf`.

Xavier Schmitt, Sylvain Kubler, Jérémy Robert, Mike Papadakis, and Yves LeTraon. A replicable comparison study of ner software: Stanfordnlp, nltk, opennlp, spacy, gate. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 338–343. IEEE, 2019.

Jetze Schuurmans, Flavius Frasincar, and E. Cambria. Intent classification for dialogue utterances. *IEEE Intelligent Systems*, 35(1):82–88, January 2020. ISSN 1541-1672. doi: 10.1109/MIS.2019.2954966. URL `https://doi.org/10.1109/MIS.2019.2954966`.

Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

Rico Sennrich and Beat Kunz. Zmorge: A german morphological lexicon extracted from wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 14)*, pages 1063–1067, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2014/pdf/116_Paper.pdf`.

Gwenaelle Cunha Sergio and Minho Lee. Stacked debert: All attention in incomplete data for text classification. *arXiv preprint arXiv:2001.00137*, 2020.

K. Shridhar, A. Dash, A. Sahu, G. G. Pihlgren, P. Alonso, V. Pondenkandath, G. Kovács, F. Simistira, and M. Liwicki. Subword semantic hashing for intent classification on small datasets. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, 2019. doi: 10.1109/IJCNN.2019.8852420.

Barbara Sommer and Ferdinans von Stumm. Fernabsatz von waren und dienstleistungen. In Wolfgang Weitnauer and Tilman Mueller-Stöfen, editors, *Beckśches Formularbuch IT-Recht*, chapter J, pages 715–761. C. H. Beck Verlag, Munich, 4 edition, 2017.

Mohammad S Sorower. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 18:1–25, 2010.

Alejandro Ramos Soto, Julio Janeiro Gallardo, and Alberto Bugarín Diz. Adapting simplenlg to spanish. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 144–148, 2017.

Caroline Sporleder and Mirella Lapata. Broad coverage paragraph segmentation across languages and domains. *ACM Trans. Speech Lang. Process.*, 3(2):1–35, July 2006. ISSN 1550-4875. doi: 10.1145/1149290.1151098. URL `https://doi.org/10.1145/1149290.1151098`.

P Sriramya and RA Karthika. Providing password security by salted password hashing using bcrypt algorithm. *ARPN journal of engineering and applied sciences*, 10(13):5551–5556, 2015.

statista. Ecommerce report 2020, 2020.

statista. Ecommerce - europe. `https://www-statista-com.eaccess.ub.tum.de/outlook/243/102/ecommerce/europe`, 2020. Accessed 2020-05-02.

Johannes Steger and Egon Stemle. KrdWrd: Architecture for Unified Processing of Web Content. In Iñaki Alegria, Igor Leturia, and Serge Sharoff, editors, *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*, pages 63–70. Elhuyar Fundazioa, 2009. URL `https://www.sigwac.org.uk/raw-attachment/wiki/WAC5/WAC5_proceedings.pdf`.

Benjamin Strickson and Beatriz De La Iglesia. Legal judgement prediction for uk courts. In *Proceedings of the 2020 The 3rd International Conference on Information Science and System*, ICISS 2020, page 204–209, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450377256. doi: 10.1145/3388176.3388183. URL `https://doi.org/10.1145/3388176.3388183`.

Ernest T Stringer. *Action research.* Sage publications, 4 edition, 2014. ISBN 9781483320731.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache, 2019.

Kyoko Sugisaki and Don Tuggener. German compound splitting using the compound productivity of morphemes. In *14th Conference on Natural Language Processing-KONVENS 2018*, pages 141–147. Austrian Academy of Sciences Press, 2018.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

P. Szymański and T. Kajdanowicz. A scikit-based Python environment for performing multi-label classification. *ArXiv e-prints*, February 2017.

Randy L. Teach and Edward H. Shortliffe. An analysis of physician attitudes regarding computer-based clinical consultation systems. *Computers and Biomedical Research*, 14(6):542 – 558, 1981. ISSN 0010-4809. doi: https://doi.org/10.1016/0010-4809(81)90012-4. URL `http://www.sciencedirect.com/science/article/pii/0010480981900124`.

Saurabh Tiwari, Santosh Singh Rathore, and Atul Gupta. Selecting requirement elicitation techniques for software projects. In *2012 CSI Sixth International Conference on Software Engineering (CONSEG)*, pages 1–10. IEEE, 2012.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003. URL `https://www.aclweb.org/anthology/W03-0419`.

Damiano Torre, Sallam Abualhaija, Mehrdad Sabetzadeh, Lionel Briand, Katrien Baetens, Peter Goes, and Sylvie Forastier. An ai-assisted approach for checking the completeness of privacy policies against gdpr. In *in Proceedings of the 28th IEEE International Requirements Engineering Conference (RE'20)*, 2020.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling coverage for neural machine translation. *arXiv preprint arXiv:1601.04811*, 2016.

Andre Valente. *Types and Roles of Legal Ontologies*, pages 65–76. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005. ISBN 978-3-540-32253-5. doi: 10.1007/978-3-540-32253-5_5. URL `https://doi.org/10.1007/978-3-540-32253-5_5`.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan, October–November 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-8643. URL `https://www.aclweb.org/anthology/W19-8643`.

Jean Vancoppenolle, Eric Tabbert, Gerlof Bouma, and Manfred Stede. A german grammar for generation in open ccg. In *Multilingual resources and multilingual applications: Proceedings of* *the Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, pages 145–150, 2011.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008, 2017.

Pierre-Luc Vaudry and Guy Lapalme. Adapting simplenlg for bilingual english-french realisation. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 183–187, 2013.

Verbraucherzentrale Bundesverband e.V. Marktwächter mahnen paypal ab. `https://www.vzbv.de/pressemitteilung/marktwaechter-mahnen-paypal-ab`, 2018a. Accessed 2020-03-30.

Verbraucherzentrale Bundesverband e.V. Jahresbericht 2017, 2018b.

Verbraucherzentrale Bundesverband e.V. Jahresbericht 2018. `https://www.vzbv.de/content/bericht-2018`, 2019a. Accessed 2020-05-04.

Verbraucherzentrale Bundesverband e.V. Jahresbericht 2018, 2019b.

Verbraucherzentrale Bundesverband e.V. Ein junger verband mit langer geschichte. `https://www.vzbv.de/ueber-uns/geschichte`, 2020a. Accessed 2020-03-20.

Verbraucherzentrale Bundesverband e.V. Jahresbericht 2019, 2020b.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700, 2015.

Purna Vithlani and CK Kumbharana. Comparative study of character recognition tools. *International Journal of Computer Applications*, 118(9), 2015.

Stephan Walter and Manfred Pinkal. Automatic extraction of definitions from German court decisions. In *Proceedings of the Workshop on Information Extraction Beyond The Document*, pages 20–28, Sydney, Australia, July 2006. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W06-0203`.

Bernhard Waltl, Thomas Reschenhofer, and Florian Matthes. Modeling, execution and analysis of formalized legal norms in model based decision structures. In *AI Approaches to the Complexity of Legal Systems*, pages 157–171. Springer, 2015.

Bernhard Waltl, Jörg Landthaler, and Florian Matthes. Differentiation and empirical analysis of reference types in legal documents. In *JURIX*, pages 211–214, 2016a.

Bernhard Waltl, Florian Matthes, Tobias Waltl, and Thomass Grass. Lexia: A data science environment for semantic analysis of german legal texts. In *IRIS 2016*, 2016b. doi: 10.38023/dd760bfd-13a0-4930-8226-62a9cf625c54.

Bernhard Waltl, Georg Bonczek, Elena Scepankova, Jörg Landthaler, and Florian Matthes. Predicting the outcome of appeal decisions in germany's tax law. In Peter Parycek, Yannis Charalabidis, Andrei V. Chugunov, Panos Panagiotopoulos, Theresa A. Pardo, Øystein Sæbø, and Efthimios Tambouris, editors, *Electronic Participation*, pages 89–99, Cham, 2017a. Springer International Publishing. ISBN 978-3-319-64322-9.

Bernhard Waltl, Jörg Landthaler, Elena Scepankova, Florian Matthes, THOMAS Geiger, Christoph Stocker, and Christian Schneider. Automated extraction of semantic information from german legal documents. In *IRIS 2017*, 2017b.

Bernhard Ernst Waltl. *Semantic Analysis and Computational Modeling of Legal Documents*. Dissertation, Technische Universität München, München, 2018.

Xiaojun Wan. Towards a unified approach to simultaneous single-document and multi-document summarizations. In *Proceedings of the 23rd international conference on computational linguistics*, pages 1137–1145. Association for Computational Linguistics, 2010.

Xiao-Lin Wang and Bao-Liang Lu. Flatten hierarchies for large-scale hierarchical text categorization. In *2010 Fifth International Conference on Digital Information Management (ICDIM)*, pages 139–144. IEEE, 2010.

J.L. Weiner. Blah, a system which explains its reasoning. *Artificial Intelligence*, 15(1):19 – 48, 1980. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(80)90021-1. URL `http://www.sciencedirect.com/science/article/pii/0004370280900211`.

Wolfgang Weitnauer and Tilman Mueller-Stöfen. *Beckšches Formularbuch IT-Recht*. C. H. Beck Verlag, Munich, 4th edition edition, 2017. ISBN 978-3-406-69303-8.

Felix Weißl. Effects of integrating assistance features in a customer support system. Master's thesis, Technical University of Munich, 2018.

Marion Weller-Di Marco. Simple compound splitting for German. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 161–166, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1722. URL `https://www.aclweb.org/anthology/W17-1722`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

Leiming Yan, Yuhui Zheng, and Jie Cao. Few-shot learning for short text classification. *Multimedia Tools and Applications*, 77(22):29799–29810, 2018.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*, 2019.

L. Richard Ye and Paul E. Johnson. The impact of explanation facilities on user acceptance of expert systems advice. *MIS Q.*, 19(2):157–172, June 1995. ISSN 0276-7783. doi: 10.2307/249686. URL `https://doi.org/10.2307/249686`.

Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. Comparative study of CNN and RNN for natural language processing. *CoRR*, abs/1702.01923, 2017. URL `http://arxiv.org/abs/1702.01923`.

Thomas Zerres. Principles of the german law on standard terms of contract. *Jurawelt*, 2014.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.466. URL `https://www.aclweb.org/anthology/2020.acl-main.466`.

Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1106–1117, 2020.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.

Abbreviations

AGB	Allgemeine Geschäftsbedingungen
AMT	Amazon Mechanical Turk
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BGB	Bürgerliches Gesetzbuch
BMJV	Bundesministerium für Justiz und Verbraucherschutz
BPMN	Business Process Model and Notation
CNN	Convolutional Neural Network
ECHR	European Court of Human Rights
ECJ	European Court of Justice
ELRA	European Language Resources Association
EU	European Union
FernAbsG	Fernabsatzgesetz

GDPR	General Data Protection Regulation
HMM	Hidden Markov Model
IE	Information Extraction
IHK	Industrie- und Handelskammer
ISLRN	International Standard Language Resource Number
KSchG	Konsumentenschutzgesetz
LSTM	Long short-term memory
ML	Machine Learning
MLP	Multilayer Perceptron
NDA	Non-Disclosure Agreement
NER	Named Entity Recognition
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
NGO	Non Governmental Organisation
OCR	Optical Character Recognition
ODR	Online Dispute Resolution
PoS	Part of Speech
RDG	Rechtsdienstleistungsgesetz
RNN	Recurrent Neural Network
RVG	Rechtsanwaltsvergütungsgesetz

SaaS	Software as a Service
SVM	Support Vector Machine
T&C	Terms and Conditions
ToS	Terms of Service
TFEU	Treaty on the Functioning of the European Union
TPU	Tensor Processing Unit
UI	User Interface
UK	United Kingdom
UKlaG	Unterlassungsklagengesetz
vzbv	Bundesverband der Verbraucherzentrale

APPENDIX A

Linguistic Annotations

This appendix gives an overview of the linguistic annotations used throughout this thesis.

A.1. Part of Speech and Constituent Tags

This work uses the Part of Speech tags as used in the Penn Treebank Projekt (Marcus et al., 1993). A list of all tags and their description is shown in Table A.1.

Tag	**Description**	**Tag**	**Description**
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present

Table A.1.: List of PoS tags as used by the Penn Treebank Project (Marcus et al., 1993)

Tag	**Description**	**Tag**	**Description**
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

Table A.1.: List of PoS tags as used by the Penn Treebank Project (Marcus et al., 1993)

A.2. Dependency Types

For the annotation of dependency trees, we use the dependency types in Table A.2 from the Stanford typed dependencies manual (de Marneffe and Manning, 2008).

Tag	Description	Tag	Description
acomp	adjectival complement	nn	noun compound modifier
advcl	adverbial clause modifier	npadvmod	noun phrase adverbial modifier
advmod	adverbial modifier	nsubj	nominal subject
agent	agent	nsubjpass	passive nominal subject
amod	adjectival modifier	num	numeric modifier
appos	appositional modifier	number	element of compound number
arg	argument	obj	object
aux	auxiliary	parataxis	parataxis
auxpass	passive auxiliary	pobj	object of preposition
cc	coordination	poss	possession modifier
ccomp	clausal complement with internal subject	possessive	possessive modifier (ś)
comp	complement	preconj	preconjunct
conj	conjunct	predet	predeterminer
cop	copula	prep	prepositional modifier
csubj	clausal subject	prt	phrasal verb particle
csubjpass	passive clausal subject	punct	punctuation
dep	dependent	quantmod	quantifier modifier
det	determiner	rcmod	relative clause modifier
dobj	direct object	ref	referent
expl	expletive (expletive “there”)	root	root
goeswith	goes with	sdep	semantic dependent
iobj	indirect object	subj	subject

Table A.2.: List of dependency types as used by the Stanford CoreNLP library (de Marneffe and Manning, 2008)

Tag	Description	Tag	Description
mark	marker (word introducing an advcl or ccomp	tmod	temporal modifier
mod	modifier	vmod	reduced, non-finite verbal modifier
mwe	multiword expression modifier	xcomp	clausal complement with external subject
neg	negation modifier	xsubj	controlling subject

Table A.2.: List of dependency types as used by the Stanford CoreNLP library (de Marneffe and Manning, 2008)

APPENDIX B

NLU Evaluation Corpora

This sections gives on overview of the data format which is used by the corpora for the evaluation of NLU services, described in Section 2.2 of this thesis and Braun et al. (2017a).

B.1. Chatbot Dataset

```
{
  "text": "what is the cheapest connection between quiddestrasse and hauptbahnhof?",
  "intent": "FindConnection",
  "entities": [
    {
      "entity": "Criterion",
      "start": 3,
      "stop": 3
    },
    {
      "entity": "StationStart",
      "start": 6,
      "stop": 6
    },
    {
      "entity": "StationDest",
      "start": 8,
      "stop": 8
    }
  ]
},
{
  "text": "when is the next u6 leaving from garching?",
  "intent": "DepartureTime",
  "entities": [
    {
      "entity": "Line",
      "start": 4,
      "stop": 4
    },
    {
      "entity": "StationStart",
      "start": 7,
      "stop": 7
    }
  ]
}
```

Listing B.1: Example entry from the Chatbot Dataset

B.2. Web Applications Dataset

```
{
  "text": "How can I delete my Twitter account?",
  "url": "http://webapps.stackexchange.com/questions/57/how-can-i-delete-my-twitter-account",
  "author": "Jared Harley",
  "answer": {
    "text": "[...]",
    "author": "Ken Pespisa"
  },
  "intent": "Delete Account",
  "entities": [
    {
      "text": "Twitter",
      "stop": 5,
      "start": 5,
      "entity": "WebService"
    }
  ]
},
{
  "text": "Is it possible to export my data from Trello to back it up?",
  "url": "http://webapps.stackexchange.com/questions/18975/is-it-possible-to-
    ↪ export-my-data-from-trello-to-back-it-up",
  "author": "Clare Macrae",
  "answer": {
    "text": "[...]",
    "author": "Daniel LeCheminant"
  },
  "intent": "Export Data",
  "entities": [
    {
      "text": "Trello",
      "stop": 8,
      "start": 8,
      "entity": "WebService"
    }
  ]
}
```

Listing B.2: Example entry from the Web Applications Dataset

B.3. Ask Ubuntu Dataset

```
{
  "text": "How do I install the HP F4280 printer?",
  "url": "http://askubuntu.com/questions/24073/how-do-i-install-the-hp-f4280-
    ↪ printer",
  "author": "ok comp",
  "answer": {
    "text": "[...]",
    "author": "nejode"
  },
  "intent": "Setup Printer",
  "entities": [
    {
      "text": "HP F4280",
      "stop": 6,
      "start": 5,
      "entity": "Printer"
    }
  ]
},
{
  "text": "What is a good MongoDB GUI client?",
  "url": "http://askubuntu.com/questions/196136/what-is-a-good-mongodb-gui-
    ↪ client",
  "author": "Eyal",
  "answer": {
    "text": "[...]",
    "author": "Eyal"
  },
  "intent": "Software Recommendation",
  "entities": [
    {
      "text": "MongoDB",
      "stop": 4,
      "start": 4,
      "entity": "SoftwareName"
    }
  ]
}
```

Listing B.3: Example entry from the Ask Ubuntu Dataset

APPENDIX C

List of Open-Sourced Results

- **MucLex**
    - Description: A German lexicon for surface realisation based on the crowd-source online lexicon Wiktionary.
    - License: Mozilla Public License 2.0
    - Repository: `https://github.com/sebischair/MucLex`
- **NLU Evaluation Corpora**
    - Description: Three annotated corpora for the evaluation of NLU services.
    - License: Create Commons BY-SA 3.0
    - Repository: `https://github.com/sebischair/NLU-Evaluation-Corpora`
- **NLU Evaluation Scripts**
    - Description: Python scripts for the automated evaluation and comparison of different NLU services.
    - License: MIT License
    - Repository: `https://github.com/sebischair/NLU-Evaluation-Scripts`
- **SimpleNLG-DE**
    - Description: A surface realiser for German based on SimpleNLG
    - License: Mozilla Public License 1.1
    - Repository: `https://github.com/sebischair/SimpleNLG-DE`

- **T&C Detection Corpus**
  - Description: A corpus containing annotated links to 5,232 T&C pages from German and English web-shops and 10,574 links to other pages from the same web-shops.
  - License: Create Commons BY-SA 3.0
  - Repository: `https://github.com/sebischair/TC-Detection-Corpus`

APPENDIX D

Generated T&C

This appendix contains the full text generated by the T&C generators introduced in Section 3.5.

D.1. Trusted Shops

Allgemeine Geschäftsbedingungen

1. Geltungsbereich Für alle Bestellungen über unseren Online-Shop gelten die nachfolgenden AGB. Unser Online-Shop richtet sich ausschließlich an Verbraucher.

Verbraucher ist jede natürliche Person, die ein Rechtsgeschäft zu Zwecken abschließt, die überwiegend weder ihrer gewerblichen noch ihrer selbständigen beruflichen Tätigkeit zugerechnet werden können. Unternehmer ist eine natürliche oder juristische Person oder eine rechtsfähige Personengesellschaft, die bei Abschluss eines Rechtsgeschäfts in Ausübung ihrer gewerblichen oder selbständigen beruflichen Tätigkeit handelt.

2. Vertragspartner, Vertragsschluss, Korrekturmöglichkeiten Der Kaufvertrag kommt zustande mit Muster GmbH.

Mit Einstellung der Produkte in den Online-Shop geben wir ein verbindliches Angebot zum Vertragsschluss über diese Artikel ab. Sie können unsere Produkte zunächst unverbindlich in den Warenkorb legen und Ihre Eingaben vor Absenden Ihrer verbindlichen Bestellung jederzeit korrigieren, indem Sie die hierfür im Bestellablauf vorgesehenen und erläuterten Korrekturhilfen nutzen. Der Vertrag kommt zustande, indem Sie durch Anklicken des Bestellbuttons das Angebot über die im Warenkorb enthaltenen Waren annehmen. Unmittelbar nach dem Absenden der Bestellung erhalten Sie noch einmal eine Bestätigung per E-Mail.

3. Vertragssprache, Vertragstextspeicherung Die für den Vertragsschluss zur Verfügung stehende(n) Sprache(n): Deutsch

Wir speichern den Vertragstext und senden Ihnen die Bestelldaten und unsere AGB in Textform zu. Der Vertragstext ist aus Sicherheitsgründen nicht mehr über das Internet zugänglich.

4. Lieferbedingungen Zuzüglich zu den angegebenen Produktpreisen kommen noch Versandkosten hinzu. Näheres zur Höhe der Versandkosten erfahren Sie bei den Angeboten.

Wir liefern nur im Versandweg. Eine Selbstabholung der Ware ist leider nicht möglich.

5. Bezahlung In unserem Shop stehen Ihnen grundsätzlich die folgenden Zahlungsarten zur Verfügung:

PayPal Im Bestellprozess werden Sie auf die Webseite des Online-Anbieters PayPal weitergeleitet. Um den Rechnungsbetrag über PayPal bezahlen zu können, müssen Sie dort registriert sein bzw. sich erst registrieren, mit Ihren Zugangsdaten legitimieren und die Zahlungsanweisung an uns bestätigen. Nach Abgabe der Bestellung im Shop fordern wir PayPal zur Einleitung der Zahlungstransaktion auf. Die Zahlungstransaktion wird durch PayPal unmittelbar danach automatisch durchgeführt. Weitere Hinweise erhalten Sie beim Bestellvorgang.

Rechnung Sie zahlen den Rechnungsbetrag nach Erhalt der Ware und der Rechnung per Überweisung auf unser Bankkonto. Wir behalten uns vor, den Kauf auf Rechnung nur nach einer erfolgreichen Bonitätsprüfung anzubieten.

6. Eigentumsvorbehalt Die Ware bleibt bis zur vollständigen Bezahlung unser Eigentum.

7. Transportschäden Werden Waren mit offensichtlichen Transportschäden angeliefert, so reklamieren Sie solche Fehler bitte möglichst sofort beim Zusteller und nehmen Sie bitte unverzüglich Kontakt zu uns auf. Die Versäumung einer Reklamation oder Kontaktaufnahme hat für Ihre gesetzlichen Ansprüche und deren Durchsetzung, insbesondere Ihre Gewährleistungsrechte, keinerlei Konsequenzen. Sie helfen uns aber, unsere eigenen Ansprüche gegenüber dem Frachtführer bzw. der Transportversicherung geltend machen zu können.

8. Gewährleistung und Garantien Soweit nicht nachstehend ausdrücklich anders vereinbart, gilt das gesetzliche Mängelhaftungsrecht. Bei gebrauchten Waren gilt: wenn der Mangel nach Ablauf eines Jahres ab Ablieferung der Ware auftritt, sind die Mängelansprüche ausgeschlossen. Mängel, die innerhalb eines Jahres ab Ablieferung der Ware auftreten, können im Rahmen der gesetzlichen Verjährungsfrist von zwei Jahren ab Ablieferung der Ware geltend gemacht werden. Die vorstehenden Einschränkungen und Fristverkürzungen gelten nicht für Ansprüche aufgrund von Schäden, die durch uns, unsere gesetzlichen Vertreter oder Erfüllungsgehilfen verursacht wurden

- bei Verletzung des Lebens, des Körpers oder der Gesundheit
- bei vorsätzlicher oder grob fahrlässiger Pflichtverletzung sowie Arglist
- bei Verletzung wesentlicher Vertragspflichten, deren Erfüllung die ordnungsgemäße Durchführung des Vertrages überhaupt erst ermöglicht und auf deren Einhaltung der Vertragspartner regelmäßig vertrauen darf (Kardinalpflichten)
- im Rahmen eines Garantieversprechens, soweit vereinbart oder
- soweit der Anwendungsbereich des Produkthaftungsgesetzes eröffnet ist. Informationen zu

gegebenenfalls geltenden zusätzlichen Garantien und deren genaue Bedingungen finden Sie jeweils beim Produkt und auf besonderen Informationsseiten im Online-Shop.

9. Haftung Für Ansprüche aufgrund von Schäden, die durch uns, unsere gesetzlichen Vertreter oder Erfüllungsgehilfen verursacht wurden, haften wir stets unbeschränkt
• bei Verletzung des Lebens, des Körpers oder der Gesundheit
• bei vorsätzlicher oder grob fahrlässiger Pflichtverletzung
• bei Garantieversprechen, soweit vereinbart, oder
• soweit der Anwendungsbereich des Produkthaftungsgesetzes eröffnet ist. Bei Verletzung wesentlicher Vertragspflichten, deren Erfüllung die ordnungsgemäße Durchführung des Vertrages überhaupt erst ermöglicht und auf deren Einhaltung der Vertragspartner regelmäßig vertrauen darf, (Kardinalpflichten) durch leichte Fahrlässigkeit von uns, unseren gesetzlichen Vertretern oder Erfüllungsgehilfen ist die Haftung der Höhe nach auf den bei Vertragsschluss vorhersehbaren Schaden begrenzt, mit dessen Entstehung typischerweise gerechnet werden muss. Im Übrigen sind Ansprüche auf Schadensersatz ausgeschlossen. 10. Streitbeilegung Die Europäische Kommission stellt eine Plattform zur Online-Streitbeilegung (OS) bereit, die Sie hier finden. Zur Teilnahme an einem Streitbeilegungsverfahren vor einer Verbraucherschlichtungsstelle sind wir nicht verpflichtet und nicht bereit.

AGB erstellt mit rechtstexter.de.

D.2. Termly

TERMS OF USE

Last updated December 06, 2019

AGREEMENT TO TERMS

These Terms of Use constitute a legally binding agreement made between you, whether personally or on behalf of an entity (“you”) and Test LLC ("Company", “we”, “us”, or “our”), concerning your access to and use of the http://www.example.com website as well as any other media form, media channel, mobile website or mobile application related, linked, or otherwise connected thereto (collectively, the “Site”). You agree that by accessing the Site, you have read, understood, and agreed to be bound by all of these Terms of Use. IF YOU DO NOT AGREE WITH ALL OF THESE TERMS OF USE, THEN YOU ARE EXPRESSLY PROHIBITED FROM USING THE SITE AND YOU MUST DISCONTINUE USE IMMEDIATELY.

Supplemental terms and conditions or documents that may be posted on the Site from time to time are hereby expressly incorporated herein by reference. We reserve the right, in our sole discretion, to make changes or modifications to these Terms of Use at any time and for any reason. We will alert you about any changes by updating the “Last updated” date of these Terms of Use, and you waive any right to receive specific notice of each such change. It is your responsibility to periodically review these Terms of Use to stay informed of updates. You will be subject to, and will be deemed to have been made aware of and to have accepted, the changes in any revised Terms of Use by your continued use of the Site after the date such revised Terms of Use are posted.

The information provided on the Site is not intended for distribution to or use by any person or entity in any jurisdiction or country where such distribution or use would be contrary to law or regulation or which would subject us to any registration requirement within such jurisdiction or country. Accordingly, those persons who choose to access the Site from other locations do so on their own initiative and are solely responsible for compliance with local laws, if and to the extent local laws are applicable.

The Site is intended for users who are at least 13 years of age. All users who are minors in the jurisdiction in which they reside (generally under the age of 18) must have the permission of, and be directly supervised by, their parent or guardian to use the Site. If you are a minor, you must have your parent or guardian read and agree to these Terms of Use prior to you using the Site.

INTELLECTUAL PROPERTY RIGHTS

Unless otherwise indicated, the Site is our proprietary property and all source code, databases, functionality, software, website designs, audio, video, text, photographs, and graphics on the Site (collectively, the "Content") and the trademarks, service marks, and logos contained therein (the "Marks") are owned or controlled by us or licensed to us, and are protected by copyright and trademark laws and various other intellectual property rights and unfair competition laws of the United States, international copyright laws, and international conventions. The Content and the Marks are provided on the Site "AS IS" for your information and personal use only. Except as expressly provided in these Terms of Use, no part of the Site and no Content or Marks may be copied, reproduced, aggregated, republished, uploaded, posted, publicly displayed, encoded, translated, transmitted, distributed, sold, licensed, or otherwise exploited for any commercial purpose whatsoever, without our express prior written permission.

Provided that you are eligible to use the Site, you are granted a limited license to access and use the Site and to download or print a copy of any portion of the Content to which you have properly gained access solely for your personal, non-commercial use. We reserve all rights not expressly granted to you in and to the Site, the Content and the Marks.

USER REPRESENTATIONS

By using the Site, you represent and warrant that: (1) all registration information you submit will be true, accurate, current, and complete; (2) you will maintain the accuracy of such information and promptly update such registration information as necessary; (3) you have the legal capacity and you agree to comply with these Terms of Use; (4) you are not under the age of 13; (5) you are not a minor in the jurisdiction in which you reside, or if a minor, you have received parental permission to use the Site; (6) you will not access the Site through automated or non-human means, whether through a bot, script, or otherwise; (7) you will not use the Site for any illegal or unauthorized purpose; and (8) your use of the Site will not violate any applicable law or regulation.

If you provide any information that is untrue, inaccurate, not current, or incomplete, we have the right to suspend or terminate your account and refuse any and all current or future use of the Site (or any portion thereof).

USER REGISTRATION

You may be required to register with the Site. You agree to keep your password confidential and will be responsible for all use of your account and password. We reserve the right to remove, reclaim, or change a username you select if we determine, in our sole discretion, that such username is inappropriate, obscene, or otherwise objectionable.

PROHIBITED ACTIVITIES

You may not access or use the Site for any purpose other than that for which we make the Site available. The Site may not be used in connection with any commercial endeavors except those that are specifically endorsed or approved by us.

As a user of the Site, you agree not to:

1. Systematically retrieve data or other content from the Site to create or compile, directly or indirectly, a collection, compilation, database, or directory without written permission from us.

USER GENERATED CONTRIBUTIONS

The Site may invite you to chat, contribute to, or participate in blogs, message boards, online forums, and other functionality, and may provide you with the opportunity to create, submit, post, display, transmit, perform, publish, distribute, or broadcast content and materials to us or on the Site, including but not limited to text, writings, video, audio, photographs, graphics, comments, suggestions, or personal information or other material (collectively, "Contributions"). Contributions may be viewable by other users of the Site and through third-party websites. As such, any Contributions you transmit may be treated as non-confidential and non-proprietary. When you create or make available any Contributions, you thereby represent and warrant that:

1. The creation, distribution, transmission, public display, or performance, and the accessing, downloading, or copying of your Contributions do not and will not infringe the proprietary rights, including but not limited to the copyright, patent, trademark, trade secret, or moral rights of any third party. 2. You are the creator and owner of or have the necessary licenses, rights, consents, releases, and permissions to use and to authorize us, the Site, and other users of the Site to use your Contributions in any manner contemplated by the Site and these Terms of Use. 3. You have the written consent, release, and/or permission of each and every identifiable individual person in your Contributions to use the name or likeness of each and every such identifiable individual person to enable inclusion and use of your Contributions in any manner contemplated by the Site and these Terms of Use. 4. Your Contributions are not false, inaccurate, or misleading. 5. Your Contributions are not unsolicited or unauthorized advertising, promotional materials, pyramid schemes, chain letters, spam, mass mailings, or other forms of solicitation. 6. Your Contributions are not obscene, lewd, lascivious, filthy, violent, harassing, libelous, slanderous, or otherwise objectionable (as determined by us). 7. Your Contributions do not ridicule, mock, disparage, intimidate, or abuse anyone. 8. Your Contributions do not advocate the violent overthrow of any government or incite, encourage, or threaten physical harm against another. 9. Your Contributions do not violate any applicable law, regulation, or rule. 10. Your Contributions do not violate the privacy or publicity rights of any third party. 11. Your Contributions do not contain any material that solicits personal information from anyone under the age of 18 or exploits people under the age of 18 in a sexual or violent manner. 12. Your Contributions do

not violate any applicable law concerning child pornography, or otherwise intended to protect the health or well-being of minors; 13. Your Contributions do not include any offensive comments that are connected to race, national origin, gender, sexual preference, or physical handicap. 14. Your Contributions do not otherwise violate, or link to material that violates, any provision of these Terms of Use, or any applicable law or regulation.

Any use of the Site in violation of the foregoing violates these Terms of Use and may result in, among other things, termination or suspension of your rights to use the Site.

CONTRIBUTION LICENSE

By posting your Contributions to any part of the Site, you automatically grant, and you represent and warrant that you have the right to grant, to us an unrestricted, unlimited, irrevocable, perpetual, non-exclusive, transferable, royalty-free, fully-paid, worldwide right, and license to host, use, copy, reproduce, disclose, sell, resell, publish, broadcast, retitle, archive, store, cache, publicly perform, publicly display, reformat, translate, transmit, excerpt (in whole or in part), and distribute such Contributions (including, without limitation, your image and voice) for any purpose, commercial, advertising, or otherwise, and to prepare derivative works of, or incorporate into other works, such Contributions, and grant and authorize sublicenses of the foregoing. The use and distribution may occur in any media formats and through any media channels.

This license will apply to any form, media, or technology now known or hereafter developed, and includes our use of your name, company name, and franchise name, as applicable, and any of the trademarks, service marks, trade names, logos, and personal and commercial images you provide. You waive all moral rights in your Contributions, and you warrant that moral rights have not otherwise been asserted in your Contributions.

We do not assert any ownership over your Contributions. You retain full ownership of all of your Contributions and any intellectual property rights or other proprietary rights associated with your Contributions. We are not liable for any statements or representations in your Contributions provided by you in any area on the Site. You are solely responsible for your Contributions to the Site and you expressly agree to exonerate us from any and all responsibility and to refrain from any legal action against us regarding your Contributions.

We have the right, in our sole and absolute discretion, (1) to edit, redact, or otherwise change any Contributions; (2) to re-categorize any Contributions to place them in more appropriate locations on the Site; and (3) to pre-screen or delete any Contributions at any time and for any reason, without notice. We have no obligation to monitor your Contributions.

GUIDELINES FOR REVIEWS

We may provide you areas on the Site to leave reviews or ratings. When posting a review, you must comply with the following criteria: (1) you should have firsthand experience with the person/entity being reviewed; (2) your reviews should not contain offensive profanity, or abusive, racist, offensive, or hate language; (3) your reviews should not contain discriminatory references based on religion, race, gender, national origin, age, marital status, sexual orientation, or disability; (4) your reviews should not contain references to illegal activity; (5) you should not be affiliated with competitors if posting negative reviews; (6) you should not make any conclusions as to the legality of conduct; (7) you may not post any false or misleading statements;

and (8) you may not organize a campaign encouraging others to post reviews, whether positive or negative.

We may accept, reject, or remove reviews in our sole discretion. We have absolutely no obligation to screen reviews or to delete reviews, even if anyone considers reviews objectionable or inaccurate. Reviews are not endorsed by us, and do not necessarily represent our opinions or the views of any of our affiliates or partners. We do not assume liability for any review or for any claims, liabilities, or losses resulting from any review. By posting a review, you hereby grant to us a perpetual, non-exclusive, worldwide, royalty-free, fully-paid, assignable, and sublicensable right and license to reproduce, modify, translate, transmit by any means, display, perform, and/or distribute all content relating to reviews.

MOBILE APPLICATION LICENSE

Use License

If you access the Site via a mobile application, then we grant you a revocable, non-exclusive, non-transferable, limited right to install and use the mobile application on wireless electronic devices owned or controlled by you, and to access and use the mobile application on such devices strictly in accordance with the terms and conditions of this mobile application license contained in these Terms of Use. You shall not: (1) decompile, reverse engineer, disassemble, attempt to derive the source code of, or decrypt the application; (2) make any modification, adaptation, improvement, enhancement, translation, or derivative work from the application; (3) violate any applicable laws, rules, or regulations in connection with your access or use of the application; (4) remove, alter, or obscure any proprietary notice (including any notice of copyright or trademark) posted by us or the licensors of the application; (5) use the application for any revenue generating endeavor, commercial enterprise, or other purpose for which it is not designed or intended; (6) make the application available over a network or other environment permitting access or use by multiple devices or users at the same time; (7) use the application for creating a product, service, or software that is, directly or indirectly, competitive with or in any way a substitute for the application; (8) use the application to send automated queries to any website or to send any unsolicited commercial e-mail; or (9) use any proprietary information or any of our interfaces or our other intellectual property in the design, development, manufacture, licensing, or distribution of any applications, accessories, or devices for use with the application.

Apple and Android Devices

The following terms apply when you use a mobile application obtained from either the Apple Store or Google Play (each an “App Distributor”) to access the Site: (1) the license granted to you for our mobile application is limited to a non-transferable license to use the application on a device that utilizes the Apple iOS or Android operating systems, as applicable, and in accordance with the usage rules set forth in the applicable App Distributor’s terms of service; (2) we are responsible for providing any maintenance and support services with respect to the mobile application as specified in the terms and conditions of this mobile application license contained in these Terms of Use or as otherwise required under applicable law, and you acknowledge that each App Distributor has no obligation whatsoever to furnish any maintenance and support services with respect to the mobile application; (3) in the event of any failure of the mobile application to conform to any applicable warranty, you may notify the applicable App Distributor, and the

App Distributor, in accordance with its terms and policies, may refund the purchase price, if any, paid for the mobile application, and to the maximum extent permitted by applicable law, the App Distributor will have no other warranty obligation whatsoever with respect to the mobile application; (4) you represent and warrant that (i) you are not located in a country that is subject to a U.S. government embargo, or that has been designated by the U.S. government as a “terrorist supporting” country and (ii) you are not listed on any U.S. government list of prohibited or restricted parties; (5) you must comply with applicable third-party terms of agreement when using the mobile application, e.g., if you have a VoIP application, then you must not be in violation of their wireless data service agreement when using the mobile application; and (6) you acknowledge and agree that the App Distributors are third-party beneficiaries of the terms and conditions in this mobile application license contained in these Terms of Use, and that each App Distributor will have the right (and will be deemed to have accepted the right) to enforce the terms and conditions in this mobile application license contained in these Terms of Use against you as a third-party beneficiary thereof.

SUBMISSIONS

You acknowledge and agree that any questions, comments, suggestions, ideas, feedback, or other information regarding the Site ("Submissions") provided by you to us are non-confidential and shall become our sole property. We shall own exclusive rights, including all intellectual property rights, and shall be entitled to the unrestricted use and dissemination of these Submissions for any lawful purpose, commercial or otherwise, without acknowledgment or compensation to you. You hereby waive all moral rights to any such Submissions, and you hereby warrant that any such Submissions are original with you or that you have the right to submit such Submissions. You agree there shall be no recourse against us for any alleged or actual infringement or misappropriation of any proprietary right in your Submissions.

THIRD-PARTY WEBSITE AND CONTENT

The Site may contain (or you may be sent via the Site) links to other websites ("Third-Party Websites") as well as articles, photographs, text, graphics, pictures, designs, music, sound, video, information, applications, software, and other content or items belonging to or originating from third parties ("Third-Party Content"). Such Third-Party Websites and Third-Party Content are not investigated, monitored, or checked for accuracy, appropriateness, or completeness by us, and we are not responsible for any Third-Party Websites accessed through the Site or any Third-Party Content posted on, available through, or installed from the Site, including the content, accuracy, offensiveness, opinions, reliability, privacy practices, or other policies of or contained in the Third-Party Websites or the Third-Party Content. Inclusion of, linking to, or permitting the use or installation of any Third-Party Websites or any Third-Party Content does not imply approval or endorsement thereof by us. If you decide to leave the Site and access the Third-Party Websites or to use or install any Third-Party Content, you do so at your own risk, and you should be aware these Terms of Use no longer govern. You should review the applicable terms and policies, including privacy and data gathering practices, of any website to which you navigate from the Site or relating to any applications you use or install from the Site. Any purchases you make through Third-Party Websites will be through other websites and from other companies, and we take no responsibility whatsoever in relation to such purchases which are exclusively between you and the applicable third party. You agree and acknowledge that we do not endorse the products or services offered on Third-Party Websites and you shall hold us harmless from any harm caused by your purchase of such products or services. Additionally, you shall hold us harmless from any losses sustained by you or harm caused to you relating to or resulting in any way from any Third-Party Content or any contact with Third-Party Websites.

SITE MANAGEMENT

We reserve the right, but not the obligation, to: (1) monitor the Site for violations of these Terms of Use; (2) take appropriate legal action against anyone who, in our sole discretion, violates the law or these Terms of Use, including without limitation, reporting such user to law enforcement authorities; (3) in our sole discretion and without limitation, refuse, restrict access to, limit the availability of, or disable (to the extent technologically feasible) any of your Contributions or any portion thereof; (4) in our sole discretion and without limitation, notice, or liability, to remove from the Site or otherwise disable all files and content that are excessive in size or are in any way burdensome to our systems; and (5) otherwise manage the Site in a manner designed to protect our rights and property and to facilitate the proper functioning of the Site.

PRIVACY POLICY

We care about data privacy and security. Please review our Privacy Policy: http://www.example.com/privacy. By using the Site, you agree to be bound by our Privacy Policy, which is incorporated into these Terms of Use. Please be advised the Site is hosted in the United Kingdom. If you access the Site from any other region of the world with laws or other requirements governing personal data collection, use, or disclosure that differ from applicable laws in the United Kingdom, then through your continued use of the Site, you are transferring your data to the United Kingdom, and you agree to have your data transferred to and processed in the United Kingdom.

COPYRIGHT INFRINGEMENTS

We respect the intellectual property rights of others. If you believe that any material available on or through the Site infringes upon any copyright you own or control, please immediately notify us using the contact information provided below (a “Notification”). A copy of your Notification will be sent to the person who posted or stored the material addressed in the Notification. Please be advised that pursuant to applicable law you may be held liable for damages if you make material misrepresentations in a Notification. Thus, if you are not sure that material located on or linked to by the Site infringes your copyright, you should consider first contacting an attorney.

TERM AND TERMINATION

These Terms of Use shall remain in full force and effect while you use the Site. WITHOUT LIMITING ANY OTHER PROVISION OF THESE TERMS OF USE, WE RESERVE THE RIGHT TO, IN OUR SOLE DISCRETION AND WITHOUT NOTICE OR LIABILITY, DENY ACCESS TO AND USE OF THE SITE (INCLUDING BLOCKING CERTAIN IP ADDRESSES), TO ANY PERSON FOR ANY REASON OR FOR NO REASON, INCLUDING WITHOUT LIMITATION FOR BREACH OF ANY REPRESENTATION, WARRANTY, OR COVENANT CONTAINED IN THESE TERMS OF USE OR OF ANY APPLICABLE LAW OR REGULATION. WE MAY TERMINATE YOUR USE OR PARTICIPATION IN THE SITE

OR DELETE YOUR ACCOUNT AND ANY CONTENT OR INFORMATION THAT YOU POSTED AT ANY TIME, WITHOUT WARNING, IN OUR SOLE DISCRETION.

If we terminate or suspend your account for any reason, you are prohibited from registering and creating a new account under your name, a fake or borrowed name, or the name of any third party, even if you may be acting on behalf of the third party. In addition to terminating or suspending your account, we reserve the right to take appropriate legal action, including without limitation pursuing civil, criminal, and injunctive redress.

MODIFICATIONS AND INTERRUPTIONS

We reserve the right to change, modify, or remove the contents of the Site at any time or for any reason at our sole discretion without notice. However, we have no obligation to update any information on our Site. We also reserve the right to modify or discontinue all or part of the Site without notice at any time. We will not be liable to you or any third party for any modification, price change, suspension, or discontinuance of the Site.

We cannot guarantee the Site will be available at all times. We may experience hardware, software, or other problems or need to perform maintenance related to the Site, resulting in interruptions, delays, or errors. We reserve the right to change, revise, update, suspend, discontinue, or otherwise modify the Site at any time or for any reason without notice to you. You agree that we have no liability whatsoever for any loss, damage, or inconvenience caused by your inability to access or use the Site during any downtime or discontinuance of the Site. Nothing in these Terms of Use will be construed to obligate us to maintain and support the Site or to supply any corrections, updates, or releases in connection therewith.

GOVERNING LAW

These conditions are governed by and interpreted following the laws of the United Kingdom, and the use of the United Nations Convention of Contracts for the International Sale of Goods is expressly excluded. If your habitual residence is in the EU, and you are a consumer, you additionally possess the protection provided to you by obligatory provisions of the law of your country of residence. Test LLC and yourself both agree to submit to the non-exclusive jurisdiction of the courts of England, which means that you may make a claim to defend your consumer protection rights in regards to these Conditions of Use in the United Kingdom, or in the EU country in which you reside.

DISPUTE RESOLUTION

The European Commission provides an online dispute resolution platform, which you can access here: https://ec.europa.eu/consumers/odr. If you would like to bring this subject to our attention, please contact us.

CORRECTIONS

There may be information on the Site that contains typographical errors, inaccuracies, or omissions, including descriptions, pricing, availability, and various other information. We reserve the right to correct any errors, inaccuracies, or omissions and to change or update the information on the Site at any time, without prior notice.

DISCLAIMER

THE SITE IS PROVIDED ON AN AS-IS AND AS-AVAILABLE BASIS. YOU AGREE THAT YOUR USE OF THE SITE AND OUR SERVICES WILL BE AT YOUR SOLE RISK. TO THE FULLEST EXTENT PERMITTED BY LAW, WE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, IN CONNECTION WITH THE SITE AND YOUR USE THEREOF, INCLUDING, WITHOUT LIMITATION, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NON-INFRINGEMENT. WE MAKE NO WARRANTIES OR REPRESENTATIONS ABOUT THE ACCURACY OR COMPLETENESS OF THE SITE'S CONTENT OR THE CONTENT OF ANY WEBSITES LINKED TO THE SITE AND WE WILL ASSUME NO LIABILITY OR RESPONSIBILITY FOR ANY (1) ERRORS, MISTAKES, OR INACCURACIES OF CONTENT AND MATERIALS, (2) PERSONAL INJURY OR PROPERTY DAMAGE, OF ANY NATURE WHATSOEVER, RESULTING FROM YOUR ACCESS TO AND USE OF THE SITE, (3) ANY UNAUTHORIZED ACCESS TO OR USE OF OUR SECURE SERVERS AND/OR ANY AND ALL PERSONAL INFORMATION AND/OR FINANCIAL INFORMATION STORED THEREIN, (4) ANY INTERRUPTION OR CESSATION OF TRANSMISSION TO OR FROM THE SITE, (5) ANY BUGS, VIRUSES, TROJAN HORSES, OR THE LIKE WHICH MAY BE TRANSMITTED TO OR THROUGH THE SITE BY ANY THIRD PARTY, AND/OR (6) ANY ERRORS OR OMISSIONS IN ANY CONTENT AND MATERIALS OR FOR ANY LOSS OR DAMAGE OF ANY KIND INCURRED AS A RESULT OF THE USE OF ANY CONTENT POSTED, TRANSMITTED, OR OTHERWISE MADE AVAILABLE VIA THE SITE. WE DO NOT WARRANT, ENDORSE, GUARANTEE, OR ASSUME RESPONSIBILITY FOR ANY PRODUCT OR SERVICE ADVERTISED OR OFFERED BY A THIRD PARTY THROUGH THE SITE, ANY HYPERLINKED WEBSITE, OR ANY WEBSITE OR MOBILE APPLICATION FEATURED IN ANY BANNER OR OTHER ADVERTISING, AND WE WILL NOT BE A PARTY TO OR IN ANY WAY BE RESPONSIBLE FOR MONITORING ANY TRANSACTION BETWEEN YOU AND ANY THIRD-PARTY PROVIDERS OF PRODUCTS OR SERVICES. AS WITH THE PURCHASE OF A PRODUCT OR SERVICE THROUGH ANY MEDIUM OR IN ANY ENVIRONMENT, YOU SHOULD USE YOUR BEST JUDGMENT AND EXERCISE CAUTION WHERE APPROPRIATE.

LIMITATIONS OF LIABILITY

IN NO EVENT WILL WE OR OUR DIRECTORS, EMPLOYEES, OR AGENTS BE LIABLE TO YOU OR ANY THIRD PARTY FOR ANY DIRECT, INDIRECT, CONSEQUENTIAL, EXEMPLARY, INCIDENTAL, SPECIAL, OR PUNITIVE DAMAGES, INCLUDING LOST PROFIT, LOST REVENUE, LOSS OF DATA, OR OTHER DAMAGES ARISING FROM YOUR USE OF THE SITE, EVEN IF WE HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. NOTWITHSTANDING ANYTHING TO THE CONTRARY CONTAINED HEREIN, OUR LIABILITY TO YOU FOR ANY CAUSE WHATSOEVER AND REGARDLESS OF THE FORM OF THE ACTION, WILL AT ALL TIMES BE LIMITED TO THE AMOUNT PAID, IF ANY, BY YOU TO US DURING THE SIX (6) MONTH PERIOD PRIOR TO ANY CAUSE OF ACTION ARISING. CERTAIN US STATE LAWS AND INTERNATIONAL LAWS DO NOT ALLOW LIMITATIONS ON IMPLIED WARRANTIES OR THE EXCLUSION OR LIMITATION OF CERTAIN DAMAGES. IF THESE LAWS APPLY

TO YOU, SOME OR ALL OF THE ABOVE DISCLAIMERS OR LIMITATIONS MAY NOT APPLY TO YOU, AND YOU MAY HAVE ADDITIONAL RIGHTS.

INDEMNIFICATION

You agree to defend, indemnify, and hold us harmless, including our subsidiaries, affiliates, and all of our respective officers, agents, partners, and employees, from and against any loss, damage, liability, claim, or demand, including reasonable attorneys' fees and expenses, made by any third party due to or arising out of: (1) your Contributions; (2) use of the Site; (3) breach of these Terms of Use; (4) any breach of your representations and warranties set forth in these Terms of Use; (5) your violation of the rights of a third party, including but not limited to intellectual property rights; or (6) any overt harmful act toward any other user of the Site with whom you connected via the Site. Notwithstanding the foregoing, we reserve the right, at your expense, to assume the exclusive defense and control of any matter for which you are required to indemnify us, and you agree to cooperate, at your expense, with our defense of such claims. We will use reasonable efforts to notify you of any such claim, action, or proceeding which is subject to this indemnification upon becoming aware of it.

USER DATA

We will maintain certain data that you transmit to the Site for the purpose of managing the performance of the Site, as well as data relating to your use of the Site. Although we perform regular routine backups of data, you are solely responsible for all data that you transmit or that relates to any activity you have undertaken using the Site. You agree that we shall have no liability to you for any loss or corruption of any such data, and you hereby waive any right of action against us arising from any such loss or corruption of such data.

ELECTRONIC COMMUNICATIONS, TRANSACTIONS, AND SIGNATURES

Visiting the Site, sending us emails, and completing online forms constitute electronic communications. You consent to receive electronic communications, and you agree that all agreements, notices, disclosures, and other communications we provide to you electronically, via email and on the Site, satisfy any legal requirement that such communication be in writing. YOU HEREBY AGREE TO THE USE OF ELECTRONIC SIGNATURES, CONTRACTS, ORDERS, AND OTHER RECORDS, AND TO ELECTRONIC DELIVERY OF NOTICES, POLICIES, AND RECORDS OF TRANSACTIONS INITIATED OR COMPLETED BY US OR VIA THE SITE. You hereby waive any rights or requirements under any statutes, regulations, rules, ordinances, or other laws in any jurisdiction which require an original signature or delivery or retention of non-electronic records, or to payments or the granting of credits by any means other than electronic means.

MISCELLANEOUS

These Terms of Use and any policies or operating rules posted by us on the Site or in respect to the Site constitute the entire agreement and understanding between you and us. Our failure to exercise or enforce any right or provision of these Terms of Use shall not operate as a waiver of such right or provision. These Terms of Use operate to the fullest extent permissible by law. We may assign any or all of our rights and obligations to others at any time. We shall not be responsible or liable for any loss, damage, delay, or failure to act caused by any cause beyond our reasonable control. If any provision or part of a provision of these Terms of Use is determined to be unlawful, void, or unenforceable, that provision or part of the provision is deemed severable from these Terms of Use and does not affect the validity and enforceability of any remaining provisions. There is no joint venture, partnership, employment or agency relationship created between you and us as a result of these Terms of Use or use of the Site. You agree that these Terms of Use will not be construed against us by virtue of having drafted them. You hereby waive any and all defenses you may have based on the electronic form of these Terms of Use and the lack of signing by the parties hereto to execute these Terms of Use.

CONTACT US

In order to resolve a complaint regarding the Site or to receive further information regarding use of the Site, please contact us at:

Test LLC Test Test AA000AA United Kingdom Phone: 00 Fax: 00 test@example.com

These terms of use were created using Termly's Terms and Conditions Generator.

APPENDIX E

Taxonomy for Clause Topics

Topic	Subtopic	Description
age		Minimum age to order
applicability		Applicability of the T&C
applicableLaw		Applicable law
arbitration		Participation in arbitration
changes		Changes to the contract
codeOfConduct		Code of conduct
conclusionOfContract		Conclusion of contract
conclusionOfContract	binding	When the contract becomes binding
conclusionOfContract	changeOfOrder	Changes and adujstments of orders
conclusionOfContract	definition	Definition of terms used in the contract
conclusionOfContract	restrictions	Restrictions to orders (e.g. the amount of ordered goods)
conclusionOfContract	steps	Steps towards contract conclusion
conclusionOfContract	withdrawal	Withdrawal of the company from the contract
delivery		Delivery
delivery	brokenPackaging	Handling of broken packaging
delivery	costs	Costs of delivery
delivery	customs	Customs handling
delivery	destination	Destinations to which goods are delivered
delivery	methods	Delivery methods
delivery	partial	Partial delivery
delivery	time	Delivery duration
description		Product descriptions

Table E.1.: Taxonomy for T&C clauses

Topic	Subtopic	Description
disposal		Disposal regulations
intellectualProperty		Intellectual property
liability		Liability
party		Contracting party
payment		Payment
payment	fee	Payment fees
payment	late	Late payment
payment	loyalty	Loyalty schemes and reward programs
payment	methods	Accepted payment methods
payment	restraint	Restraint of payment
payment	vouchers	Vouchers
personalData		Personal data
personalData	cookies	Cookie regulations
personalData	duration	Storage duration for personal data
personalData	information	Which information is processed/stored
personalData	reason	Reason for storing / processing personal data
personalData	update	Updates of personal data
personalData	usage	Usage of personal data
placeOfJurisdiction		Place of jurisdiction
prices		Prices
prices	currency	Currency of prices
prices	vat	VAT
retentionOfTitle		Retention of title
severability		Severability clause
textStorage		Storage of contract text
warranty		Warranty
warranty	options	Options in case of warranty
warranty	period	Warranty period
withdrawal		Withdrawal
withdrawal	compensation	Compensation for product usage
withdrawal	effects	Effects of withdrawal
withdrawal	exclusion	Cases excluded from the right to withdraw
withdrawal	form	Allowed / disallowed to submit a withdrawal
withdrawal	model	Withdrawal model form
withdrawal	period	Time period for withdrawal
withdrawal	shippingCosts	Shipping costs for withdrawal
withdrawal	shippingMethod	Shipping method for withdrawal

Table E.1.: Taxonomy for T&C clauses

APPENDIX F

Detailed Results Clause Topic Classification

This appendix provides detailed results, broken down on topics (Section F.1) and subtopics (Section F.2), for the evaluation of the different topic classification approaches presented in Section 6.3.

F.1. Topics

F.1.1. Rule-based

Clause Topic	Support	TP	FP	FN	A	P	R	F1
age	38	9	0	29	0.237	1.000	0.237	0.383
applicability	253	134	77	119	0.406	0.635	0.530	0.578
applicableLaw	136	27	11	109	0.184	0.711	0.199	0.310
arbitration	155	109	8	46	0.669	0.932	0.703	0.801
changes	13	3	0	10	0.231	1.000	0.231	0.375
codeOfConduct	55	44	0	11	0.800	1.000	0.800	0.889
conclusionOfContract	797	384	152	413	0.405	0.716	0.482	0.576
delivery	836	471	130	365	0.488	0.784	0.563	0.656
description	86	0	0	86	0.000	0.000	0.000	0.000
disposal	51	39	8	12	0.661	0.830	0.765	0.796
intellectualProperty	45	5	1	40	0.109	0.833	0.111	0.196
language	124	50	79	74	0.246	0.388	0.403	0.395
liability	438	284	218	154	0.433	0.566	0.648	0.604
party	157	44	72	113	0.192	0.379	0.280	0.322
payment	897	763	162	134	0.720	0.825	0.851	0.838
personalData	213	89	9	124	0.401	0.908	0.418	0.572
placeOfJurisdiction	116	62	53	54	0.367	0.539	0.534	0.537
prices	158	111	303	47	0.241	0.268	0.703	0.388
retentionOfTitle	222	211	40	11	0.805	0.841	0.950	0.892
severability	42	13	8	29	0.260	0.619	0.310	0.413
textStorage	152	76	60	76	0.358	0.559	0.500	0.528
warranty	538	370	100	168	0.580	0.787	0.688	0.734
withdrawal	479	322	22	157	0.643	0.936	0.672	0.783
TOTAL	6001	3620	1513	2381	0.482	0.705	0.603	0.650

Table F.1.: Results of the rule-based clause topic classification in German using the paragraph title as input on clauses with paragraph title present

Clause Topic	**Support**	**TP**	**FP**	**FN**	**A**	**P**	**R**	**F1**
age	38	9	0	29	0.237	1.000	0.237	0.383
applicability	253	134	77	119	0.406	0.635	0.530	0.578
applicableLaw	137	27	11	110	0.182	0.711	0.197	0.309
arbitration	155	109	8	46	0.669	0.932	0.703	0.801
changes	13	3	0	10	0.231	1.000	0.231	0.375
codeOfConduct	55	44	0	11	0.800	1.000	0.800	0.889
conclusionOfContract	800	384	152	416	0.403	0.716	0.480	0.575
delivery	839	471	130	368	0.486	0.784	0.561	0.654
description	86	0	0	86	0.000	0.000	0.000	0.000
disposal	51	39	8	12	0.661	0.830	0.765	0.796
intellectualProperty	45	5	1	40	0.109	0.833	0.111	0.196
language	124	50	79	74	0.246	0.388	0.403	0.395
liability	439	284	218	155	0.432	0.566	0.647	0.604
party	157	44	72	113	0.192	0.379	0.280	0.322
payment	898	763	162	135	0.720	0.825	0.850	0.837
personalData	213	89	9	124	0.401	0.908	0.418	0.572
placeOfJurisdiction	117	62	53	55	0.365	0.539	0.530	0.534
prices	158	111	303	47	0.241	0.268	0.703	0.388
retentionOfTitle	222	211	40	11	0.805	0.841	0.950	0.892
severability	42	13	8	29	0.260	0.619	0.310	0.413
textStorage	152	76	60	76	0.358	0.559	0.500	0.528
warranty	540	370	100	170	0.578	0.787	0.685	0.733
withdrawal	484	322	22	162	0.636	0.936	0.665	0.778
TOTAL	6018	3620	1513	2398	0.481	0.705	0.602	0.649

Table F.2.: Results of the rule-based clause topic classification in German using the paragraph title as input on all clauses

Clause Topic	Support	TP	FP	FN	A	P	R	F1
age	2	0	0	2	0.000	0.000	0.000	0.000
applicability	14	3	0	11	0.214	1.000	0.214	0.353
applicableLaw	3	0	0	3	0.000	0.000	0.000	0.000
arbitration	10	9	0	1	0.900	1.000	0.900	0.947
changes	1	0	0	1	0.000	0.000	0.000	0.000
codeOfConduct	1	1	0	0	1.000	1.000	1.000	1.000
conclusionOfContract	29	12	0	17	0.414	1.000	0.414	0.585
delivery	21	10	5	11	0.385	0.667	0.476	0.556
description	1	0	0	1	0.000	0.000	0.000	0.000
disposal	5	5	0	0	1.000	1.000	1.000	1.000
intellectualProperty	4	1	0	3	0.250	1.000	0.250	0.400
language	2	2	0	0	1.000	1.000	1.000	1.000
liability	20	6	5	14	0.240	0.545	0.300	0.387
party	19	4	0	15	0.211	1.000	0.211	0.348
payment	66	60	16	6	0.732	0.789	0.909	0.845
personalData	50	19	1	31	0.373	0.950	0.380	0.543
placeOfJurisdiction	2	1	0	1	0.500	1.000	0.500	0.667
prices	2	1	0	1	0.500	1.000	0.500	0.667
retentionOfTitle	2	2	1	0	0.667	0.667	1.000	0.800
severability	0	0	0	0	0.000	0.000	0.000	0.000
textStorage	3	2	0	1	0.667	1.000	0.667	0.800
warranty	41	26	0	15	0.634	1.000	0.634	0.776
withdrawal	169	156	1	13	0.918	0.994	0.923	0.957
TOTAL	467	320	29	147	0.645	0.917	0.685	0.784

Table F.3.: Results of the rule-based clause topic classification in German using the clause title as input on clauses with clause title present

Clause Topic	**Support**	**TP**	**FP**	**FN**	**A**	**P**	**R**	**F1**
age	38	0	0	38	0.000	0.000	0.000	0.000
applicability	253	3	0	250	0.012	1.000	0.012	0.023
applicableLaw	137	0	0	137	0.000	0.000	0.000	0.000
arbitration	155	9	0	146	0.058	1.000	0.058	0.110
changes	13	0	0	13	0.000	0.000	0.000	0.000
codeOfConduct	55	1	0	54	0.018	1.000	0.018	0.036
conclusionOfContract	800	12	0	788	0.015	1.000	0.015	0.030
delivery	839	10	5	829	0.012	0.667	0.012	0.023
description	86	0	0	86	0.000	0.000	0.000	0.000
disposal	51	5	0	46	0.098	1.000	0.098	0.179
intellectualProperty	45	1	0	44	0.022	1.000	0.022	0.043
language	124	2	0	122	0.016	1.000	0.016	0.032
liability	439	6	5	433	0.014	0.545	0.014	0.027
party	157	4	0	153	0.025	1.000	0.025	0.050
payment	898	60	16	838	0.066	0.789	0.067	0.123
personalData	213	19	1	194	0.089	0.950	0.089	0.163
placeOfJurisdiction	117	1	0	116	0.009	1.000	0.009	0.017
prices	158	1	0	157	0.006	1.000	0.006	0.013
retentionOfTitle	222	2	1	220	0.009	0.667	0.009	0.018
severability	42	0	0	42	0.000	0.000	0.000	0.000
textStorage	152	2	0	150	0.013	1.000	0.013	0.026
warranty	540	26	0	514	0.048	1.000	0.048	0.092
withdrawal	484	156	1	328	0.322	0.994	0.322	0.487
TOTAL	6018	320	29	5698	0.053	0.917	0.053	0.101

Table F.4.: Results of the rule-based clause topic classification in German using the clause title as input on all clauses

Clause Topic	Support	TP	FP	FN	A	P	R	F1
age	38	9	0	29	0.237	1.000	0.237	0.383
applicability	253	136	77	117	0.412	0.638	0.538	0.584
applicableLaw	136	27	11	109	0.184	0.711	0.199	0.310
arbitration	155	116	8	39	0.712	0.935	0.748	0.832
changes	13	3	0	10	0.231	1.000	0.231	0.375
codeOfConduct	55	45	0	10	0.818	1.000	0.818	0.900
conclusionOfContract	797	396	152	401	0.417	0.723	0.497	0.589
delivery	836	474	135	362	0.488	0.778	0.567	0.656
description	86	0	0	86	0.000	0.000	0.000	0.000
disposal	51	43	8	8	0.729	0.843	0.843	0.843
intellectualProperty	45	6	1	39	0.130	0.857	0.133	0.231
language	124	52	79	72	0.256	0.397	0.419	0.408
liability	438	285	218	153	0.434	0.567	0.651	0.606
party	157	46	72	111	0.201	0.390	0.293	0.335
payment	897	766	172	131	0.717	0.817	0.854	0.835
personalData	213	94	10	119	0.422	0.904	0.441	0.593
placeOfJurisdiction	116	62	53	54	0.367	0.539	0.534	0.537
prices	158	112	303	46	0.243	0.270	0.709	0.391
retentionOfTitle	222	212	41	10	0.806	0.838	0.955	0.893
severability	42	13	8	29	0.260	0.619	0.310	0.413
textStorage	152	78	60	74	0.368	0.565	0.513	0.538
warranty	538	370	100	168	0.580	0.787	0.688	0.734
withdrawal	484	331	23	153	0.653	0.935	0.684	0.790
TOTAL	6006	3676	1531	2330	0.488	0.706	0.612	0.656

Table F.5.: Results of the rule-based clause topic classification in German using both titles (paragraph and clause) as input on clauses with paragraph or clause title present

Clause Topic	Support	TP	FP	FN	A	P	R	F1
age	38	9	0	29	0.237	1.000	0.237	0.383
applicability	253	136	77	117	0.412	0.638	0.538	0.584
applicableLaw	137	27	11	110	0.182	0.711	0.197	0.309
arbitration	155	116	8	39	0.712	0.935	0.748	0.832
changes	13	3	0	10	0.231	1.000	0.231	0.375
codeOfConduct	55	45	0	10	0.818	1.000	0.818	0.900
conclusionOfContract	800	396	152	404	0.416	0.723	0.495	0.588
delivery	839	474	135	365	0.487	0.778	0.565	0.655
description	86	0	0	86	0.000	0.000	0.000	0.000
disposal	51	43	8	8	0.729	0.843	0.843	0.843
intellectualProperty	45	6	1	39	0.130	0.857	0.133	0.231
language	124	52	79	72	0.256	0.397	0.419	0.408
liability	439	285	218	154	0.434	0.567	0.649	0.605
party	157	46	72	111	0.201	0.390	0.293	0.335
payment	898	766	172	132	0.716	0.817	0.853	0.834
personalData	213	94	10	119	0.422	0.904	0.441	0.593
placeOfJurisdiction	117	62	53	55	0.365	0.539	0.530	0.534
prices	158	112	303	46	0.243	0.270	0.709	0.391
retentionOfTitle	222	212	41	10	0.806	0.838	0.955	0.893
severability	42	13	8	29	0.260	0.619	0.310	0.413
textStorage	152	78	60	74	0.368	0.565	0.513	0.538
warranty	540	370	100	170	0.578	0.787	0.685	0.733
withdrawal	484	331	23	153	0.653	0.935	0.684	0.790
TOTAL	6018	3676	1531	2342	0.487	0.706	0.611	0.655

Table F.6.: Results of the rule-based clause topic classification in German using both titles (paragraph and clause) as input on all clauses

Clause Topic	**Support**	**TP**	**FP**	**FN**	**A**	**P**	**R**	**F1**
age	38	31	1	7	0.795	0.969	0.816	0.886
applicability	253	208	99	45	0.591	0.678	0.822	0.743
applicableLaw	137	83	7	54	0.576	0.922	0.606	0.731
arbitration	155	154	0	1	0.994	1.000	0.994	0.997
changes	13	9	8	4	0.429	0.529	0.692	0.600
codeOfConduct	55	55	1	0	0.982	0.982	1.000	0.991
conclusionOfContract	800	511	252	289	0.486	0.670	0.639	0.654
delivery	839	708	316	131	0.613	0.691	0.844	0.760
description	86	75	20	11	0.708	0.789	0.872	0.829
disposal	51	45	7	6	0.776	0.865	0.882	0.874
intellectualProperty	45	22	0	23	0.489	1.000	0.489	0.657
language	124	122	1	2	0.976	0.992	0.984	0.988
liability	439	310	43	129	0.643	0.878	0.706	0.783
party	157	112	160	45	0.353	0.412	0.713	0.522
payment	898	755	149	143	0.721	0.835	0.841	0.838
personalData	213	158	34	55	0.640	0.823	0.742	0.780
placeOfJurisdiction	117	105	0	12	0.897	1.000	0.897	0.946
prices	158	143	84	15	0.591	0.630	0.905	0.743
retentionOfTitle	222	188	3	34	0.836	0.984	0.847	0.910
severability	42	39	1	3	0.907	0.975	0.929	0.951
textStorage	152	126	7	26	0.792	0.947	0.829	0.884
warranty	540	408	67	132	0.672	0.859	0.756	0.804
withdrawal	484	417	167	67	0.641	0.714	0.862	0.781
TOTAL	6018	4784	1427	1234	0.643	0.770	0.795	0.782

Table F.7.: Results of the rule-based clause topic classification in German using the clause text as input on all clauses

Clause Topic	Support	TP	FP	FN	A	P	R	F1
age	38	31	1	7	0.795	0.969	0.816	0.886
applicability	253	208	107	45	0.578	0.660	0.822	0.732
applicableLaw	136	83	18	53	0.539	0.822	0.610	0.700
arbitration	155	154	8	1	0.945	0.951	0.994	0.972
changes	13	9	8	4	0.429	0.529	0.692	0.600
codeOfConduct	55	55	1	0	0.982	0.982	1.000	0.991
conclusionOfContract	797	525	269	272	0.492	0.661	0.659	0.660
delivery	836	723	380	113	0.595	0.655	0.865	0.746
description	86	75	20	11	0.708	0.789	0.872	0.829
disposal	51	45	8	6	0.763	0.849	0.882	0.865
intellectualProperty	45	22	1	23	0.478	0.957	0.489	0.647
language	124	122	79	2	0.601	0.607	0.984	0.751
liability	438	312	43	126	0.649	0.879	0.712	0.787
party	157	112	161	45	0.352	0.410	0.713	0.521
payment	897	779	191	118	0.716	0.803	0.868	0.834
personalData	213	166	37	47	0.664	0.818	0.779	0.798
placeOfJurisdiction	116	109	53	7	0.645	0.673	0.940	0.784
prices	158	144	92	14	0.576	0.610	0.911	0.731
retentionOfTitle	222	188	3	34	0.836	0.984	0.847	0.910
severability	42	39	1	3	0.907	0.975	0.929	0.951
textStorage	152	142	63	10	0.660	0.693	0.934	0.796
warranty	538	499	168	39	0.707	0.748	0.928	0.828
withdrawal	484	431	173	53	0.656	0.714	0.890	0.792
TOTAL	6006	4973	1885	1033	0.630	0.725	0.828	0.773

Table F.8.: Results of the rule-based clause topic classification in German using the clause text and both titles (paragraph and clause) as input on clauses with paragraph or clause title present

Clause Topic	Support	TP	FP	FN	A	P	R	F1
age	38	31	1	7	0.795	0.969	0.816	0.886
applicability	253	208	107	45	0.578	0.660	0.822	0.732
applicableLaw	137	83	18	54	0.535	0.822	0.606	0.697
arbitration	155	154	8	1	0.945	0.951	0.994	0.972
changes	13	9	8	4	0.429	0.529	0.692	0.600
codeOfConduct	55	55	1	0	0.982	0.982	1.000	0.991
conclusionOfContract	800	526	269	274	0.492	0.662	0.657	0.660
delivery	839	726	380	113	0.596	0.656	0.865	0.747
description	86	75	20	11	0.708	0.789	0.872	0.829
disposal	51	45	8	6	0.763	0.849	0.882	0.865
intellectualProperty	45	22	1	23	0.478	0.957	0.489	0.647
language	124	122	79	2	0.601	0.607	0.984	0.751
liability	439	313	43	126	0.649	0.879	0.713	0.787
party	157	112	161	45	0.352	0.410	0.713	0.521
payment	898	779	191	119	0.715	0.803	0.867	0.834
personalData	213	166	37	47	0.664	0.818	0.779	0.798
placeOfJurisdiction	117	110	53	7	0.647	0.675	0.940	0.786
prices	158	144	92	14	0.576	0.610	0.911	0.731
retentionOfTitle	222	188	3	34	0.836	0.984	0.847	0.910
severability	42	39	1	3	0.907	0.975	0.929	0.951
textStorage	152	142	63	10	0.660	0.693	0.934	0.796
warranty	540	501	168	39	0.708	0.749	0.928	0.829
withdrawal	484	431	174	53	0.655	0.712	0.890	0.792
TOTAL	6018	4981	1886	1037	0.630	0.725	0.828	0.773

Table F.9.: Results of the rule-based clause topic classification in German using the clause text and both titles (paragraph and clause) as input on all clauses

Clause Topic	**Support**	**TP**	**FP**	**FN**	**A**	**P**	**R**	**F1**
age	5	1	0	4	0.200	1.000	0.200	0.333
applicability	27	7	1	20	0.250	0.875	0.259	0.400
applicableLaw	22	15	3	7	0.600	0.833	0.682	0.750
arbitration	13	8	11	5	0.333	0.421	0.615	0.500
changes	10	1	0	9	0.100	1.000	0.100	0.182
codeOfConduct	1	1	0	0	1.000	1.000	1.000	1.000
conclusionOfContract	133	104	70	29	0.512	0.598	0.782	0.678
delivery	160	124	32	36	0.646	0.795	0.775	0.785
description	27	8	2	19	0.276	0.800	0.296	0.432
disposal	16	16	0	0	1.000	1.000	1.000	1.000
intellectualProperty	24	18	3	6	0.667	0.857	0.750	0.800
language	11	3	3	8	0.214	0.500	0.273	0.353
liability	136	86	18	50	0.558	0.827	0.632	0.717
party	18	4	0	14	0.222	1.000	0.222	0.364
payment	102	84	21	18	0.683	0.800	0.824	0.812
personalData	45	27	1	18	0.587	0.964	0.600	0.740
placeOfJurisdiction	18	15	4	3	0.682	0.789	0.833	0.811
prices	53	42	32	11	0.494	0.568	0.792	0.661
retentionOfTitle	13	7	5	6	0.389	0.583	0.538	0.560
severability	12	3	3	9	0.200	0.500	0.250	0.333
textStorage	10	2	2	8	0.167	0.500	0.200	0.286
warranty	23	14	13	9	0.389	0.519	0.609	0.560
withdrawal	202	184	17	18	0.840	0.915	0.911	0.913
TOTAL	1081	774	241	307	0.585	0.763	0.716	0.739

Table F.10.: Results of the rule-based clause topic classification in English using the paragraph title as input on clauses with paragraph title present

Clause Topic	Support	TP	FP	FN	A	P	R	F1
age	5	1	0	4	0.200	1.000	0.200	0.333
applicability	33	7	1	26	0.206	0.875	0.212	0.341
applicableLaw	23	15	3	8	0.577	0.833	0.652	0.732
arbitration	13	8	11	5	0.333	0.421	0.615	0.500
changes	12	1	0	11	0.083	1.000	0.083	0.154
codeOfConduct	1	1	0	0	1.000	1.000	1.000	1.000
conclusionOfContract	146	104	70	42	0.481	0.598	0.712	0.650
delivery	164	124	32	40	0.633	0.795	0.756	0.775
description	30	8	2	22	0.250	0.800	0.267	0.400
disposal	16	16	0	0	1.000	1.000	1.000	1.000
intellectualProperty	24	18	3	6	0.667	0.857	0.750	0.800
language	11	3	3	8	0.214	0.500	0.273	0.353
liability	139	86	18	53	0.548	0.827	0.619	0.708
party	21	4	0	17	0.190	1.000	0.190	0.320
payment	112	84	21	28	0.632	0.800	0.750	0.774
personalData	49	27	1	22	0.540	0.964	0.551	0.701
placeOfJurisdiction	19	15	4	4	0.652	0.789	0.789	0.789
prices	56	42	32	14	0.477	0.568	0.750	0.646
retentionOfTitle	13	7	5	6	0.389	0.583	0.538	0.560
severability	12	3	3	9	0.200	0.500	0.250	0.333
textStorage	11	2	2	9	0.154	0.500	0.182	0.267
warranty	25	14	13	11	0.368	0.519	0.560	0.538
withdrawal	203	184	17	19	0.836	0.915	0.906	0.911
TOTAL	1138	774	241	364	0.561	0.763	0.680	0.719

Table F.11.: Results of the rule-based clause topic classification in English using the paragraph title as input on all clauses

Clause Topic	**Support**	**TP**	**FP**	**FN**	**A**	**P**	**R**	**F1**
age	0	0	0	0	0.000	0.000	0.000	0.000
applicability	0	0	0	0	0.000	0.000	0.000	0.000
applicableLaw	0	0	0	0	0.000	0.000	0.000	0.000
arbitration	3	3	0	0	1.000	1.000	1.000	1.000
changes	0	0	0	0	0.000	0.000	0.000	0.000
codeOfConduct	0	0	0	0	0.000	0.000	0.000	0.000
conclusionOfContract	10	10	2	0	0.833	0.833	1.000	0.909
delivery	5	4	0	1	0.800	1.000	0.800	0.889
description	0	0	0	0	0.000	0.000	0.000	0.000
disposal	0	0	0	0	0.000	0.000	0.000	0.000
intellectualProperty	0	0	0	0	0.000	0.000	0.000	0.000
language	0	0	0	0	0.000	0.000	0.000	0.000
liability	1	0	0	1	0.000	0.000	0.000	0.000
party	0	0	0	0	0.000	0.000	0.000	0.000
payment	13	11	0	2	0.846	1.000	0.846	0.917
personalData	1	0	0	1	0.000	0.000	0.000	0.000
placeOfJurisdiction	0	0	0	0	0.000	0.000	0.000	0.000
prices	0	0	0	0	0.000	0.000	0.000	0.000
retentionOfTitle	0	0	0	0	0.000	0.000	0.000	0.000
severability	0	0	0	0	0.000	0.000	0.000	0.000
textStorage	0	0	0	0	0.000	0.000	0.000	0.000
warranty	2	1	0	1	0.500	1.000	0.500	0.667
withdrawal	33	33	0	0	1.000	1.000	1.000	1.000
TOTAL	68	62	2	6	0.886	0.969	0.912	0.939

Table F.12.: Results of the rule-based clause topic classification in English using the clause title as input on clauses with clause title present

Clause Topic	Support	TP	FP	FN	A	P	R	F1
age	5	0	0	5	0.000	0.000	0.000	0.000
applicability	33	0	0	33	0.000	0.000	0.000	0.000
applicableLaw	23	0	0	23	0.000	0.000	0.000	0.000
arbitration	13	3	0	10	0.231	1.000	0.231	0.375
changes	12	0	0	12	0.000	0.000	0.000	0.000
codeOfConduct	1	0	0	1	0.000	0.000	0.000	0.000
conclusionOfContract	146	10	2	136	0.068	0.833	0.068	0.127
delivery	164	4	0	160	0.024	1.000	0.024	0.048
description	30	0	0	30	0.000	0.000	0.000	0.000
disposal	16	0	0	16	0.000	0.000	0.000	0.000
intellectualProperty	24	0	0	24	0.000	0.000	0.000	0.000
language	11	0	0	11	0.000	0.000	0.000	0.000
liability	139	0	0	139	0.000	0.000	0.000	0.000
party	21	0	0	21	0.000	0.000	0.000	0.000
payment	112	11	0	101	0.098	1.000	0.098	0.179
personalData	49	0	0	49	0.000	0.000	0.000	0.000
placeOfJurisdiction	19	0	0	19	0.000	0.000	0.000	0.000
prices	56	0	0	56	0.000	0.000	0.000	0.000
retentionOfTitle	13	0	0	13	0.000	0.000	0.000	0.000
severability	12	0	0	12	0.000	0.000	0.000	0.000
textStorage	11	0	0	11	0.000	0.000	0.000	0.000
warranty	25	1	0	24	0.040	1.000	0.040	0.077
withdrawal	203	33	0	170	0.163	1.000	0.163	0.280
TOTAL	1138	62	2	1076	0.054	0.969	0.054	0.103

Table F.13.: Results of the rule-based clause topic classification in English using the clause title as input on all clauses

Clause Topic	**Support**	**TP**	**FP**	**FN**	**A**	**P**	**R**	**F1**
age	5	1	0	4	0.200	1.000	0.200	0.333
applicability	27	7	1	20	0.250	0.875	0.259	0.400
applicableLaw	22	15	3	7	0.600	0.833	0.682	0.750
arbitration	13	9	11	4	0.375	0.450	0.692	0.545
changes	10	1	0	9	0.100	1.000	0.100	0.182
codeOfConduct	1	1	0	0	1.000	1.000	1.000	1.000
conclusionOfContract	133	104	70	29	0.512	0.598	0.782	0.678
delivery	160	125	32	35	0.651	0.796	0.781	0.789
description	27	8	2	19	0.276	0.800	0.296	0.432
disposal	16	16	0	0	1.000	1.000	1.000	1.000
intellectualProperty	24	18	3	6	0.667	0.857	0.750	0.800
language	11	3	3	8	0.214	0.500	0.273	0.353
liability	136	86	18	50	0.558	0.827	0.632	0.717
party	18	4	0	14	0.222	1.000	0.222	0.364
payment	102	84	21	18	0.683	0.800	0.824	0.812
personalData	45	27	1	18	0.587	0.964	0.600	0.740
placeOfJurisdiction	18	15	4	3	0.682	0.789	0.833	0.811
prices	53	42	32	11	0.494	0.568	0.792	0.661
retentionOfTitle	13	7	5	6	0.389	0.583	0.538	0.560
severability	12	3	3	9	0.200	0.500	0.250	0.333
textStorage	10	2	2	8	0.167	0.500	0.200	0.286
warranty	23	14	13	9	0.389	0.519	0.609	0.560
withdrawal	202	189	17	13	0.863	0.917	0.936	0.926
TOTAL	1081	781	241	300	0.591	0.764	0.722	0.743

Table F.14.: Results of the rule-based clause topic classification in English using both titles (paragraph and clause) as input on clauses with paragraph or clause title present

Clause Topic	Support	TP	FP	FN	A	P	R	F1
age	5	1	0	4	0.200	1.000	0.200	0.333
applicability	33	7	1	26	0.206	0.875	0.212	0.341
applicableLaw	23	15	3	8	0.577	0.833	0.652	0.732
arbitration	13	9	11	4	0.375	0.450	0.692	0.545
changes	12	1	0	11	0.083	1.000	0.083	0.154
codeOfConduct	1	1	0	0	1.000	1.000	1.000	1.000
conclusionOfContract	146	104	70	42	0.481	0.598	0.712	0.650
delivery	164	125	32	39	0.638	0.796	0.762	0.779
description	30	8	2	22	0.250	0.800	0.267	0.400
disposal	16	16	0	0	1.000	1.000	1.000	1.000
intellectualProperty	24	18	3	6	0.667	0.857	0.750	0.800
language	11	3	3	8	0.214	0.500	0.273	0.353
liability	139	86	18	53	0.548	0.827	0.619	0.708
party	21	4	0	17	0.190	1.000	0.190	0.320
payment	112	84	21	28	0.632	0.800	0.750	0.774
personalData	49	27	1	22	0.540	0.964	0.551	0.701
placeOfJurisdiction	19	15	4	4	0.652	0.789	0.789	0.789
prices	56	42	32	14	0.477	0.568	0.750	0.646
retentionOfTitle	13	7	5	6	0.389	0.583	0.538	0.560
severability	12	3	3	9	0.200	0.500	0.250	0.333
textStorage	11	2	2	9	0.154	0.500	0.182	0.267
warranty	25	14	13	11	0.368	0.519	0.560	0.538
withdrawal	203	189	17	14	0.859	0.917	0.931	0.924
TOTAL	1138	781	241	357	0.566	0.764	0.686	0.723

Table F.15.: Results of the rule-based clause topic classification in English using both titles (paragraph and clause) as input on all clauses

Clause Topic	**Support**	**TP**	**FP**	**FN**	**A**	**P**	**R**	**F1**
age	5	1	0	4	0.200	1.000	0.200	0.333
applicability	33	17	28	16	0.279	0.378	0.515	0.436
applicableLaw	23	15	2	8	0.600	0.882	0.652	0.750
arbitration	13	12	0	1	0.923	1.000	0.923	0.960
changes	12	7	0	5	0.583	1.000	0.583	0.737
codeOfConduct	1	1	0	0	1.000	1.000	1.000	1.000
conclusionOfContract	146	77	59	69	0.376	0.566	0.527	0.546
delivery	164	144	142	20	0.471	0.503	0.878	0.640
description	30	16	8	14	0.421	0.667	0.533	0.593
disposal	16	15	3	1	0.789	0.833	0.938	0.882
intellectualProperty	24	13	2	11	0.500	0.867	0.542	0.667
language	11	8	0	3	0.727	1.000	0.727	0.842
liability	139	108	72	31	0.512	0.600	0.777	0.677
party	21	6	1	15	0.273	0.857	0.286	0.429
payment	112	75	71	37	0.410	0.514	0.670	0.581
personalData	49	34	24	15	0.466	0.586	0.694	0.636
placeOfJurisdiction	19	16	1	3	0.800	0.941	0.842	0.889
prices	56	48	70	8	0.381	0.407	0.857	0.552
retentionOfTitle	13	11	0	2	0.846	1.000	0.846	0.917
severability	12	9	0	3	0.750	1.000	0.750	0.857
textStorage	11	8	1	3	0.667	0.889	0.727	0.800
warranty	25	14	24	11	0.286	0.368	0.560	0.444
withdrawal	203	151	30	52	0.648	0.834	0.744	0.786
TOTAL	1138	806	538	332	0.481	0.600	0.708	0.649

Table F.16.: Results of the rule-based clause topic classification in English using the clause text as input on all clauses

Clause Topic	**Support**	**TP**	**FP**	**FN**	**A**	**P**	**R**	**F1**
age	5	1	0	4	0.200	1.000	0.200	0.333
applicability	27	16	29	11	0.286	0.356	0.593	0.444
applicableLaw	22	17	2	5	0.708	0.895	0.773	0.829
arbitration	13	12	0	1	0.923	1.000	0.923	0.960
changes	10	6	0	4	0.600	1.000	0.600	0.750
codeOfConduct	1	1	0	0	1.000	1.000	1.000	1.000
conclusionOfContract	133	80	62	53	0.410	0.563	0.602	0.582
delivery	160	151	164	9	0.466	0.479	0.944	0.636
description	27	18	9	9	0.500	0.667	0.667	0.667
disposal	16	16	3	0	0.842	0.842	1.000	0.914
intellectualProperty	24	21	5	3	0.724	0.808	0.875	0.840
language	11	10	3	1	0.714	0.769	0.909	0.833
liability	136	125	86	11	0.563	0.592	0.919	0.720
party	18	5	1	13	0.263	0.833	0.278	0.417
payment	102	81	68	21	0.476	0.544	0.794	0.645
personalData	45	39	23	6	0.574	0.629	0.867	0.729
placeOfJurisdiction	18	16	5	2	0.696	0.762	0.889	0.821
prices	53	49	96	4	0.329	0.338	0.925	0.495
retentionOfTitle	13	11	0	2	0.846	1.000	0.846	0.917
severability	12	10	3	2	0.667	0.769	0.833	0.800
textStorage	10	9	3	1	0.692	0.750	0.900	0.818
warranty	23	14	26	9	0.286	0.350	0.609	0.444
withdrawal	202	176	33	26	0.749	0.842	0.871	0.856
TOTAL	1081	884	621	197	0.519	0.587	0.818	0.684

Table F.17.: Results of the rule-based clause topic classification in English using the clause text and both titles (paragraph and clause) as input on clauses with paragraph or clause title present

Clause Topic	**Support**	**TP**	**FP**	**FN**	**A**	**P**	**R**	**F1**
age	5	1	0	4	0.200	1.000	0.200	0.333
applicability	33	20	29	13	0.323	0.408	0.606	0.488
applicableLaw	23	17	2	6	0.680	0.895	0.739	0.810
arbitration	13	12	0	1	0.923	1.000	0.923	0.960
changes	12	8	0	4	0.667	1.000	0.667	0.800
codeOfConduct	1	1	0	0	1.000	1.000	1.000	1.000
conclusionOfContract	146	85	66	61	0.401	0.563	0.582	0.572
delivery	164	153	170	11	0.458	0.474	0.933	0.628
description	30	20	10	10	0.500	0.667	0.667	0.667
disposal	16	16	3	0	0.842	0.842	1.000	0.914
intellectualProperty	24	21	5	3	0.724	0.808	0.875	0.840
language	11	10	3	1	0.714	0.769	0.909	0.833
liability	139	128	86	11	0.569	0.598	0.921	0.725
party	21	6	1	15	0.273	0.857	0.286	0.429
payment	112	85	71	27	0.464	0.545	0.759	0.634
personalData	49	40	25	9	0.541	0.615	0.816	0.702
placeOfJurisdiction	19	17	5	2	0.708	0.773	0.895	0.829
prices	56	52	98	4	0.338	0.347	0.929	0.505
retentionOfTitle	13	11	0	2	0.846	1.000	0.846	0.917
severability	12	10	3	2	0.667	0.769	0.833	0.800
textStorage	11	9	3	2	0.643	0.750	0.818	0.783
warranty	25	16	27	9	0.308	0.372	0.640	0.471
withdrawal	203	177	35	26	0.744	0.835	0.872	0.853
TOTAL	1138	915	642	223	0.514	0.588	0.804	0.679

Table F.18.: Results of the rule-based clause topic classification in English using the clause text and both titles (paragraph and clause) as input on all clauses

F.1.2. Logistic Regression

Clause Topic	Support	P	R	F1
age	8	1.00	1.00	0.86
applicability	51	0.94	0.94	0.75
applicableLaw	27	1.00	1.00	0.96
arbitration	31	1.00	1.00	0.98
changes	3	0.00	0.00	0.00
codeOfConduct	11	1.00	1.00	0.90
conclusionOfContract	160	0.89	0.89	0.81
delivery	168	0.93	0.93	0.86
description	17	0.80	0.80	0.59
disposal	10	1.00	1.00	0.89
intellectualProperty	9	1.00	1.00	0.71
language	25	1.00	1.00	0.94
liability	88	0.94	0.94	0.82
party	31	0.84	0.84	0.64
payment	180	0.99	0.99	0.92
personalData	43	0.89	0.89	0.81
placeOfJurisdiction	23	1.00	1.00	0.98
prices	32	1.00	1.00	0.90
retentionOfTitle	44	0.97	0.97	0.92
severability	8	1.00	1.00	0.93
textStorage	30	0.94	0.94	0.95
warranty	108	0.94	0.94	0.88
withdrawal	97	0.99	0.99	0.87
micro avg	1204	0.95	0.95	0.86
macro avg	1204	0.92	0.92	0.82
weighted avg	1204	0.94	0.94	0.86
samples avg	1204	0.85	0.85	0.83

Table F.19.: Results of the Logistic Regression clause topic classification in German using the clause text as input

Clause Topic	**Support**	**P**	**R**	**F1**
age	8	1.00	1.00	0.86
applicability	51	0.95	0.95	0.86
applicableLaw	27	1.00	1.00	0.96
arbitration	31	1.00	1.00	0.97
changes	3	0.00	0.00	0.00
codeOfConduct	11	1.00	1.00	0.95
conclusionOfContract	160	0.92	0.92	0.83
delivery	168	0.94	0.94	0.85
description	17	0.80	0.80	0.59
disposal	10	1.00	1.00	0.89
intellectualProperty	9	1.00	1.00	0.50
language	25	1.00	1.00	0.94
liability	88	0.88	0.88	0.80
party	31	0.86	0.86	0.69
payment	180	0.98	0.98	0.93
personalData	43	0.94	0.94	0.85
placeOfJurisdiction	23	1.00	1.00	0.95
prices	32	1.00	1.00	0.88
retentionOfTitle	44	0.98	0.98	0.94
severability	8	1.00	1.00	0.93
textStorage	30	0.93	0.93	0.93
warranty	108	0.93	0.93	0.88
withdrawal	97	0.96	0.96	0.88
micro avg	1204	0.95	0.95	0.87
macro avg	1204	0.92	0.92	0.82
weighted avg	1204	0.94	0.94	0.87
samples avg	1204	0.87	0.87	0.85

Table F.20.: Results of the Logistic Regression clause topic classification in German using the clause text and both titles (paragraph and clause) as input

Clause Topic	Support	P	R	F1
age	1	0.00	0.00	0.00
applicability	7	0.80	0.80	0.67
applicableLaw	5	0.83	0.83	0.91
arbitration	3	1.00	1.00	1.00
changes	2	1.00	1.00	1.00
codeOfConduct	0	0.00	0.00	0.00
conclusionOfContract	29	0.63	0.63	0.69
delivery	33	0.76	0.76	0.78
description	6	1.00	1.00	0.80
disposal	3	0.00	0.00	0.00
intellectualProperty	5	1.00	1.00	0.57
language	2	0.00	0.00	0.00
liability	28	1.00	1.00	0.81
party	4	1.00	1.00	0.67
payment	22	0.80	0.80	0.65
personalData	10	0.83	0.83	0.62
placeOfJurisdiction	4	1.00	1.00	1.00
prices	11	1.00	1.00	0.90
retentionOfTitle	3	1.00	1.00	0.80
severability	2	1.00	1.00	0.67
textStorage	2	1.00	1.00	0.67
warranty	5	1.00	1.00	0.75
withdrawal	41	0.97	0.97	0.88
micro avg	228	0.85	0.85	0.77
macro avg	228	0.77	0.77	0.64
weighted avg	228	0.85	0.85	0.75
samples avg	228	0.71	0.71	0.70

Table F.21.: Results of the Logistic Regression clause topic classification in English using the clause text as input

Clause Topic	**Support**	**P**	**R**	**F1**
age	1	0.00	0.00	0.00
applicability	7	0.80	0.80	0.67
applicableLaw	5	0.83	0.83	0.91
arbitration	3	1.00	1.00	1.00
changes	2	1.00	1.00	1.00
codeOfConduct	0	0.00	0.00	0.00
conclusionOfContract	29	0.71	0.71	0.76
delivery	33	0.82	0.82	0.82
description	6	1.00	1.00	0.80
disposal	3	0.00	0.00	0.00
intellectualProperty	5	1.00	1.00	0.57
language	2	0.00	0.00	0.00
liability	28	1.00	1.00	0.81
party	4	1.00	1.00	0.67
payment	22	0.88	0.88	0.74
personalData	10	1.00	1.00	0.67
placeOfJurisdiction	4	1.00	1.00	1.00
prices	11	0.83	0.83	0.87
retentionOfTitle	3	1.00	1.00	0.80
severability	2	1.00	1.00	0.67
textStorage	2	1.00	1.00	0.67
warranty	5	1.00	1.00	0.57
withdrawal	41	1.00	1.00	0.94
micro avg	228	0.88	0.88	0.80
macro avg	228	0.78	0.78	0.65
weighted avg	228	0.88	0.88	0.78
samples avg	228	0.75	0.75	0.74

Table F.22.: Results of the Logistic Regression clause topic classification in English using the clause text and both titles (paragraph and clause) as input

F.1.3. Random Forest

Clause Topic	Support	P	R	F1
age	8.0	1.000	1.000	0.769
applicability	51.0	1.000	1.000	0.756
applicableLaw	27.0	1.000	1.000	0.962
arbitration	31.0	1.000	1.000	1.000
changes	3.0	0.000	0.000	0.000
codeOfConduct	11.0	1.000	1.000	1.000
conclusionOfContract	160.0	0.950	0.950	0.819
delivery	168.0	0.967	0.967	0.818
description	17.0	0.667	0.667	0.552
disposal	10.0	1.000	1.000	0.824
intellectualProperty	9.0	0.000	0.000	0.000
language	25.0	1.000	1.000	0.913
liability	88.0	0.950	0.950	0.770
party	31.0	1.000	1.000	0.558
payment	180.0	0.973	0.973	0.881
personalData	43.0	1.000	1.000	0.677
placeOfJurisdiction	23.0	1.000	1.000	0.905
prices	32.0	1.000	1.000	0.745
retentionOfTitle	44.0	0.951	0.951	0.918
severability	8.0	1.000	1.000	0.857
textStorage	30.0	1.000	1.000	0.929
warranty	108.0	0.920	0.920	0.821
withdrawal	97.0	0.986	0.986	0.840
micro avg	1204.0	0.966	0.966	0.826
macro avg	1204.0	0.885	0.885	0.753
weighted avg	1204.0	0.956	0.956	0.817
samples avg	1204.0	0.820	0.820	0.791

Table F.23.: Results of the Random Forest clause topic classification in German

Clause Topic	Support	P	R	F1
age	1.0	0.000	0.000	0.000
applicability	7.0	0.500	0.500	0.222
applicableLaw	5.0	0.800	0.800	0.800
arbitration	3.0	1.000	1.000	1.000
changes	2.0	0.000	0.000	0.000
codeOfConduct	0.0	0.000	0.000	0.000
conclusionOfContract	29.0	0.704	0.704	0.679
delivery	33.0	0.867	0.867	0.825
description	6.0	1.000	1.000	0.286
disposal	3.0	0.000	0.000	0.000
intellectualProperty	5.0	1.000	1.000	0.571
language	2.0	0.000	0.000	0.000
liability	28.0	1.000	1.000	0.756
party	4.0	0.000	0.000	0.000
payment	22.0	0.889	0.889	0.516
personalData	10.0	1.000	1.000	0.667
placeOfJurisdiction	4.0	1.000	1.000	1.000
prices	11.0	0.875	0.875	0.737
retentionOfTitle	3.0	1.000	1.000	0.500
severability	2.0	0.000	0.000	0.000
textStorage	2.0	1.000	1.000	0.667
warranty	5.0	0.000	0.000	0.000
withdrawal	41.0	0.944	0.944	0.883
micro avg	228.0	0.875	0.875	0.700
macro avg	228.0	0.590	0.590	0.439
weighted avg	228.0	0.813	0.813	0.660
samples avg	228.0	0.599	0.599	0.592

Table F.24.: Results of the Random Forest clause topic classification in English

F.1.4. Multilayer Perceptron

Clause Topic	Support	P	R	F1
age	9.0	1.000	1.000	0.714
applicability	48.0	0.893	0.893	0.658
applicableLaw	28.0	0.923	0.923	0.889
arbitration	31.0	0.931	0.931	0.900
changes	3.0	0.000	0.000	0.000
codeOfConduct	12.0	1.000	1.000	0.857
conclusionOfContract	164.0	0.910	0.910	0.854
delivery	160.0	0.855	0.855	0.813
description	17.0	0.333	0.333	0.174
disposal	10.0	1.000	1.000	0.889
intellectualProperty	9.0	1.000	1.000	0.500
language	25.0	1.000	1.000	0.810
liability	96.0	0.805	0.805	0.717
party	30.0	0.778	0.778	0.583
payment	180.0	0.923	0.923	0.891
personalData	43.0	0.688	0.688	0.587
placeOfJurisdiction	24.0	1.000	1.000	0.829
prices	32.0	0.800	0.800	0.381
retentionOfTitle	45.0	1.000	1.000	0.977
severability	9.0	1.000	1.000	0.875
textStorage	29.0	0.931	0.931	0.931
warranty	106.0	0.922	0.922	0.847
withdrawal	100.0	0.884	0.884	0.862
micro avg	1210.0	0.891	0.891	0.810
macro avg	1210.0	0.851	0.851	0.719
weighted avg	1210.0	0.882	0.882	0.799
samples avg	1210.0	0.891	0.891	0.837

Table F.25.: Results of the Multilayer Perceptron clause topic classification in German using Tf-idf vectors as input

Clause Topic	Support	P	R	F1
age	1.0	0.000	0.000	0.000
applicability	6.0	0.714	0.714	0.769
applicableLaw	4.0	0.250	0.250	0.250
arbitration	3.0	0.750	0.750	0.857
changes	2.0	1.000	1.000	0.667
codeOfConduct	0.0	0.000	0.000	0.000
conclusionOfContract	27.0	0.808	0.808	0.792
delivery	33.0	0.839	0.839	0.813
description	5.0	0.500	0.500	0.545
disposal	3.0	0.500	0.500	0.400
intellectualProperty	5.0	0.833	0.833	0.909
language	3.0	1.000	1.000	0.500
liability	29.0	0.821	0.821	0.807
party	4.0	0.500	0.500	0.333
payment	22.0	0.875	0.875	0.737
personalData	9.0	0.750	0.750	0.706
placeOfJurisdiction	4.0	1.000	1.000	0.857
prices	10.0	0.727	0.727	0.762
retentionOfTitle	3.0	1.000	1.000	0.800
severability	2.0	1.000	1.000	0.667
textStorage	3.0	0.500	0.500	0.400
warranty	5.0	0.750	0.750	0.667
withdrawal	41.0	0.902	0.902	0.902
micro avg	224.0	0.802	0.802	0.770
macro avg	224.0	0.697	0.697	0.615
weighted avg	224.0	0.806	0.806	0.765
samples avg	224.0	0.802	0.802	0.778

Table F.26.: Results of the Multilayer Perceptron clause topic classification in English using Word2Vec vectors as input

F.1.5. Convolutional Neural Network

Clause Topic	Support	P	R	F1
age	9.0	1.000	1.000	0.500
applicability	48.0	0.939	0.939	0.765
applicableLaw	28.0	1.000	1.000	0.880
arbitration	31.0	1.000	1.000	0.893
changes	3.0	0.000	0.000	0.000
codeOfConduct	11.0	0.900	0.900	0.857
conclusionOfContract	167.0	0.919	0.919	0.825
delivery	161.0	0.833	0.833	0.820
description	17.0	0.556	0.556	0.385
disposal	10.0	0.750	0.750	0.429
intellectualProperty	9.0	0.000	0.000	0.000
language	25.0	1.000	1.000	0.837
liability	96.0	0.892	0.892	0.720
party	30.0	0.929	0.929	0.591
payment	180.0	0.942	0.942	0.875
personalData	42.0	0.952	0.952	0.635
placeOfJurisdiction	24.0	1.000	1.000	0.829
prices	32.0	1.000	1.000	0.439
retentionOfTitle	47.0	1.000	1.000	0.881
severability	9.0	1.000	1.000	0.714
textStorage	29.0	1.000	1.000	0.906
warranty	105.0	0.885	0.885	0.802
withdrawal	101.0	0.987	0.987	0.847
micro avg	1214.0	0.920	0.920	0.797
macro avg	1214.0	0.847	0.847	0.671
weighted avg	1214.0	0.914	0.914	0.785
samples avg	1214.0	0.845	0.845	0.796

Table F.27.: Results of the CNN clause topic classification in German using GloVe vectors as input

Clause Topic	**Support**	**P**	**R**	**F1**
age	1.0	0.000	0.000	0.000
applicability	7.0	0.500	0.500	0.222
applicableLaw	5.0	1.000	1.000	1.000
arbitration	3.0	1.000	1.000	0.800
changes	2.0	1.000	1.000	1.000
codeOfConduct	0.0	0.000	0.000	0.000
conclusionOfContract	29.0	0.700	0.700	0.712
delivery	33.0	0.920	0.920	0.793
description	6.0	1.000	1.000	0.667
disposal	3.0	0.000	0.000	0.000
intellectualProperty	5.0	1.000	1.000	0.750
language	2.0	0.000	0.000	0.000
liability	28.0	0.957	0.957	0.863
party	4.0	0.500	0.500	0.333
payment	22.0	0.842	0.842	0.780
personalData	10.0	0.700	0.700	0.700
placeOfJurisdiction	4.0	1.000	1.000	0.400
prices	11.0	0.750	0.750	0.632
retentionOfTitle	3.0	1.000	1.000	0.800
severability	2.0	0.000	0.000	0.000
textStorage	2.0	1.000	1.000	0.667
warranty	5.0	0.000	0.000	0.000
withdrawal	41.0	0.923	0.923	0.900
micro avg	228.0	0.854	0.854	0.749
macro avg	228.0	0.643	0.643	0.523
weighted avg	228.0	0.809	0.809	0.721
samples avg	228.0	0.717	0.717	0.693

Table F.28.: Results of the CNN clause topic classification in English using GloVe vectors as input

F.1.6. Recurrent Neural Network

Clause Topic	Support	P	R	F1
age	14.0	1.000	1.000	0.526
applicability	49.0	0.861	0.861	0.729
applicableLaw	33.0	1.000	1.000	0.969
arbitration	26.0	1.000	1.000	0.960
changes	3.0	0.000	0.000	0.000
codeOfConduct	4.0	1.000	1.000	1.000
conclusionOfContract	162.0	0.871	0.871	0.808
delivery	163.0	0.888	0.888	0.771
description	9.0	0.000	0.000	0.000
disposal	9.0	0.857	0.857	0.750
intellectualProperty	12.0	0.833	0.833	0.556
language	20.0	1.000	1.000	0.750
liability	90.0	0.833	0.833	0.774
party	38.0	0.938	0.938	0.556
payment	182.0	0.936	0.936	0.864
personalData	54.0	0.842	0.842	0.696
placeOfJurisdiction	19.0	1.000	1.000	0.914
prices	33.0	0.950	0.950	0.717
retentionOfTitle	37.0	0.946	0.946	0.946
severability	11.0	1.000	1.000	0.842
textStorage	37.0	1.000	1.000	0.912
warranty	112.0	0.900	0.900	0.802
withdrawal	86.0	0.837	0.837	0.837
micro avg	1203.0	0.901	0.901	0.803
macro avg	1203.0	0.848	0.848	0.725
weighted avg	1203.0	0.894	0.894	0.794
samples avg	1203.0	0.875	0.875	0.823

Table F.29.: Results of the LSTM clause topic classification in German using domain specific embeddings vectors as input

Clause Topic	**Support**	**P**	**R**	**F1**
age	1.0	0.000	0.000	0.000
applicability	7.0	0.714	0.714	0.714
applicableLaw	5.0	0.000	0.000	0.000
arbitration	3.0	1.000	1.000	1.000
changes	2.0	1.000	1.000	1.000
codeOfConduct	0.0	0.000	0.000	0.000
conclusionOfContract	29.0	0.786	0.786	0.772
delivery	33.0	0.929	0.929	0.852
description	6.0	1.000	1.000	0.500
disposal	3.0	1.000	1.000	0.500
intellectualProperty	5.0	0.833	0.833	0.909
language	2.0	0.000	0.000	0.000
liability	28.0	0.870	0.870	0.784
party	4.0	1.000	1.000	0.667
payment	22.0	0.895	0.895	0.829
personalData	10.0	0.700	0.700	0.700
placeOfJurisdiction	4.0	0.667	0.667	0.800
prices	11.0	0.750	0.750	0.783
retentionOfTitle	3.0	0.500	0.500	0.400
severability	2.0	0.333	0.333	0.400
textStorage	2.0	0.333	0.333	0.400
warranty	5.0	0.429	0.429	0.500
withdrawal	41.0	0.804	0.804	0.851
micro avg	228.0	0.800	0.800	0.767
macro avg	228.0	0.632	0.632	0.581
weighted avg	228.0	0.792	0.792	0.752
samples avg	228.0	0.792	0.792	0.769

Table F.30.: Results of the LSTM clause topic classification in English using GloVe vectors as input

F.1.7. BERT

Clause Topic	Support	P	R	F1
age	9.0	1.000	1.000	0.941
applicability	48.0	0.927	0.927	0.854
applicableLaw	28.0	1.000	1.000	1.000
arbitration	31.0	0.969	0.969	0.984
changes	3.0	1.000	1.000	0.500
codeOfConduct	11.0	1.000	1.000	0.952
conclusionOfContract	167.0	0.903	0.903	0.898
delivery	161.0	0.917	0.917	0.902
description	17.0	0.800	0.800	0.750
disposal	10.0	0.875	0.875	0.778
intellectualProperty	9.0	0.875	0.875	0.824
language	25.0	1.000	1.000	0.980
liability	96.0	0.907	0.907	0.857
party	30.0	0.870	0.870	0.755
payment	180.0	0.939	0.939	0.942
personalData	42.0	0.778	0.778	0.805
placeOfJurisdiction	24.0	1.000	1.000	0.957
prices	32.0	0.938	0.938	0.938
retentionOfTitle	47.0	0.902	0.902	0.939
severability	9.0	1.000	1.000	0.875
textStorage	29.0	0.933	0.933	0.949
warranty	105.0	0.934	0.934	0.938
withdrawal	101.0	0.989	0.989	0.932
micro avg	1214.0	0.926	0.926	0.908
macro avg	1214.0	0.933	0.933	0.880
weighted avg	1214.0	0.926	0.926	0.907
samples avg	1214.0	0.939	0.939	0.923

Table F.31.: Results of the BERT clause topic classification in German

Clause Topic	Support	P	R	F1
age	1.0	0.000	0.000	0.000
applicability	7.0	0.714	0.714	0.714
applicableLaw	5.0	0.833	0.833	0.909
arbitration	3.0	1.000	1.000	1.000
changes	2.0	1.000	1.000	0.667
codeOfConduct	0.0	0.000	0.000	0.000
conclusionOfContract	29.0	0.730	0.730	0.818
delivery	33.0	0.969	0.969	0.954
description	6.0	1.000	1.000	0.667
disposal	3.0	1.000	1.000	1.000
intellectualProperty	5.0	0.750	0.750	0.667
language	2.0	0.000	0.000	0.000
liability	28.0	0.920	0.920	0.868
party	4.0	1.000	1.000	0.857
payment	22.0	0.905	0.905	0.884
personalData	10.0	1.000	1.000	0.667
placeOfJurisdiction	4.0	1.000	1.000	1.000
prices	11.0	0.818	0.818	0.818
retentionOfTitle	3.0	1.000	1.000	0.800
severability	2.0	1.000	1.000	0.667
textStorage	2.0	0.000	0.000	0.000
warranty	5.0	1.000	1.000	0.571
withdrawal	41.0	0.949	0.949	0.925
micro avg	228.0	0.890	0.890	0.851
macro avg	228.0	0.765	0.765	0.672
weighted avg	228.0	0.884	0.884	0.837
samples avg	228.0	0.833	0.833	0.823

Table F.32.: Results of the BERT clause topic classification in English

F.2. Subtopics

F.2.1. Rule-based

Clause Subtopic	Support	TP	FP	FN	A	P	R	F1
con.OfContract::binding	328	192	26	136	0.542	0.881	0.585	0.703
con.OfContract::changeOfOrder	58	43	28	15	0.5	0.606	0.741	0.667
con.OfContract::definition	103	93	36	10	0.669	0.721	0.903	0.802
con.OfContract::restrictions	42	19	15	23	0.333	0.559	0.452	0.5
con.OfContract::steps	256	129	81	127	0.383	0.614	0.504	0.554
con.OfContract::withdrawal	95	44	16	51	0.396	0.733	0.463	0.568
delivery:brokenPackaging	134	99	10	35	0.688	0.908	0.739	0.815
delivery:costs	247	70	10	177	0.272	0.875	0.283	0.428
delivery:customs	43	36	7	7	0.72	0.837	0.837	0.837
delivery:destination	96	34	11	62	0.318	0.756	0.354	0.482
delivery:methods	160	81	38	79	0.409	0.681	0.506	0.581
delivery:partial	32	24	5	8	0.649	0.828	0.75	0.787
delivery:time	143	84	35	59	0.472	0.706	0.587	0.641
payment:fee	50	25	45	25	0.263	0.357	0.5	0.417
payment:late	48	36	29	12	0.468	0.554	0.75	0.637
payment:methods	435	248	123	187	0.444	0.668	0.57	0.615
payment:loyalty	7	1	0	6	0.143	1.0	0.143	0.25
payment:restraint	46	45	10	1	0.804	0.818	0.978	0.891
payment:vouchers	301	287	9	14	0.926	0.97	0.953	0.961
personalData:cookies	6	6	2	0	0.75	0.75	1.0	0.857
personalData:duration	8	1	0	7	0.125	1.0	0.125	0.222
personalData:information	48	3	1	45	0.061	0.75	0.062	0.115
personalData:reason	50	22	18	28	0.324	0.55	0.44	0.489
personalData:update	7	0	1	7	0.0	0.0	0.0	0
personalData:usage	57	3	107	54	0.018	0.027	0.053	0.036
prices:currency	17	13	12	4	0.448	0.52	0.765	0.619
prices:vat	119	84	5	35	0.677	0.944	0.706	0.808
warranty:options	69	44	27	25	0.458	0.62	0.638	0.629
warranty:period	155	96	22	59	0.542	0.814	0.619	0.703
withdrawal:compensation	94	70	5	24	0.707	0.933	0.745	0.828
withdrawal:effects	97	4	3	93	0.04	0.571	0.041	0.077
withdrawal:exclusion	100	53	4	47	0.51	0.93	0.53	0.675
withdrawal:form	131	82	0	49	0.626	1.0	0.626	0.77
withdrawal:model	41	0	0	41	0.0	0	0.0	0
withdrawal:period	126	48	2	78	0.375	0.96	0.381	0.545
withdrawal:shippingCosts	118	90	20	28	0.652	0.818	0.763	0.789
withdrawal:shippingMethod	74	5	1	69	0.067	0.833	0.068	0.125
TOTAL	3941	2214	764	1727	0.471	0.743	0.562	0.64

Table F.33.: Results of the rule-based clause subtopic classification in German using the clause text as input

Clause Subtopic	Support	TP	FP	FN	A	P	R	F1
con.OfContract::binding	39	26	28	13	0.388	0.481	0.667	0.559
con.OfContract::changeOfOrder	6	1	6	5	0.083	0.143	0.167	0.154
con.OfContract::definition	4	3	0	1	0.75	1.0	0.75	0.857
con.OfContract::restrictions	7	2	0	5	0.286	1.0	0.286	0.444
con.OfContract::steps	58	29	39	29	0.299	0.426	0.5	0.46
con.OfContract::withdrawal	20	12	6	8	0.462	0.667	0.6	0.632
delivery:brokenPackaging	10	3	5	7	0.2	0.375	0.3	0.333
delivery:costs	57	43	29	14	0.5	0.597	0.754	0.667
delivery:customs	6	3	2	3	0.375	0.6	0.5	0.545
delivery:destination	16	5	10	11	0.192	0.333	0.312	0.323
delivery:methods	17	7	165	10	0.038	0.041	0.412	0.074
delivery:partial	5	3	1	2	0.5	0.75	0.6	0.667
delivery:time	41	17	33	24	0.23	0.34	0.415	0.374
payment:fee	3	2	6	1	0.222	0.25	0.667	0.364
payment:late	1	1	1	0	0.5	0.5	1.0	0.667
payment:methods	53	48	47	5	0.48	0.505	0.906	0.649
payment:loyalty	22	9	6	13	0.321	0.6	0.409	0.486
payment:restraint	1	1	3	0	0.25	0.25	1.0	0.4
payment:vouchers	14	5	6	9	0.25	0.455	0.357	0.4
personalData:cookies	3	3	1	0	0.75	0.75	1.0	0.857
personalData:duration	1	1	0	0	1.0	1.0	1.0	1.0
personalData:information	12	2	0	10	0.167	1.0	0.167	0.286
personalData:reason	11	0	0	11	0.0	0	0.0	0
personalData:update	4	3	3	1	0.429	0.5	0.75	0.6
personalData:usage	16	8	9	8	0.32	0.471	0.5	0.485
prices:currency	13	9	18	4	0.29	0.333	0.692	0.45
prices:vat	24	9	0	15	0.375	1.0	0.375	0.545
warranty:options	5	2	4	3	0.222	0.333	0.4	0.364
warranty:period	10	4	1	6	0.364	0.8	0.4	0.533
withdrawal:compensation	27	11	6	16	0.333	0.647	0.407	0.5
withdrawal:effects	12	8	61	4	0.11	0.116	0.667	0.198
withdrawal:exclusion	27	5	0	22	0.185	1.0	0.185	0.312
withdrawal:form	37	6	4	31	0.146	0.6	0.162	0.255
withdrawal:model	2	1	2	1	0.25	0.333	0.5	0.4
withdrawal:period	40	18	16	22	0.321	0.529	0.45	0.486
withdrawal:shippingCosts	43	25	14	18	0.439	0.641	0.581	0.61
withdrawal:shippingMethod	13	1	0	12	0.077	1.0	0.077	0.143
TOTAL	680	336	532	344	0.277	0.387	0.494	0.434

Table F.34.: Results of the rule-based clause subtopic classification in English using the clause text as input

F.2.2. Logistic Regression

Clause Subtopic	Support	P	R	F1
	441	0.91	0.91	0.90
conclusionOfContract:binding	66	0.95	0.95	0.88
conclusionOfContract:changeOfOrder	12	0.75	0.75	0.60
conclusionOfContract:definition	21	1.00	1.00	0.95
conclusionOfContract:restrictions	8	1.00	1.00	0.22
conclusionOfContract:steps	51	0.74	0.74	0.63
conclusionOfContract:withdrawal	19	0.82	0.82	0.60
delivery:brokenPackaging	27	1.00	1.00	0.88
delivery:costs	49	0.94	0.94	0.76
delivery:customs	9	1.00	1.00	0.88
delivery:destination	19	0.78	0.78	0.50
delivery:methods	32	0.88	0.88	0.61
delivery:partial	6	1.00	1.00	0.50
delivery:time	29	0.78	0.78	0.60
payment:fee	10	1.00	1.00	0.57
payment:late	10	1.00	1.00	0.57
payment:methods	1	1.00	1.00	1.00
payment:loyalty	87	0.95	0.95	0.88
payment:restraint	9	1.00	1.00	0.94
payment:vouchers	60	1.00	1.00	0.92
personalData:cookies	1	0.00	0.00	0.00
personalData:duration	2	1.00	1.00	0.67
personalData:information	10	1.00	1.00	0.46
personalData:reason	10	0.80	0.80	0.53
personalData:update	1	0.00	0.00	0.00
personalData:usage	11	0.86	0.86	0.67
prices:currency	4	0.00	0.00	0.00
prices:vat	24	0.96	0.96	0.94
warranty:options	14	1.00	1.00	0.83
warranty:period	31	0.90	0.90	0.87
withdrawal:compensation	19	0.85	0.85	0.87
withdrawal:effects	20	0.95	0.95	0.92
withdrawal:exclusion	20	0.94	0.94	0.86
withdrawal:form	28	0.89	0.89	0.87
withdrawal:model	8	1.00	1.00	1.00
withdrawal:period	25	0.92	0.92	0.90
withdrawal:shippingCosts	24	0.79	0.79	0.79
withdrawal:shippingMethod	15	0.82	0.82	0.69
micro avg	1233	0.91	0.91	0.83
macro avg	1233	0.85	0.85	0.69
weighted avg	1233	0.90	0.90	0.82
samples avg	1233	0.81	0.81	0.80

Table F.35.: Results of the Logistic Regression clause subtopic classification

Clause Subtopic	Support	P	R	F1
	441	0.92	0.92	0.91
conclusionOfContract:binding	66	0.95	0.95	0.90
conclusionOfContract:changeOfOrder	12	0.75	0.75	0.60
conclusionOfContract:definition	21	1.00	1.00	0.95
conclusionOfContract:restrictions	8	1.00	1.00	0.22
conclusionOfContract:steps	51	0.76	0.76	0.64
conclusionOfContract:withdrawal	19	0.88	0.88	0.52
delivery:brokenPackaging	27	1.00	1.00	0.90
delivery:costs	49	0.90	0.90	0.67
delivery:customs	9	1.00	1.00	0.88
delivery:destination	19	0.83	0.83	0.65
delivery:methods	32	0.95	0.95	0.73
delivery:partial	6	1.00	1.00	0.29
delivery:time	29	0.81	0.81	0.68
payment:fee	10	1.00	1.00	0.46
payment:late	10	1.00	1.00	0.57
payment:methods	1	1.00	1.00	1.00
payment:loyalty	87	0.94	0.94	0.88
payment:restraint	9	1.00	1.00	0.94
payment:vouchers	60	1.00	1.00	0.90
personalData:cookies	1	0.00	0.00	0.00
personalData:duration	2	1.00	1.00	0.67
personalData:information	10	1.00	1.00	0.46
personalData:reason	10	0.80	0.80	0.53
personalData:update	1	0.00	0.00	0.00
personalData:usage	11	0.83	0.83	0.59
prices:currency	4	0.00	0.00	0.00
prices:vat	24	1.00	1.00	0.96
warranty:options	14	1.00	1.00	0.83
warranty:period	31	0.88	0.88	0.89
withdrawal:compensation	19	0.89	0.89	0.89
withdrawal:effects	20	0.95	0.95	0.92
withdrawal:exclusion	20	0.89	0.89	0.87
withdrawal:form	28	0.92	0.92	0.89
withdrawal:model	8	1.00	1.00	0.93
withdrawal:period	25	0.88	0.88	0.88
withdrawal:shippingCosts	24	0.78	0.78	0.77
withdrawal:shippingMethod	15	0.75	0.75	0.67
micro avg	1233	0.91	0.91	0.84
macro avg	1233	0.85	0.85	0.68
weighted avg	1233	0.91	0.91	0.83
samples avg	1233	0.82	0.82	0.81

Table F.36.: Results of the Logistic Regression clause subtopic classification in German using the clause text and both titles (paragraph and clause) as input

Clause Subtopic	Support	P	R	F1
	101	0.79	0.79	0.79
conclusionOfContract:binding	8	0.71	0.71	0.67
conclusionOfContract:changeOfOrder	1	1.00	1.00	1.00
conclusionOfContract:definition	1	0.00	0.00	0.00
conclusionOfContract:restrictions	1	0.00	0.00	0.00
conclusionOfContract:steps	12	0.78	0.78	0.67
conclusionOfContract:withdrawal	4	1.00	1.00	0.67
delivery:brokenPackaging	2	0.00	0.00	0.00
delivery:costs	11	0.88	0.88	0.74
delivery:customs	1	0.00	0.00	0.00
delivery:destination	3	1.00	1.00	0.50
delivery:methods	3	1.00	1.00	0.80
delivery:partial	1	0.00	0.00	0.00
delivery:time	8	1.00	1.00	0.22
payment:fee	1	0.00	0.00	0.00
payment:late	0	0.00	0.00	0.00
payment:methods	4	0.00	0.00	0.00
payment:loyalty	11	0.71	0.71	0.56
payment:restraint	0	0.00	0.00	0.00
payment:vouchers	3	0.00	0.00	0.00
personalData:cookies	1	0.00	0.00	0.00
personalData:duration	0	0.00	0.00	0.00
personalData:information	2	1.00	1.00	0.67
personalData:reason	2	1.00	1.00	1.00
personalData:update	1	1.00	1.00	1.00
personalData:usage	3	1.00	1.00	0.50
prices:currency	3	0.00	0.00	0.00
prices:vat	5	1.00	1.00	0.89
warranty:options	1	0.00	0.00	0.00
warranty:period	2	1.00	1.00	0.67
withdrawal:compensation	5	1.00	1.00	0.89
withdrawal:effects	2	0.00	0.00	0.00
withdrawal:exclusion	5	0.00	0.00	0.00
withdrawal:form	7	1.00	1.00	0.83
withdrawal:model	0	0.00	0.00	0.00
withdrawal:period	8	0.60	0.60	0.46
withdrawal:shippingCosts	9	0.67	0.67	0.33
withdrawal:shippingMethod	3	1.00	1.00	0.50
micro avg	235	0.80	0.80	0.67
macro avg	235	0.50	0.50	0.38
weighted avg	235	0.73	0.73	0.63
samples avg	235	0.60	0.60	0.59

Table F.37.: Results of the Logistic Regression clause subtopic classification in English using the clause text as input

Clause Subtopic	Support	P	R	F1
	101	0.78	0.78	0.78
conclusionOfContract:binding	8	0.80	0.80	0.62
conclusionOfContract:changeOfOrder	1	1.00	1.00	1.00
conclusionOfContract:definition	1	0.00	0.00	0.00
conclusionOfContract:restrictions	1	0.00	0.00	0.00
conclusionOfContract:steps	12	0.73	0.73	0.70
conclusionOfContract:withdrawal	4	1.00	1.00	0.67
delivery:brokenPackaging	2	0.00	0.00	0.00
delivery:costs	11	0.80	0.80	0.76
delivery:customs	1	0.00	0.00	0.00
delivery:destination	3	1.00	1.00	0.50
delivery:methods	3	1.00	1.00	0.80
delivery:partial	1	0.00	0.00	0.00
delivery:time	8	1.00	1.00	0.55
payment:fee	1	0.00	0.00	0.00
payment:late	0	0.00	0.00	0.00
payment:methods	4	1.00	1.00	0.67
payment:loyalty	11	0.80	0.80	0.76
payment:restraint	0	0.00	0.00	0.00
payment:vouchers	3	0.00	0.00	0.00
personalData:cookies	1	0.00	0.00	0.00
personalData:duration	0	0.00	0.00	0.00
personalData:information	2	1.00	1.00	0.67
personalData:reason	2	1.00	1.00	1.00
personalData:update	1	1.00	1.00	1.00
personalData:usage	3	1.00	1.00	0.80
prices:currency	3	0.00	0.00	0.00
prices:vat	5	1.00	1.00	0.89
warranty:options	1	0.00	0.00	0.00
warranty:period	2	0.00	0.00	0.00
withdrawal:compensation	5	1.00	1.00	0.75
withdrawal:effects	2	0.00	0.00	0.00
withdrawal:exclusion	5	0.00	0.00	0.00
withdrawal:form	7	1.00	1.00	0.60
withdrawal:model	0	0.00	0.00	0.00
withdrawal:period	8	0.50	0.50	0.33
withdrawal:shippingCosts	9	0.75	0.75	0.46
withdrawal:shippingMethod	3	1.00	1.00	0.50
micro avg	235	0.80	0.80	0.68
macro avg	235	0.50	0.50	0.39
weighted avg	235	0.74	0.74	0.64
samples avg	235	0.62	0.62	0.60

Table F.38.: Results of the Logistic Regression clause subtopic classification in English using the clause text and both titles (paragraph and clause) as input

F.2.3. Random Forest

Clause Subtopic	Support	P	R	F1
	442.0	0.898	0.898	0.905
conclusionOfContract:binding	55.0	0.891	0.891	0.891
conclusionOfContract:changeOfOrder	8.0	0.375	0.375	0.375
conclusionOfContract:definition	23.0	1.000	1.000	0.878
conclusionOfContract:restrictions	7.0	1.000	1.000	0.250
conclusionOfContract:steps	48.0	0.935	0.935	0.734
conclusionOfContract:withdrawal	23.0	0.833	0.833	0.345
delivery:brokenPackaging	25.0	1.000	1.000	0.649
delivery:costs	54.0	1.000	1.000	0.633
delivery:customs	9.0	1.000	1.000	0.800
delivery:destination	20.0	0.750	0.750	0.250
delivery:methods	25.0	0.917	0.917	0.595
delivery:partial	14.0	1.000	1.000	0.353
delivery:time	28.0	0.750	0.750	0.333
payment:fee	15.0	1.000	1.000	0.333
payment:late	12.0	0.000	0.000	0.000
payment:loyalty	1.0	0.000	0.000	0.000
payment:methods	96.0	0.932	0.932	0.805
payment:restraint	12.0	1.000	1.000	0.909
payment:vouchers	42.0	1.000	1.000	0.895
personalData:cookies	1.0	0.000	0.000	0.000
personalData:duration	3.0	0.000	0.000	0.000
personalData:information	12.0	0.000	0.000	0.000
personalData:reason	16.0	0.000	0.000	0.000
personalData:update	0.0	0.000	0.000	0.000
personalData:usage	18.0	0.000	0.000	0.000
prices:currency	5.0	1.000	1.000	0.333
prices:vat	22.0	1.000	1.000	0.900
warranty:options	18.0	1.000	1.000	0.500
warranty:period	30.0	0.933	0.933	0.622
withdrawal:compensation	19.0	0.938	0.938	0.857
withdrawal:effects	20.0	0.941	0.941	0.865
withdrawal:exclusion	23.0	1.000	1.000	0.850
withdrawal:form	32.0	0.952	0.952	0.755
withdrawal:model	9.0	0.900	0.900	0.947
withdrawal:period	28.0	1.000	1.000	0.809
withdrawal:shippingCosts	24.0	0.786	0.786	0.579
withdrawal:shippingMethod	14.0	0.714	0.714	0.476
micro avg	1253.0	0.914	0.914	0.773
macro avg	1253.0	0.722	0.722	0.511
weighted avg	1253.0	0.870	0.870	0.732
samples avg	1253.0	0.743	0.743	0.727

Table F.39.: Results of the Random Forest clause subtopic classification in German

Clause Subtopic	Support	P	R	F1
	99.0	0.846	0.846	0.811
conclusionOfContract:binding	10.0	1.000	1.000	0.333
conclusionOfContract:changeOfOrder	2.0	1.000	1.000	0.667
conclusionOfContract:definition	1.0	0.000	0.000	0.000
conclusionOfContract:restrictions	1.0	0.000	0.000	0.000
conclusionOfContract:steps	8.0	0.800	0.800	0.615
conclusionOfContract:withdrawal	6.0	1.000	1.000	0.286
delivery:brokenPackaging	1.0	0.000	0.000	0.000
delivery:costs	15.0	1.000	1.000	0.333
delivery:customs	0.0	0.000	0.000	0.000
delivery:destination	5.0	0.000	0.000	0.000
delivery:methods	5.0	0.000	0.000	0.000
delivery:partial	2.0	1.000	1.000	0.667
delivery:time	8.0	0.000	0.000	0.000
payment:fee	0.0	0.000	0.000	0.000
payment:late	0.0	0.000	0.000	0.000
payment:loyalty	6.0	0.000	0.000	0.000
payment:methods	8.0	1.000	1.000	0.667
payment:restraint	0.0	0.000	0.000	0.000
payment:vouchers	5.0	0.000	0.000	0.000
personalData:cookies	1.0	0.000	0.000	0.000
personalData:duration	0.0	0.000	0.000	0.000
personalData:information	1.0	0.000	0.000	0.000
personalData:reason	1.0	1.000	1.000	1.000
personalData:update	1.0	1.000	1.000	1.000
personalData:usage	5.0	0.000	0.000	0.000
prices:currency	2.0	0.000	0.000	0.000
prices:vat	4.0	1.000	1.000	0.667
warranty:options	1.0	0.000	0.000	0.000
warranty:period	2.0	0.000	0.000	0.000
withdrawal:compensation	6.0	0.000	0.000	0.000
withdrawal:effects	4.0	0.000	0.000	0.000
withdrawal:exclusion	7.0	0.000	0.000	0.000
withdrawal:form	3.0	1.000	1.000	0.500
withdrawal:model	1.0	0.000	0.000	0.000
withdrawal:period	12.0	1.000	1.000	0.286
withdrawal:shippingCosts	8.0	0.500	0.500	0.333
withdrawal:shippingMethod	3.0	1.000	1.000	0.800
micro avg	244.0	0.852	0.852	0.568
macro avg	244.0	0.372	0.372	0.236
weighted avg	244.0	0.661	0.661	0.483
samples avg	244.0	0.470	0.470	0.464

Table F.40.: Results of the Random Forest clause subtopic classification in English

F.2.4. Multilayer Perceptron

Clause Topic	Support	P	R	F1
	441.0	0.910	0.910	0.913
conclusionOfContract:binding	66.0	0.846	0.846	0.840
conclusionOfContract:changeOfOrder	12.0	0.000	0.000	0.000
conclusionOfContract:definition	21.0	0.905	0.905	0.905
conclusionOfContract:restrictions	8.0	0.600	0.600	0.462
conclusionOfContract:steps	51.0	0.714	0.714	0.506
conclusionOfContract:withdrawal	19.0	0.636	0.636	0.467
delivery:brokenPackaging	27.0	0.864	0.864	0.776
delivery:costs	49.0	0.839	0.839	0.650
delivery:customs	9.0	1.000	1.000	0.714
delivery:destination	19.0	0.500	0.500	0.296
delivery:methods	32.0	0.625	0.625	0.625
delivery:partial	6.0	1.000	1.000	0.286
delivery:time	29.0	0.696	0.696	0.615
payment:fee	10.0	0.714	0.714	0.588
payment:late	10.0	0.625	0.625	0.556
payment:loyalty	1.0	0.000	0.000	0.000
payment:methods	87.0	0.962	0.962	0.909
payment:restraint	9.0	1.000	1.000	0.875
payment:vouchers	60.0	0.963	0.963	0.912
personalData:cookies	1.0	0.000	0.000	0.000
personalData:duration	2.0	0.000	0.000	0.000
personalData:information	10.0	0.000	0.000	0.000
personalData:reason	10.0	0.000	0.000	0.000
personalData:update	1.0	0.000	0.000	0.000
personalData:usage	11.0	0.273	0.273	0.273
prices:currency	4.0	0.000	0.000	0.000
prices:vat	24.0	0.944	0.944	0.810
warranty:options	14.0	0.714	0.714	0.476
warranty:period	31.0	0.781	0.781	0.794
withdrawal:compensation	19.0	0.000	0.000	0.000
withdrawal:effects	20.0	0.000	0.000	0.000
withdrawal:exclusion	20.0	1.000	1.000	0.710
withdrawal:form	28.0	0.000	0.000	0.000
withdrawal:model	8.0	1.000	1.000	0.933
withdrawal:period	25.0	0.000	0.000	0.000
withdrawal:shippingCosts	24.0	0.500	0.500	0.143
withdrawal:shippingMethod	15.0	0.000	0.000	0.000
micro avg	1233.0	0.859	0.859	0.746
macro avg	1233.0	0.516	0.516	0.422
weighted avg	1233.0	0.748	0.748	0.691
samples avg	1233.0	0.817	0.817	0.784

Table F.41.: Results of the Multilayer Perceptron clause subtopic classification in German using Tf-idf vectors as input

Clause Topic	Support	P	R	F1
	101.0	0.819	0.819	0.835
conclusionOfContract:binding	8.0	0.800	0.800	0.615
conclusionOfContract:changeOfOrder	1.0	0.000	0.000	0.000
conclusionOfContract:definition	1.0	0.000	0.000	0.000
conclusionOfContract:restrictions	1.0	0.000	0.000	0.000
conclusionOfContract:steps	12.0	0.692	0.692	0.720
conclusionOfContract:withdrawal	4.0	0.750	0.750	0.750
delivery:brokenPackaging	2.0	1.000	1.000	0.667
delivery:costs	11.0	0.900	0.900	0.857
delivery:customs	1.0	1.000	1.000	1.000
delivery:destination	3.0	0.500	0.500	0.400
delivery:methods	3.0	1.000	1.000	0.500
delivery:partial	1.0	0.000	0.000	0.000
delivery:time	8.0	0.667	0.667	0.706
payment:fee	1.0	0.000	0.000	0.000
payment:late	0.0	0.000	0.000	0.000
payment:loyalty	4.0	1.000	1.000	1.000
payment:methods	11.0	0.700	0.700	0.667
payment:restraint	0.0	0.000	0.000	0.000
payment:vouchers	3.0	0.750	0.750	0.857
personalData:cookies	1.0	0.000	0.000	0.000
personalData:duration	0.0	0.000	0.000	0.000
personalData:information	2.0	0.000	0.000	0.000
personalData:reason	2.0	0.000	0.000	0.000
personalData:update	1.0	0.000	0.000	0.000
personalData:usage	3.0	0.500	0.500	0.400
prices:currency	3.0	0.000	0.000	0.000
prices:vat	5.0	0.667	0.667	0.727
warranty:options	1.0	0.000	0.000	0.000
warranty:period	2.0	0.000	0.000	0.000
withdrawal:compensation	5.0	0.429	0.429	0.500
withdrawal:effects	2.0	0.000	0.000	0.000
withdrawal:exclusion	5.0	1.000	1.000	0.571
withdrawal:form	7.0	1.000	1.000	0.727
withdrawal:model	0.0	0.000	0.000	0.000
withdrawal:period	8.0	0.750	0.750	0.500
withdrawal:shippingCosts	9.0	0.400	0.400	0.286
withdrawal:shippingMethod	3.0	1.000	1.000	0.500
micro avg	235.0	0.764	0.764	0.708
macro avg	235.0	0.430	0.430	0.363
weighted avg	235.0	0.718	0.718	0.673
samples avg	235.0	0.738	0.738	0.713

Table F.42.: Results of the Multilayer Perceptron clause subtopic classification in English using Word2Vec vectors as input

F.2.5. Convolutional Neural Network

Clause Topic	Support	P	R	F1
	441.0	0.903	0.903	0.895
conclusionOfContract:binding	66.0	0.851	0.851	0.900
conclusionOfContract:changeOfOrder	12.0	1.000	1.000	0.286
conclusionOfContract:definition	21.0	1.000	1.000	0.950
conclusionOfContract:restrictions	8.0	0.000	0.000	0.000
conclusionOfContract:steps	51.0	0.840	0.840	0.553
conclusionOfContract:withdrawal	19.0	0.600	0.600	0.250
delivery:brokenPackaging	27.0	0.944	0.944	0.756
delivery:costs	49.0	0.688	0.688	0.543
delivery:customs	9.0	0.833	0.833	0.667
delivery:destination	19.0	0.400	0.400	0.167
delivery:methods	32.0	0.938	0.938	0.625
delivery:partial	9.0	0.000	0.000	0.000
delivery:time	29.0	0.750	0.750	0.324
payment:fee	10.0	0.000	0.000	0.000
payment:late	10.0	0.000	0.000	0.000
payment:loyalty	1.0	0.000	0.000	0.000
payment:methods	87.0	0.951	0.951	0.784
payment:restraint	9.0	1.000	1.000	1.000
payment:vouchers	64.0	0.905	0.905	0.898
personalData:cookies	0.0	0.000	0.000	0.000
personalData:duration	2.0	0.000	0.000	0.000
personalData:information	10.0	0.000	0.000	0.000
personalData:reason	10.0	0.000	0.000	0.000
personalData:update	1.0	0.000	0.000	0.000
personalData:usage	12.0	0.000	0.000	0.000
prices:currency	3.0	0.000	0.000	0.000
prices:vat	24.0	1.000	1.000	0.703
warranty:options	14.0	1.000	1.000	0.353
warranty:period	31.0	0.929	0.929	0.881
withdrawal:compensation	19.0	0.000	0.000	0.000
withdrawal:effects	19.0	0.000	0.000	0.000
withdrawal:exclusion	20.0	1.000	1.000	0.462
withdrawal:form	26.0	0.000	0.000	0.000
withdrawal:model	8.0	1.000	1.000	0.857
withdrawal:period	25.0	0.000	0.000	0.000
withdrawal:shippingCosts	24.0	0.000	0.000	0.000
withdrawal:shippingMethod	15.0	0.000	0.000	0.000
micro avg	1236.0	0.894	0.894	0.720
macro avg	1236.0	0.461	0.461	0.338
weighted avg	1236.0	0.738	0.738	0.644
samples avg	1236.0	0.737	0.737	0.712

Table F.43.: Results of the CNN clause subtopic classification in German using Word2Vec vectors as input

Clause Topic	Support	P	R	F1
	99.0	0.788	0.788	0.808
conclusionOfContract:binding	10.0	1.000	1.000	0.462
conclusionOfContract:changeOfOrder	2.0	0.000	0.000	0.000
conclusionOfContract:definition	1.0	0.000	0.000	0.000
conclusionOfContract:restrictions	1.0	0.000	0.000	0.000
conclusionOfContract:steps	8.0	0.556	0.556	0.588
conclusionOfContract:withdrawal	6.0	1.000	1.000	0.667
delivery:brokenPackaging	1.0	0.000	0.000	0.000
delivery:costs	15.0	0.818	0.818	0.692
delivery:customs	0.0	0.000	0.000	0.000
delivery:destination	5.0	0.000	0.000	0.000
delivery:methods	5.0	0.000	0.000	0.000
delivery:partial	2.0	0.000	0.000	0.000
delivery:time	8.0	1.000	1.000	0.222
payment:fee	0.0	0.000	0.000	0.000
payment:late	0.0	0.000	0.000	0.000
payment:loyalty	5.0	1.000	1.000	0.750
payment:methods	9.0	0.833	0.833	0.667
payment:restraint	0.0	0.000	0.000	0.000
payment:vouchers	5.0	0.000	0.000	0.000
personalData:cookies	1.0	0.000	0.000	0.000
personalData:duration	0.0	0.000	0.000	0.000
personalData:information	1.0	0.000	0.000	0.000
personalData:reason	1.0	1.000	1.000	1.000
personalData:update	1.0	0.000	0.000	0.000
personalData:usage	4.0	0.000	0.000	0.000
prices:currency	2.0	0.000	0.000	0.000
prices:vat	4.0	0.750	0.750	0.750
warranty:options	1.0	0.000	0.000	0.000
warranty:period	2.0	0.000	0.000	0.000
withdrawal:compensation	6.0	1.000	1.000	0.286
withdrawal:effects	4.0	0.000	0.000	0.000
withdrawal:exclusion	7.0	0.000	0.000	0.000
withdrawal:form	3.0	0.333	0.333	0.333
withdrawal:model	1.0	0.000	0.000	0.000
withdrawal:period	12.0	0.625	0.625	0.500
withdrawal:shippingCosts	8.0	0.714	0.714	0.667
withdrawal:shippingMethod	3.0	0.000	0.000	0.000
micro avg	243.0	0.770	0.770	0.623
macro avg	243.0	0.300	0.300	0.221
weighted avg	243.0	0.640	0.640	0.548
samples avg	243.0	0.599	0.599	0.579

Table F.44.: Results of the CNN clause subtopic classification in English using Word2Vec vectors as input

F.2.6. Recurrent Neural Network

Clause Topic	Support	P	R	F1
	441.0	0.911	0.911	0.908
conclusionOfContract:binding	66.0	0.803	0.803	0.859
conclusionOfContract:changeOfOrder	12.0	1.000	1.000	0.154
conclusionOfContract:definition	21.0	0.905	0.905	0.905
conclusionOfContract:restrictions	8.0	1.000	1.000	0.667
conclusionOfContract:steps	51.0	0.643	0.643	0.456
conclusionOfContract:withdrawal	19.0	0.444	0.444	0.286
delivery:brokenPackaging	27.0	0.909	0.909	0.816
delivery:costs	49.0	0.697	0.697	0.561
delivery:customs	9.0	1.000	1.000	0.714
delivery:destination	19.0	0.455	0.455	0.333
delivery:methods	32.0	0.750	0.750	0.643
delivery:partial	9.0	0.000	0.000	0.000
delivery:time	29.0	0.842	0.842	0.667
payment:fee	10.0	0.250	0.250	0.143
payment:late	10.0	0.667	0.667	0.308
payment:loyalty	1.0	0.000	0.000	0.000
payment:methods	87.0	0.835	0.835	0.795
payment:restraint	9.0	1.000	1.000	0.941
payment:vouchers	64.0	0.965	0.965	0.909
personalData:cookies	0.0	0.000	0.000	0.000
personalData:duration	2.0	0.000	0.000	0.000
personalData:information	10.0	0.000	0.000	0.000
personalData:reason	10.0	0.000	0.000	0.000
personalData:update	1.0	0.000	0.000	0.000
personalData:usage	12.0	0.250	0.250	0.125
prices:currency	3.0	0.000	0.000	0.000
prices:vat	24.0	0.895	0.895	0.791
warranty:options	14.0	0.625	0.625	0.455
warranty:period	31.0	0.720	0.720	0.643
withdrawal:compensation	19.0	0.000	0.000	0.000
withdrawal:effects	19.0	0.000	0.000	0.000
withdrawal:exclusion	20.0	0.600	0.600	0.240
withdrawal:form	26.0	0.000	0.000	0.000
withdrawal:model	8.0	1.000	1.000	0.222
withdrawal:period	25.0	0.000	0.000	0.000
withdrawal:shippingCosts	24.0	1.000	1.000	0.222
withdrawal:shippingMethod	15.0	0.000	0.000	0.000
micro avg	1236.0	0.852	0.852	0.721
macro avg	1236.0	0.504	0.504	0.362
weighted avg	1236.0	0.738	0.738	0.658
samples avg	1236.0	0.765	0.765	0.738

Table F.45.: Results of the LSTM clause subtopic classification in German using Word2Vec vectors as input

F.2.7. BERT

Clause Topic	Support	P	R	F1
	441.0	0.958	0.958	0.916
conclusionOfContract:binding	66.0	0.928	0.928	0.948
conclusionOfContract:changeOfOrder	12.0	0.909	0.909	0.870
conclusionOfContract:definition	21.0	0.950	0.950	0.927
conclusionOfContract:restrictions	8.0	0.800	0.800	0.615
conclusionOfContract:steps	51.0	0.892	0.892	0.750
conclusionOfContract:withdrawal	19.0	0.647	0.647	0.611
delivery:brokenPackaging	27.0	0.950	0.950	0.809
delivery:costs	49.0	0.897	0.897	0.795
delivery:customs	9.0	1.000	1.000	0.800
delivery:destination	19.0	0.895	0.895	0.895
delivery:methods	32.0	0.833	0.833	0.806
delivery:partial	9.0	1.000	1.000	0.364
delivery:time	29.0	0.706	0.706	0.762
payment:fee	10.0	0.625	0.625	0.556
payment:late	10.0	0.833	0.833	0.625
payment:loyalty	1.0	0.000	0.000	0.000
payment:methods	87.0	0.963	0.963	0.935
payment:restraint	9.0	0.818	0.818	0.900
payment:vouchers	64.0	1.000	1.000	0.984
personalData:cookies	0.0	0.000	0.000	0.000
personalData:duration	2.0	0.000	0.000	0.000
personalData:information	10.0	0.667	0.667	0.632
personalData:reason	10.0	0.889	0.889	0.842
personalData:update	1.0	0.000	0.000	0.000
personalData:usage	12.0	0.700	0.700	0.636
prices:currency	3.0	0.000	0.000	0.000
prices:vat	24.0	0.846	0.846	0.880
warranty:options	14.0	0.722	0.722	0.813
warranty:period	31.0	0.806	0.806	0.806
withdrawal:compensation	19.0	0.773	0.773	0.829
withdrawal:effects	19.0	0.708	0.708	0.791
withdrawal:exclusion	20.0	0.846	0.846	0.667
withdrawal:form	26.0	0.767	0.767	0.821
withdrawal:model	8.0	0.875	0.875	0.875
withdrawal:period	25.0	0.893	0.893	0.943
withdrawal:shippingCosts	24.0	0.710	0.710	0.800
withdrawal:shippingMethod	15.0	0.727	0.727	0.615
micro avg	1236.0	0.891	0.891	0.860
macro avg	1236.0	0.725	0.725	0.679
weighted avg	1236.0	0.891	0.891	0.854
samples avg	1236.0	0.876	0.876	0.860

Table F.46.: Results of the BERT clause subtopic classification in German

Clause Topic	Support	P	R	F1
	99.0	0.815	0.815	0.850
conclusionOfContract:binding	10.0	0.750	0.750	0.667
conclusionOfContract:changeOfOrder	2.0	0.000	0.000	0.000
conclusionOfContract:definition	1.0	0.000	0.000	0.000
conclusionOfContract:restrictions	1.0	0.000	0.000	0.000
conclusionOfContract:steps	8.0	0.500	0.500	0.600
conclusionOfContract:withdrawal	6.0	1.000	1.000	0.667
delivery:brokenPackaging	1.0	0.000	0.000	0.000
delivery:costs	15.0	0.813	0.813	0.839
delivery:customs	0.0	0.000	0.000	0.000
delivery:destination	5.0	0.000	0.000	0.000
delivery:methods	5.0	0.000	0.000	0.000
delivery:partial	2.0	0.000	0.000	0.000
delivery:time	8.0	1.000	1.000	0.769
payment:fee	0.0	0.000	0.000	0.000
payment:late	0.0	0.000	0.000	0.000
payment:loyalty	5.0	1.000	1.000	1.000
payment:methods	9.0	1.000	1.000	0.875
payment:restraint	0.0	0.000	0.000	0.000
payment:vouchers	5.0	0.667	0.667	0.500
personalData:cookies	1.0	0.000	0.000	0.000
personalData:duration	0.0	0.000	0.000	0.000
personalData:information	1.0	0.000	0.000	0.000
personalData:reason	1.0	1.000	1.000	1.000
personalData:update	1.0	0.000	0.000	0.000
personalData:usage	4.0	1.000	1.000	0.400
prices:currency	2.0	0.000	0.000	0.000
prices:vat	4.0	0.800	0.800	0.889
warranty:options	1.0	0.000	0.000	0.000
warranty:period	2.0	0.000	0.000	0.000
withdrawal:compensation	6.0	0.667	0.667	0.444
withdrawal:effects	4.0	0.000	0.000	0.000
withdrawal:exclusion	7.0	1.000	1.000	0.444
withdrawal:form	3.0	0.500	0.500	0.571
withdrawal:model	1.0	0.000	0.000	0.000
withdrawal:period	12.0	0.667	0.667	0.741
withdrawal:shippingCosts	8.0	0.625	0.625	0.625
withdrawal:shippingMethod	3.0	0.000	0.000	0.000
micro avg	243.0	0.783	0.783	0.720
macro avg	243.0	0.363	0.363	0.313
weighted avg	243.0	0.697	0.697	0.664
samples avg	243.0	0.741	0.741	0.713

Table F.47.: Results of the BERT clause subtopic classification in English