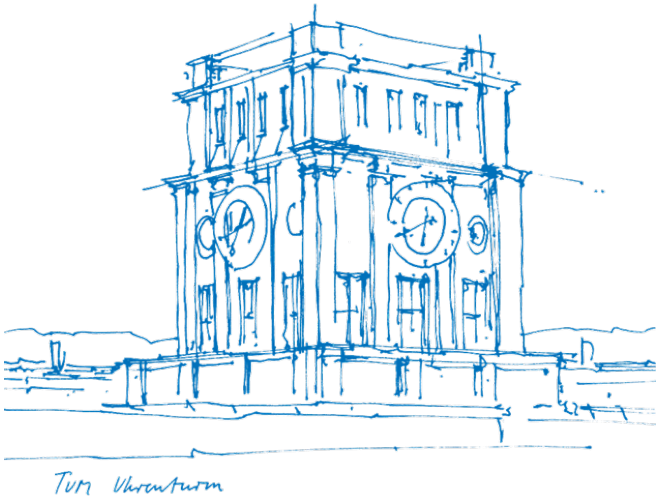
TUM School of Computation, Information and Technology
Technical University of Munich

Opinion Mining for Qualitative Content Studies

Gerhard Johann Hagerer



Opinion Mining for Qualitative Content Studies

Gerhard Johann Hagerer

TUM School of Computation, Information and Technology
Technische Universität München

Opinion Mining for Qualitative Content Studies

Gerhard Johann Hagerer

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:
Prof. Dr. Gudrun J. Klinker

Prüfer*innen der Dissertation:
1. apl. Prof. Dr. Georg Groh
2. Prof. Dr. Jürgen Pfeffer

Die Dissertation wurde am 03.06.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 12.04.2023 angenommen.

For an innocent boy residing in his place
of meadows, animals, music, and friends.

Abstract

Social media provides rapidly growing content in the form of texts and multimedia files, which are rich data sources for many research disciplines. With regard to processing these sources, there is a substantial evolution of data science related disciplines, such as, machine learning (ML), natural language processing (NLP), et cetera. However, the integration of these innovations into the workflow of domain experts to manually analyze social media big data in application domains is understudied. For instance, in qualitative research, data analyses, such as, content studies, are still carried out entirely manually on small amounts of thoughtfully selected data. Thereby, the level of detail provided by expert researchers is still unmet by computer algorithms. Thus, a professional, handcrafted analysis is the gold standard for precise explanations and theories about individuals and their comments on the world as they perceive it.

However, qualitative studies are carried out on small datasets, which are condemned to be incomplete and biased. For example, a specific data source, e.g., a domain-specific online forum about a particular topic, might be frequented by individuals with a predetermined mindset, which is commonly known as filter bubble. This can be imposed by many factors, among other, culture, language, community guidelines, and so on. Consequently, it is desirable to investigate as much data from as various data sources as possible to be able to discover all possible discussed themes and expressed opinions of a domain in a more representative manner.

Luckily, social media provides manifold sources from all sorts of interest groups and mindsets. At the same time, the recent advances in NLP and deep learning provide computational methods with unforeseen performance with regard to semantic coherence and accuracy. More than ever before, computer algorithms offer a highly connected and contextualized understanding of unstructured amounts of texts, leading to improved summarization of big social media data. Tailoring that technology for the end user throws the spotlight on the person who uses the methods in order to gain insight from social media data, which is the so-called text miner or domain expert. State-of-the-art (SOTA) methods need to add value to the tasks carried out by text miners to improve understanding of the investigated data domains and to enable better research.

This thesis aims at leveraging SOTA NLP methods to support a domain expert in the big data opinion mining process. It is dedicated to the exploration of an extended range of discussed themes as well as to providing representative quantitative statistics from social media texts. We show the potential to improve domain exploration by providing tools and methods unveiling the big picture of social media, i.e., by displaying opinions about as many topics and aspects as possible at multiple languages at a time. Further, our studies demonstrate how to complement or replace cross-cultural representative surveys by using opinion mining technology. Last but not least, we improve the SOTA of existing NLP to transfer explanations, theories, and structured knowledge provided by domain experts from limited data to big data using predictive ML models.

Zusammenfassung

Soziale Medien stellen rapide wachsende Inhalte in Form von Texten und Multimediadateien zur Verfügung, die für viele Forschungsdisziplinen reichhaltige Datenquellen darstellen. Im Hinblick auf die Verarbeitung dieser Quellen gibt es eine starke Entwicklung der mit den Datenwissenschaften verbundenen Disziplinen, wie z.B. maschinelles Lernen, Computerlinguistik, Text Mining, und so weiter. Die Integration dieser Innovationen in den Arbeitsablauf von Fachleuten zur manuellen Analyse von Big Data sozialen Medien in Anwendungsbereichen ist jedoch noch nicht ausreichend untersucht worden. In der qualitativen Forschung beispielsweise wird die Datenanalyse immer noch vollständig manuell an nicht repräsentativen Datenstichproben durchgeführt. Gleichzeitig wird der Detaillierungsgrad eines erfahrenen Forschers von Computeralgorithmen noch nicht erreicht. Eine professionelle, handwerkliche Analyse ist der Goldstandard für präzise Erklärungen und Theorien darüber, wie Menschen zu ihren Schlussfolgerungen und Meinungen über die Welt, die sie umgibt, kommen.

Erkenntnisse, die aus nicht repräsentativen, kleinen Daten gewonnen werden, sind jedoch dazu verurteilt, unvollständig und verzerrt zu sein. So mag eine bestimmte Datenquelle, z. B. ein spezifisches Online-Forum zu einem bestimmten Thema, primär von Personen mit einer bestimmten Denkweise frequentiert werden, die gemeinhin als Filterblase bezeichnet wird. Dies kann durch viele Faktoren bedingt sein, unter anderem durch Kultur, Sprache, Gemeinschaftsrichtlinien und so fort. Daher ist es wünschenswert, möglichst viele Daten aus verschiedenen Datenquellen zu untersuchen, um alle möglichen diskutierten Themen und geäußerten Meinungen eines Bereichs repräsentativ zu erfassen.

Glücklicherweise bieten die sozialen Medien vielfältige Quellen aus allen möglichen Interessengruppen. Gleichzeitig bieten die jüngsten Fortschritte in den Bereichen NLP und Deep Learning Computermethoden mit ungeahnten Leistungen in Bezug auf semantische Kohärenz und Genauigkeit. Mehr als je zuvor bieten Computeralgorithmen ein hochgradig vernetztes und kontextualisiertes Verständnis von unstrukturierten Textmengen, was zu einer verbesserten Zusammenfassung großer Datenmengen in sozialen Medien führt. Die Anpassung dieser Technologie an den Endnutzer rückt die Person in den Mittelpunkt, die die Methoden anwendet, um Erkenntnisse aus Social-Media-Daten zu gewinnen, d. h. den so genannten Text Miner oder Domänenexperten. Moderne Methoden müssen einen Mehrwert für die von Text Minern durchgeführten Aufgaben bieten, um das Verständnis der untersuchten Datendomäne zu verbessern und eine bessere Forschung zu ermöglichen.

Diese Arbeit zielt darauf ab, Computerlinguistik-Methoden auf dem Stand der Technik zu nutzen, um einen Domänenexperten im Big Data Opinion Mining-Prozess zu unterstützen. Sie widmet sich der Erkundung eines erweiterten Spektrums an diskutierten Themen sowie der Bereitstellung repräsentativer quantitativer Statistiken aus Social Media Texten. Wir zeigen das Potenzial, die Domänenexploration zu verbessern, indem wir Werkzeuge und Methoden bereitstellen, die das Gesamtbild der sozialen Medien enthüllen, d. h. indem wir Meinungen zu möglichst vielen Themen und Aspekten in mehreren Sprachen gleichzeitig anzeigen. Darüber hinaus zeigen unsere Studien, wie kulturübergreifende repräsentative Umfragen durch den Einsatz von Opinion-Mining-Technologie ergänzt oder

ersetzt werden können. Nicht zuletzt verbessern wir den Stand der Technik, um Erklärungen, Theorien und strukturiertes Wissen von Domänenexperten von begrenzten Daten auf Big Data unter Verwendung prädiktiver Modelle des maschinellen Lernens zu übertragen.

List of Contents

List of Figures

List of Tables

List of Abbreviations

ABSA aspect-based sentiment analysis. 6, 9, 10, 24, 41, 45, 47, 48, 58, 93, 100

CSAT customer satisfaction. 34, 35

DCA directed content analysis. xvii, 22, 24, 40, 48

LDA latent Dirichlet allocation. 11, 12

LSI latent semantic indexing. 11

ML machine learning. ix, 4, 5, 14, 22, 23, 34, 38, 46, 48, 59, 198

NLP natural language processing. ix, xi, 4–7, 10, 11, 19, 21, 23, 28–31, 34, 36, 38, 40, 45, 46, 52, 53, 57–59, 93, 112, 198

NMF non-negative matrix factorization. 11, 12, 50

SOTA state-of-the-art. ix, 5, 47, 48, 57, 58, 100, 126

USE Universal Sentence Encoder. 20, 46, 85, 188

XLING Universal Sentence Encoder Cross-Lingual. 46, 85, 188

List of Glossaries

aspect extraction The automatic classification of user-generated texts into aspect categories.. 10

code Synonym for annotation in the context of grounded theory and content analysis.. xv, 13, 14

content analysis The manual study of textual documents in social sciences using consistent coding or labeling and evaluating them statistically or qualitatively by deriving theories, e.g., using grounded theory.. 13

data domain Synonym for data domain.. ix, 31, 33, 59

domain expert Used as a synonym for text miner, but rather in contexts of qualitative research methodologies in social sciences, e.g., grounded theory.. ix, xv, 4, 5, 7, 14, 18, 19, 21–23, 27, 36, 38, 39, 41, 43, 44, 55, 57, 58, 198

domain of interest A domain of interest or data domain is the area from which a data source is chosen to crawl or download a corresponding textual corpus. Domains are defined by many aspects, such as, the type of medium (social media, product reviews, evaluation questionnaires, et cetera) and the topics discussed therein.. 7, 27, 31, 39

F1 score The harmonic mean of precision and recall. Used to estimate the quality of the predictions of a predictive model, e.g., a machine learning classifier.. 10, 11

grounded theory A qualitative research methodology used to manually derive theories and hypotheses by collecting and analyzing qualitative data, i.e., interviews, questionnaires with open-ended questions, or social media texts.. xv, 13, 14, 18, 19, 22–24, 35, 39, 43, 51, 57

original post The first comment in a social media thread, which actually starts a discussion. This can be a simple question, a newspaper article, an experience report, or any other form of comment.. 31

text miner Used as a synonym for domain expert, but rather in contexts of applied NLP.. ix, 7, 10, 12, 18, 21, 27, 57, 59

List of Publications

This thesis includes an introduction and the following publications. Publication properties, such as, peer-review status, full or other type of paper, conference or journal, relevance to the examination, splitted first authorships, and license, are commented onto each bibliography item. The contribution, authorship, and leadership role of Gerhard Johann Hagerer to the respective publications and, additionally, the according reprinting permission to include the paper as part of this thesis have been confirmed by all co-authors and all publishers. The latter give explicit permission either by directly licensing papers under Creative Commons (CC) licenses for open access or by granting reprint permission only for this thesis, i.e., all rights reserved.

Hannah Danner, Gerhard Johann Hagerer, Florian Kasischke, and Georg Groh. Combining content analysis and neural networks to analyze discussion topics in online comments about organic food. In *Conference Proceedings of "3rd International Conference on Advanced Research Methods and Analytics"*, pages 211–219, Valencia, Spain, July 2020. Editorial Universitat Politecnica de Valencia. ISBN 9788490488324. doi: 10.4995/CARMA2020.2020.11632. URL http://inis.iaea.org/search/search.aspx?orig_q=RN:51077731. Full paper, peer-reviewed, international conference; licensed under CC BY-NC-ND 4.0.

Hannah Danner, Gerhard Johann Hagerer, Yan Pan, and Georg Groh. The news media and its audience: Agenda-setting on organic food in the united states and germany. Full paper, peer-reviewed, international journal; accepted for publication in Journal of Cleaner Production; ©Elsevier., 2022.

Gerhard Johann Hagerer, Abdul Moeed, Yang An, and Georg Groh. Evaluation metrics for headline generation using deep pre-trained embeddings. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1796–1802, Marseille, France, May 2020a. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.222. Full paper, peer-reviewed, international conference; **relevant to the examination**; first authorship splitted equally among three first authors; licensed under CC BY-NC 2.0.

Gerhard Johann Hagerer, Abdul Moeed, Sumit Dugar, Sarthak Gupta, Mainak Ghosh, Hannah Danner, Oliver Mitevski, Andreas Nawroth, and Georg Groh. An evaluation of progressive neural networks for transfer learning in natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1376–1381, Marseille, France, 5 2020b. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.172. Full paper, peer-reviewed, international conference; **relevant to the examination**; first authorship splitted equally among five first authors; licensed under CC BY 2.0.

Gerhard Johann Hagerer, Martin Kirchhoff, Hannah Danner, Mainak Ghosh, Archishman Roy, Jiaxi Zhao, and Georg Groh. SocialVisTUM: An interactive visualization toolkit for correlated neural topic models on social media opinion mining. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 475–482, Varna, Bulgaria, September 2021a. INCOMA Ltd. URL https://arxiv.org/abs/2110.10575. Demo paper, peer-reviewed, international conference; licensed under CC BY 4.0.

Gerhard Johann Hagerer, Laura Lahesoo, Miriam Anschütz, and Stephan Krusche. An analysis of programming course evaluations before and after the introduction of an autograder. In *2021*

19th International Conference on Information Technology Based Higher Education and Training (ITHET), 2021b. URL https://arxiv.org/abs/2110.15134. Full paper, peer-reviewed, international conference; **relevant to the examination**; © IEEE 2021.

Gerhard Johann Hagerer, Wenbin Le, Hannah Danner, and Georg Groh. Classification of consumer belief statements from social media, June 2021c. URL https://arxiv.org/abs/2105.01466. arxiv report; licensed under CC BY 4.0.

Gerhard Johann Hagerer, Wing Sheung Leung, Hannah Danner, and Georg Groh. A case study and qualitative analysis of simple cross-lingual opinion mining. In *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR*, pages 17–26. INSTICC, SciTePress, November 2021d. ISBN 978-989-758-533-3. doi: 10.5220/0010649500003064. URL https://arxiv.org/abs/2111.02259. Full paper, peer-reviewed, international conference; **relevant to the examination**; licensed under CC BY-NC-ND 4.0.

Gerhard Johann Hagerer, David Szabo, Andreas Koch, Maria Luisa Ripoll Dominguez, Christian Widmer, Maximilian Wich, Hannah Danner, and Georg Groh. End-to-end annotator bias approximation on crowdsourced single-label sentiment analysis. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing*, Trento, Italy, November 2021e. Association for Computational Linguistics. URL https://arxiv.org/abs/2111.02326. Full paper, peer-reviewed, international conference; **relevant to the examination**; licensed under CC BY 4.0.

Lukas Stappen, Jason Thies, Gerhard Johann Hagerer, Björn W. Schuller, and Georg Groh. Graphtmt: Unsupervised graph-based topic modeling from video transcripts. In *2021 IEEE Seventh International Conference on Multimedia Big Data (BigMM)*, September 2021. URL https://arxiv.org/abs/2105.01466. Full paper, peer-reviewed, international conference; © IEEE 2021.

List of Supervised Theses

Ibrar Arshad. Contextualized data augmentation for sentiment analysis using deep pre-trained language models. Master's thesis, Technical University of Munich, Arcisstraße 21, 8 2020. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Shyam Arumugaswamy. Generative language models for opinion mining. Master's thesis, Technical University of Munich, Arcisstraße 21, 10 2019. Guided research under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Preethi Ballal. Implementation of an opinion mining framework for downloading and analyzing social media comments using pre-trained word embeddings. Master's thesis, Technical University of Munich, Arcisstraße 21, 4 2020. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Michael Brauweiler and Elisabeth Neumann. Topic-related sentiment analysis on open-ended student feedback in programming courses. Master's thesis, Technical University of Munich, Arcisstraße 21, 8 2021. Lab course report under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Sudeshna Dasgupta. A survey on aspect based sentiment analysis on social conversational text. Master's thesis, Technical University of Munich, Arcisstraße 21, 10 2019. Guided research under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Sudeshna Dasgupta. Deep learning approaches for opinion mining on conversational social media texts. Master's thesis, Technical University of Munich, Arcisstraße 21, 10 2020. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Kamal Datta. Aspect extraction using multi-instance learning and the google transformer architecture. Master's thesis, Technical University of Munich, Arcisstraße 21, 6 2019. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Sumit Dugar. Aspect-based sentiment analysis using deep neural networks and transfer learning. Master's thesis, Technical University of Munich, Arcisstraße 21, 3 2019. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Sebastian Erhardt. Generating language-independent neural sentence embeddings for natural language classification tasks. Master's thesis, Technical University of Munich, Arcisstraße 21, 3 2019. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Mainak Ghosh. Multilingual opinion mining on social media comments using unsupervised neural clustering methods. Master's thesis, Technical University of Munich, Arcisstraße 21, 10 2019. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Sarthak Gupta. Neural transfer learning for natural language processing. Master's thesis, Technical University of Munich, Arcisstraße 21, 4 2019. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Shayoni Halder. Textual similarity embeddings for cross-lingual aspect extraction. Master's thesis, Technical University of Munich, Arcisstraße 21, 10 2019. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Liyan Jiang. Deep active learning methods for opinion mining. Master's thesis, Technical University of Munich, Arcisstraße 21, 5 2020. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Florian Kasischke. Descriptive analytics for a nlp-based opinion mining. Master's thesis, Technical University of Munich, Arcisstraße 21, 3 2019. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., Hannah Danner, and Prof. Dr. Georg Groh.

Martin Kirchhoff. Summarizing opinions using neural topic modeling and graph-based visualizations. Master's thesis, Technical University of Munich, Arcisstraße 21, 11 2019. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Andreas Koch. Deep end-to-end learning for noisy annotations and crowdsourcing in natural language processing. Master's thesis, Technical University of Munich, Arcisstraße 21, 5 2021. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Laura Lahesoo. Applied topic modeling on results of lecture evaluation surveys at tum. Master's thesis, Technical University of Munich, Arcisstraße 21, 03 2021. Guided research under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Laura Lahesoo and Miriam Anschütz. Mining opinions from comments in the evaluation of a programming course. Master's thesis, Technical University of Munich, Arcisstraße 21, 08 2020. Lab course report under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Wenbin Le. Cluster distributions as document representations to analyze the class structure of a qualitative market research study. Master's thesis, Technical University of Munich, Arcisstraße 21, 03 2021. Guided research under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Wing Sheung Leung. Multi-class and multi-label text classification for industrial scenarios using pre-trained word embeddings. Master's thesis, Technical University of Munich, Arcisstraße 21, 03 2021. Guided research under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Phuong Mai. Transfer learning for aspect-based sentiment analysis using siamese networks. Master's thesis, Technical University of Munich, Arcisstraße 21, 10 2019. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Juanita Mendonca. Siamese networks for opinion mining using deep contextualized textual embeddings. Master's thesis, Technical University of Munich, Arcisstraße 21, 12 2019. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Abhilasha Mohania. Unsupervised aspect-based sentiment analysis applied on opinion research using deep learning and recent embedding techniques. Master's thesis, Technical University of Munich, Arcisstraße 21, 6 2019. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Ahmed Mosharafa. Deep learning approaches for cross-domain aspect-based sentiment analysis based on deep pre-trained embeddings and multi-task learning. Master's thesis, Technical University of Munich, Arcisstraße 21, 10 2019. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Farrukh Mushtaq. Predicting perception uncertainty for aspect-based sentiment analysis. Master's thesis, Technical University of Munich, Arcisstraße 21, 2 2020. Guided research under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Yan Pan. Topic modeling for opinion mining. Master's thesis, Technical University of Munich, Arcisstraße 21, 09 2020. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., Hannah Danner, and Prof. Dr. Georg Groh.

Maria Luisa Ripoll Dominguez. Interpretability in a question answering system via attention visualizations. Master's thesis, Technical University of Munich, Arcisstraße 21, 03 2021. Lab course report under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Benjamin Rösch. Multi-class and multi-label text classification for industrial scenarios using pretrained word embeddings. Master's thesis, Technical University of Munich, Arcisstraße 21, 11 2021. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Shabnam Sadegharmaki. Classification and class structure analysis for end-to-end opinion mining problems. Master's thesis, Technical University of Munich, Arcisstraße 21, 11 2020. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Anjali Sasihithlu. End-to-end aspect extraction on social media comments using attention models. Master's thesis, Technical University of Munich, Arcisstraße 21, 6 2020. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Felix Schober. Master thesis title: Transfer and multitask learning for aspect-based sentiment analysis using the google transformer architecture. Master's thesis, Technical University of Munich, Arcisstraße 21, 4 2019. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Omar Shouman. Transfer learning, language modelling, and deep neural networks for aspect-based sentiment analysis. Master's thesis, Technical University of Munich, Arcisstraße 21, 7 2019. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

David Szabo. Annotation bias in natural language processing and sentiment analysis. Master's thesis, Technical University of Munich, Arcisstraße 21, 03 2021. Guided research under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Jason Thies. Unsupervised topic and aspect modelling on massive video transcriptions. Master's thesis, Technical University of Munich, Arcisstraße 21, 3 2021. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Christian Widmer. Topic modeling for opinion mining. Master's thesis, Technical University of Munich, Arcisstraße 21, 04 2018. Bachelor's thesis under supervision of Dietrich Trautmann, M.Sc., Hannah Danner, and Prof. Dr. Georg Groh.

Haixiang Zhang. Predicting youtube user-engagement from textual representations and emotional signals. Master's thesis, Technical University of Munich, Arcisstraße 21, 06 2021. Master's thesis under supervision of Gerhard Johann Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Part I

Mantle Part

1 Introduction

1.1 Motivation

Social media, with its billion participators, is a mass phenomenon and one of the biggest collective encounters in the development of mankind. The individual experience is exposed and shared with an unprecedented large number of other humans. This work is attempting to bring the individual person closer to the collective social media consciousness, in the dualistic sense of two related but different social media analysis methodologies. The 'individual' one is interested in manually finding theoretical descriptions about specific individual motives and encounters of the world. The 'collective' one is machine-driven and aims at giving abstract high-level descriptions of big amounts of social media data. From a more general perspective, both conceptions are a recurring narrative since the Age of Enlightenment. Both are elementary views on existence and orthogonal in their truth claims, which is explained by Kant [1788] as follows:

> *"Two things fill the mind with ever new and increasing admiration and awe, the oftener and the more steadily we reflect on them: the starry heavens above and the moral law within. [...] The former begins from the place I occupy in the external world of sense [...]. The second begins from my invisible self, my personality [...]. The former view of a countless multitude of worlds annihilates my importance as an animal creature [...]. The second, on the contrary, infinitely elevates my worth as an intelligence by my personality, in which the moral law reveals to me a life independent of animality and even of the whole sensible world [...]. "*
>
> — Kant [1788]

By the *"starry heavens above"*, Kant means *"world of external sense"*, which includes the idea of finding objective truth using methods, such as, evidence of senses, reductionism, and the scientific method. Kant, an astronomer himself [Kant, 1755], refers to it as starry heavens due to its relation to astrophysics, i.e., a metaphor for the mathematical and natural sciences. At that time, astrophysics had an important influence on the age of enlightenment due to the Copernican revolution and the change from geocentrism to heliocentrism. In the context of this work, we relate the idea of the natural science view on the world to our perspective on big data and social media, using statistics, machine learning, and so on. We use it to depict the distribution of discussed high-level themes on data of a large number of samples. As these procedures are automatized, descriptive, and performed on many data sources, they are supposed to offer an unbiased, representative, synoptical view on the social media world. In the figurative sense, we use text mining instead of telescopes, and social media texts instead of planets and stars. This is what we refer to as *quantitative analysis*.

By the moral law within, Kant refers to the idea of the categorical imperative. It is used to evaluate individual motivations for action regarding their moral value. A motivation for

action is supposed to be good, i.e., matching the requirements of the categorical imperative, if its maxim *"should become a universal law"* [Kant, 1870]. We find that discussions on social media are also dealing with the question of how one acts as an individual. This is not limited to decisions of which products to buy and why or which properties of things are beneficial or not. Recently, ethical problems are increasingly considered when debating about consumption, e.g., ecological footprint in terms of carbon dioxide emissions or plastic waste et cetera. These problems are manifestly some of the most urgent ethical problems in the existence of mankind and play an increasing role in our everyday life. How these issues influence our thinking, behavior, and decisions is reflected in social media discussions as well. A deep understanding of how (specific groups of) individuals perform their own respective reasoning requires detailed manual investigation. In the end, a human expert is needed to empathically reconstruct all possible details about how subjects come to their conclusions based on a limited data sample. This approach lent from the social sciences is called *qualitative analysis*.

Even though both perspectives follow two completely different epistemologies, i.e., evidence of senses versus moral reasoning, they can be seen to be overlapping occasionally in modern research disciplines. This work elaborates on the overlap between computer science and sociology, or, more precisely, opinion mining and qualitative content analysis on textual data from social media. The underlying research question is how can state-of-the-art machine learning technology support a human domain expert finding, refining, and proofing theories to understand and explain human behavior based on social media big data.

1.2 Research Questions

The central research question of this thesis is how a domain expert can be supported in her qualitative content analysis using complementary opinion mining methods based on algorithms and machine learning (ML). Opinion mining methods have the obvious advantage of being able to process large amounts of social media data automatically, which would not be possible by a human. This is supposed to bring potential *benefits* to content studies, which we define in the following as one of our central research objectives.

The first and foremost objective is the methodical requirement that the automatically mined themes from social media are meaningful and correct, such that a domain expert would be able to work with it. To meet that requirement, all parts of the present work provide a mix of two kinds of evaluation methods. On the one hand, there are established objective measurements borrowed from natural language processing (NLP) and statistics applicable to clustering, classification, and so forth. On the other hand, qualitative interpretations should be in line with automatically mined themes and sentiments, such that the themes are assigned correctly according to the impression of the domain expert and can be interpreted meaningfully.

Secondly, it can be expected that automated NLP methods on big data discover the discussed themes and aspects such that a domain expert would come up with similar results. Since a domain expert cannot process that much data, the manually crafted qualitative themes should be a subset of the automatically mined big data themes. This principle is referred to as *perspective widening*, since the *totality* of concepts is supposed to be captured and acquired from big social media data, which would not be possible manually. At the same time, NLP methods show how the data is distributed among the themes, e.g.,

which topic, aspect, or sentiment occurs how often. This principle is called *quantitative depth*, where one might wonder if representative polling might become obsolete due to big data driven opinion mining. Between both principles, perspective widening and quantitative depth, the question is targeted which principle gives *additional value* for the domain expert and how. The expert goes through several, well-defined stages while performing content analysis manually, and each stage can benefit from automatized big data analysis routines. Here, additional value compared to traditional, manual content mining is crucial for the adoption by domain experts, since computational methods bring methodical complexity and produce information loss which needs to be recompensed. Thus, this work particularly aims at establishing simple methods which alleviate typical hurdles in adoption, such as, complex implementations or hyperparameter optimization routines. As an example, we provide a visualization tool incorporating state-of-the-art (SOTA) correlated neural topic models with automatized hyperparameter optimization and sentiment analysis. Its results are in line with an expert-conducted content study, and the topic visualizations provide an overview of meaningful themes and their statistical distributions Hagerer et al. [2021a].

Thirdly, we always aim at incorporating the SOTA of NLP into our methods. While the respective literature continuously introduces revolutionary new technologies, its interdisciplinary application is lagging behind. Especially in qualitative content studies, there is skepticism or non-consideration about recent computational methods. Establishing trust and interest by shading light on the specific advantages and potential of the technical SOTA for qualitative research is another goal of this thesis. For instance, we show how to perform cross-cultural content studies on big social media corpora using only one single, integrated model, which is a deep neural network with explainable properties [Danner et al., 2022].

Fourthly, we propose *crowdsourcing* as an alternative to the commonly applied expert-driven approach for qualitative content studies on social media text corpora. Many non-expert coders are able to improve machine learning in NLP while saving costs, which is why we investigate to what extent this can be helpful to support the qualitative text mining process of a domain expert. Especially automatic classification of texts with regard to topics, themes, entities, sentiments et cetera has a lot of potential yet unused by the humanities. However, non-expert annotator personnel does not have the same qualifications as domain experts, which raises the question of where is the limitation in terms of annotated quality and details. Exempli gratia, we show that detailed, fine-grained annotations appear to be generally difficult for being predicted by ML models, independent of the annotators being domain experts Le [2021] or non-experts Hagerer et al. [2020b]. It can be concluded that there is a need for simplicity, consistency, and noise reduction, such that automatized NLP routines really support the text mining process. We find this field of annotation collection and algorithmically supported, qualitative text mining to be understudied and depict some common challenges.

In that regard, misleading expectations raised by SOTA NLP methods shall be clarified and counter-strategies illustrated. Against the background of the deep learning revolution, it is easily forgotten that a proper knowledge about qualitative research and annotation methodology is mandatory for impactful interdisciplinary research. In that regard, we highlight the interdisciplinary gaps between NLP and qualitative research and aim at filling them up with innovative problem solutions, e.g., to overcome the forgetting effect in

transfer learning for aspect-based sentiment analysis (ABSA) [Hagerer et al., 2020b] or to remedy annotator bias in sentiment analysis [Hagerer et al., 2021e].

The denoted research goals are deliberately formulated in an abstract manner to bring the entirety down to a common denominator. The nature of the problems is diverse and interdisciplinary, such that a variety of disciplines and respective methods can be applied. We find that the methodologies from NLP and qualitative research are fundamentally different, and investigating the overlap of both is open to a range of differing problem formulations and approaches. Each of the publications at the end of this thesis addresses concrete research questions which refer to the previously denoted questions in their own way. Clear connections between the papers and their addressed problems are given at several stages in this dissertation, in particular in section 2.3 and section 3.2.

2 Methodology

This section describes the research topic of *opinion mining*, its definition, and its relevance for its application alongside of *qualitative content studies* conducted in sociological disciplines. We start by identifying the dichotomy of quantitative opinion mining and qualitative content studies, which are different approaches applied to the same type of textual data. The goal of the present thesis is to find how both disciplines can potentially benefit from each other, with a focus on innovative methods from NLP.

The term opinion mining is defined and distinguished from sentiment analysis by referring to the literature, showing that opinion mining has a broader meaning. This, for instance, is also implied by the word relationship between opinion mining on the one hand and text mining and data mining on the other hand. The latter two methodologies include methods to structure data which are naturally unstructured. This is suitable for the steps taking place before actual sentiment classification to support the collection, filtering, grouping, and descriptive analytics of text corpora. It supports the text miner to become informed about the domain of interest before and while labeling it, i.e., tagging text elements regarding their relevance, subjectivity, and opinion targets. If the whole opinion mining process consists of four steps: data collection, exploration of concept space, grouping or annotating of texts, and sentiment analysis, we see that actual opinion mining methods can be a part of each step, whereas sentiment analysis as such only takes place at the end. As the present work is not only focussed on sentiment analysis but also on all other steps, we prefer the term opinion mining over sentiment analysis.

Thirdly, we show how this work incorporates these steps. The process and the respective role of a text miner or domain expert is explained, who, during that process, gains insights into a domain of interest. This dissertation aims at solving problems occurring in this kind of manual research practice along the way, i.e., when the denoted steps are executed by a domain expert. Therefore, the different existing types of methodologies for qualitative content studies are explained. Eventually, we show how the publications included in this work are complementary and supportive for these kinds of research processes.

2.1 Opinion Mining

The term opinion mining is frequently used in the context of sentiment analysis. Some works use both terms synonymously for each other, defining it as *"the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes"* [Liu, 2012]. This definition discriminates between affect-related information and entity-related information. It implies that an expressed opinion consists of two things, i.e., a *subjective evaluation* of an *existing object*. This is intuitively clear, since an affect statement without a target is only an expression of an emotional state of mind, and an objective description of a thing is only a factual statement. However, when both are related with each other in a meaningful formulation, this establishes what we call an *opinion* in the context of this work. An opinion is assumed to be formalized as a piece of human-generated text. The task of automatically analyzing this textual information in view of extracting opinions by computational methods is what is called opinion mining. As a sidenote, it shall be noted that the process of finding opinions in social media texts manually is referred to as qualitative (content) analysis in this work, which is described in detail in a later section.

2.1.1 Sentiment Analysis

The definition of opinion mining can be further differentiated. In contrast to the previous definition, opinion mining and sentiment analysis can also be seen as two different but related research disciplines:

> *"Previous works on mining opinions can be divided into two directions: sentiment classification and* ***sentiment related information extraction****. The former is a task of identifying positive and negative sentiments from a text which can be a passage, a sentence, a phrase and even a word (Somasundaran et al., 2008; Pang et al., 2002; Dave et al., 2003; Kim and Hovy, 2004; Takamura et al., 2005).* ***The latter focuses on extracting the elements composing a sentiment text****. The elements include source of opinions who expresses an opinion (Choi et al., 2005);* ***target of opinions*** *which is a receptor of an opinion (Popescu and Etzioni, 2005); opinion expression which delivers an opinion (Wilson et al., 2005b). Some researchers refer this information extraction task as opinion extraction or* ***opinion mining****. Comparing with the former one, opinion mining usually produces richer information."*
>
> — Wu et al. [2009]

According to this definition, sentiment analysis is focussed on specifically detecting sentimental information from texts. In the following paragraph, several such examples from the literature are presented. Subjectivity is detected before or during actual sentiment analysis as a filter to detect if a text contains an opinion [Pang and Lee, 2004, Bird, 2006]. If it does not, the text can be discarded for sentiment analysis or a neutral sentiment can be assigned. Then, there are different ways to explain if a text is positive or negative. Sentiment or sentiment polarity classification resolves this task as a binary (positive/negative) or a multi-class (e.g., positive/negative/neutral) classification problem [Singh et al., 2013]. Sentiment detection can also be resolved as a regression task [Saad and Yang, 2019],

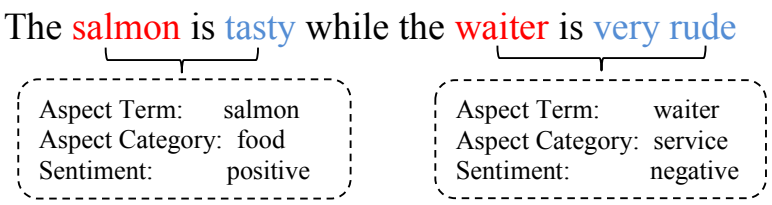


Figure 2.1 An example for aspect-based sentiment analysis and several involved sub-tasks, such as, aspect category classification, aspect term extraction, and sentiment polarity classification. Taken from Wu et al. [2020, p. 2].

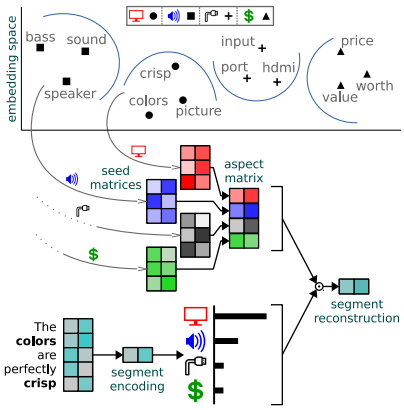
e.g., with a probability for positivity and negativity. For valence prediction, this is modeled as a single floating point number ranging from –1 or 0 for negative to +1 for positive [Eyben et al., 2017]. Valence is a psychological term used in emotion recognition. Therefore, good and bad feelings are recognized to estimate the sentiment of individuals while they are exposed to products [Kossaifi et al., 2021].

2.1.2 Target-Based Sentiment Analysis

As explained initially, a sentiment utterance refers to a corresponding target to form an actual opinion. Thus, the target can be seen as sentiment-related information, too. This can be an entity, e.g., a product, or an attribute of an entity, which is a part or an aspect of it. The classification of entity-sentiment tuples or entity-attribute-sentiment triplets is referred to as aspect-based sentiment analysis (ABSA). It is defined as *"mining opinions from text about specific entities and their aspects"* [Pontiki et al., 2015a]. Synonyms for ABSA are entity or (opinion) target based sentiment analysis, depending on the specific context. The opinion target in this case is represented by an entity and, if available, by its aspect:

> *"Aspect Based Sentiment Analysis (ABSA) systems receive as input a set of texts (e.g., product reviews or messages from social media) discussing a particular entity (e.g., a new model of a mobile phone). The systems attempt to detect the main (e.g., the most frequently discussed) aspects (features) of the entity (e.g., 'battery', 'screen') and to estimate the average sentiment of the texts per aspect (e.g., how positive or negative the opinions are on average for each aspect)."*
>
> — Pavlopoulos [2014, p. 1]

In the context of ABSA, this means an opinion can be annotated as a triple of entity, attribute, and sentiment, with the attribute being defined as "general" if not available [Pontiki et al., 2016]. These opinion triples are given as annotations on sub-sentence, sentence, and paragraph level. The problem is a multi-class multi-label classification problem, where the annotated opinion triplets are predicted. The ABSA task can be divided, such that recognizing the opinion target is done in a separate first step. This can be achieved in several ways, such as, classifying the category of the entity and the attribute (aspect category classification, Xue et al. [2017]), recognizing the entity and attribute words and their positions in the given text (aspect term extraction, Ma et al. [2019]), detecting the start and end of each formulated opinion (opinion target expression extraction, Al-Smadi et al. [2019]), or modeling these sub-tasks jointly with one single model [Nguyen and Shirai,



Aspect	Top Terms
Image	picture color quality black bright
Sound	sound speaker quality bass loud
Connectivity	hdmi port computer input component
Price	price value money worth paid
Apps	netflix user file hulu apps
Ease of Use	easy remote setup user menu
Service	paid support service week replace
Size & Look	size big bigger difference screen
General	tv bought hdtv happy problem

Figure 2.2 & Table 2.1 Aspects, their top words, and how these inform aspect extraction models. Taken from Angelidis and Lapata [2018b].

2018]. Concluding, ABSA can produce high level and low level features, from categorizing whole product reviews based on their product category down to sub-sentence chunking and grammatical feature generation.

2.1.3 Trends in Unsupervised Opinion Target Extraction

Aspect category classification, hereafter called aspect extraction, attracts significant attention from the literature. Its output can immediately be used to calculate the distribution of the opinion targets of a dataset, which helps to display its contents. It is also helpful as additional input for sentiment polarity classification and improves it. High accuracies and coherence can be achieved even with simple baseline techniques, which makes it favorable for many application domains, such as, market research [Stappen et al., 2021], teaching course evaluations [Hagerer et al., 2021b], and media agenda setting [Hagerer et al., 2021d, Danner et al., 2022], to mention just a few. Thus, it offers a pragmatic technical solution for text miners with an interest in social media, voice of the customer material, and other kinds of user-generated content.

Modern supervised methods, however, require the text miner to provide manual annotations in order to train and evaluate an aspect extraction algorithm. Annotations are difficult to prepare, as they involve much manual effort and high expertise [Danner and Menapace, 2020b]. Thus, recent natural language processing (NLP) research aims at removing the necessity of annotations for aspect extraction. For this purpose, weakly supervised and unsupervised methods are an emerging trend and increasingly successful. For instance, the annotation effort can be omitted by defining each aspect with a set of seed words. These are provided as additional input to train an aspect extraction model [Angelidis and Lapata, 2018b, Karamanolakis et al., 2019] — see, for example, Figure 2.2. They are used as prior to find text clusters in the corpus which correspond to the seed word lists. These clusters form the eventual aspect classes. For fine-grained aspect extraction in sentences, the F1 score drops by 10% for weakly supervised learning compared to fully supervised learning, while no annotations are provided during training [Karamanolakis et al., 2019]. Such models can even be used on previously unseen languages by leveraging transfer learning [Karamanolakis et al., 2019]. The initial seed words are not necessary when using unsupervised algorithms. Modern unsupervised neural attention architectures

are able to cluster clauses, sentences, and short texts with unforeseen semantic coherence [He et al., 2017b, Luo et al., 2019a]. Manual intervention is only necessary to map the clusters to the related aspect classes. The connection between clusters and classes is determined by inspecting a list of the most representative words — the so-called top words — of a cluster, which is then mapped manually to the best fitting aspect class. The classification accuracy for recognizing aspect categories is on average consistently above 70% F1 score on several datasets — a decent result considering that there were no manual annotation efforts at all [He et al., 2017b]. Based on these findings, we aim to investigate to which extent manual annotation labor can be omitted to find opinion target categories in an unsupervised or weakly supervised manner. The cost of reduced accuracy is small compared to the advantage of being able to explore new social media domains with low effort and high coherence.

2.1.4 Traditional Unsupervised Methods

Historically, unsupervised text clustering as means for opinion mining has always been a widely applied technique since the appearance of traditional topic modeling. These kinds of methods, such as, latent semantic indexing (LSI) by Hofmann [1999], latent Dirichlet allocation (LDA) by Blei et al. [2003], or non-negative matrix factorization (NMF) by Kim et al. [2007], have been used successfully to explore the discussed themes and their related sentiments in social media since 2007 [Mei et al., 2007, Denecke et al., 2009]. Until today, these traditional NLP techniques are used in many opinion mining domains on texts produced by users, customers, clients, et cetera. However, they still have several technical difficulties and limitations. They require to hold all co-occurrence information within memory during their calculation, making it unsuitable for today's large-scale datasets. Using them to form semantically coherent clusters on short texts and specifically sentences or clauses is still problematic [He et al., 2017b]. Particularly coherence is essential for fine-grained information, such as, entities and respective attributes. Moreover, it is difficult to apply transfer learning, and cross-lingual usage is hard to achieve without significant manual intervention for translation. Despite all these limitations, there is a lot of recent research activity addressing the automated analysis of satisfaction from students with their teaching courses at universities [Grönberg, 2020, Nasim et al., 2017], which is described as follows in one of our previous works:

> *"The denoted methods can be applied on the text answers of many course evaluation questionnaires at once to harmonize their outcome and to make them comparable [Hujala et al., 2020, Gottipati et al., 2017]. This has potential for new applications in educational marketing, e.g., comparing courses on a faculty or university level to improve the overall teaching quality, and giving insights about the perception towards an institution [Srinivas and Rajendran, 2019]. Furthermore, the methods are used in automated tools such as Palaute [Grönberg et al., 2020, Grönberg, 2020] and SMF [Nitin et al., 2015], visualizing the underlying topic distribution with NMF or LDA and students' sentiment [Cunningham-Nelson et al., 2019, Gottipati et al., 2017]. The topic distributions in free-text comments correlate significantly with Likert scale answers [Hujala et al., 2020], demonstrating that topic models are able to depict correct distributions of opinions which have been explicitly being asked for in*

separate questions."

— Hagerer et al. [2021b]

This shows that, despite their limitations, traditional unsupervised topic modeling methods can produce meaningful distributions about topic-related opinions. These distributions are able to reflect true sentiments towards experiences, which were encountered by polled individuals.

2.1.5 Opinion Mining as Superordinate Discipline

From the previous explanations about unsupervised and weakly supervised methods, we see that topic modeling as well as unsupervised aspect extraction are able to automatically find opinion targets, such as, topics and aspects, on large scale datasets of user-generated texts. This is helpful, because for humans it is too time-consuming to read and structure that much data.

The methods find new, previously unknown information in terms of *patterns* in unstructured texts. Words occurring in similar contexts form own, distinct topics, which are discovered algorithmically. That principle leads to insights about opinions, including, opinion targets and sentiments. So, insights about *opinions* are *mined* from texts without providing any prior knowledge to the algorithms, i.e., it is a means to explore and inform about opinions from a customer domain. This is how we refer to *opinion mining*, since it also encompasses steps that need to be carried out before actual sentiment analysis to gain domain-specific insights. It supports a text miner to inform herself about a domain of interest, leading to knowledge about the space of possible opinion targets for actual sentiment analysis.

2.2 Qualitative Content Analysis

We refer to computational opinion mining not only for sentiment classification, but also *to find which are the existing opinion targets in the first place*. It is necessary to explore textual data to find out which set of coherent entities and aspects actually occur in the given social media comments. This is a difficile, separate task, since the same opinion targets are often formulated differently from each user. For example, an entity might have various names with many possible ways to (mis-) spell them, not to speak of describing it implicitly instead of mentioning it directly. This is amplified by the fact that a situation is experienced and thus expressed in various different ways. For instance, a restaurant has different tables, waiters, and chefs, of which all can contribute to different experiences with different perceptions of the same thing, e.g., the same dish at the same restaurant. The focus within these experiences is potentially manifold, leading to a plethora of possible formulations. The context in which the formulations are expressed, for example, open-ended forum discussions versus reviews, can further impact the variety of discussed aspects. Grouping the numerous kinds of formulations unambiguously to one distinct concept on a limited dataset is one of the problems which is solved manually without algorithmic support by qualitative research methodologies, such as, grounded theory and content analysis. The former explains the theory and the conceptual framework, the latter is its manual application to textual contents, e.g., subjective consumer experiences in a written form. The following paragraphs, firstly, explain different implementations of this qualitative research method, and, secondly, show how it is connected to opinion mining.

2.2.1 Inductive Content Analysis

The most common grounded theory method is called *inductive content analysis*, which is depicted in Figure 2.3. According to Gorra [2007], it contains the following steps: open coding, focused coding, building categories, and deriving theories. During open coding, single words or statements are tagged as phenomena, for example, events, objects, persons, actions, or ideas, which are encountered by the subject. Thereafter, longer text parts are marked with focused codes as the smallest common denominator of the phenomena. These codes can be grouped to categories and sub-categories. Eventually, theories are derived to explain how categories appear and how they relate to each other in the dataset.

In the following, we outline an example of that process given by a content study conducted by our collaborators:

> *"This exploratory study aims at building an inventory of consumer beliefs about organic food. To reach this objective, we conducted a content analysis of online comments about organic food posted on news websites and forums in German-speaking countries (n = 1094) and the United States (n = 1069). The main result of this study is a comprehensive category system of 65 organic food beliefs and their relative frequencies. "*
>
> — Danner and Menapace [2020b, p. 1]

Table 2.2 shows the main themes, sub-themes, and a small selection of example belief statements. Main themes are the most abstract categories, containing several subthemes, which in turn contain several belief statements. The latter are focused codes, which are assigned to social media comments. If a comment expresses the respective

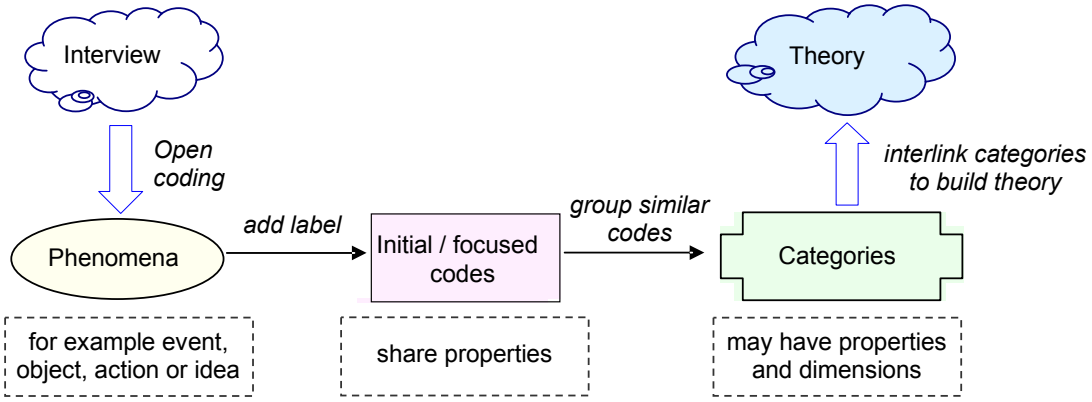


Figure 2.3 Coding process in the grounded theory framework. The *interview* represents an original textual corpus. In *open coding*, single words or statements are tagged as *phenomena*. Thereafter, longer text parts are marked with *focused codes* as the smallest common denominator of the phenomena. These codes can be grouped to *categories*. Eventually, theories are derived to explain how categories relate to each other. Taken from Gorra, 2007.

consumer belief, it gets tagged with that belief. As to deriving theories, the authors derive one out of many possible conclusions as follows: *"the findings in Table 2.2 show that authenticity of organic food products depends not merely on the organic label; commenters sought further authenticity cues originating, e.g., from origin, and other factors"* [Danner and Menapace, 2020b, p. 9]. This observation is more the case for consumers from the USA than from German-speaking regions. The development and marketing of organic products can benefit by considering these interests from consumers located in different market regions accordingly. This short example shows how a fine-grained labeling system for opinions from a qualitative content study can give meaningful theoretical insights into a domain of interest, e.g., how and why needs of consumers of organic food products vary in different market regions.

2.2.2 Directed Content Analysis

When predefined codes are assigned to the texts of a new dataset, Hsieh and Shannon [2005] refer to it as *directed content analysis*. In that case, the codes stem from the labeling system of a previous study, e.g., an inductive content analysis. This paradigm corresponds to supervised classification, where a machine learning (ML) algorithm correlates textual features with the predefined codes and assigns them accordingly to a new corpus — see Table 2.3.

A directed content analysis can generate several findings. Firstly, a domain expert might find that the predefined coding system is not sufficient either in the quantity of the available codes or in their definition. In that case, new codes would be added or existing code definitions would be adapted. Thus, the outcome would be that the existing theoretical model is expanded.

Secondly, *"the findings from a directed content analysis offer supporting and nonsupporting evidence for a theory"* [Hsieh and Shannon, 2005, p. 7]. For instance, consider text examples labeled with a specific code, which are drawn from a new corpus and from the old, original corpus. If the texts from the new corpus have the same or different meanings

Main Theme	Sub-Themes	Example Belief Statements
Product	Food safety, Price, Healthiness, Taste, Nutritional value, GMO, Quality, Naturalness, Availability	Organic products taste good or better than conventional products; Organic products are easily available
Food System	System integrity, Food security, Production scale, Farmer welfare	I disapprove the large-scale organic industry; Organic food is profitable
Authenticity	Organic labels, Organic labels, Retail outlet/brand, Product category, Packaging	Organic products of local origin are more organic; Organic labels cannot be trusted
Production	Environment, Animal welfare, Biodiversity, Working conditions	Organic farming uses no or less chemicals compared to conventional farming; Conventional farming sufficiently protects the environment

Table 2.2 Coding hierarchy from a manual content analysis implemented by our collaborators Danner and Menapace [2020b] on social media comments about organic food consumption.

and connotations as the texts from the old, original corpus, this serves as supporting or contradicting evidence for the pre-existing theory.

Lastly, a directed content analysis is the obvious choice to quantify a given theory or model on another domain of interest. Here, the evidence is supposed to be descriptive, i.e., showing the distribution of codes on a new corpus. If the distribution is similar to the distribution from the old, original corpus, then a given theory might apply there, too, or it would not otherwise. This approach is also referred to as *quantitative content analysis*.

In a summary, *"the main strength of a directed approach to content analysis is that existing theory can be supported and extended"* [Hsieh and Shannon, 2005, p. 8]. However, researchers approach the data with an informed but, nonetheless, potentially strong bias. For instance, researchers might be more likely to find evidence that is supportive rather than non-supportive of a theory.

2.2.3 Summative Content Analysis

"Typically, a study using a summative approach to qualitative content analysis starts with identifying and quantifying the occurrences of certain keywords or concepts in text with the purpose of understanding and exploring their contextual usage. [. . .] Frequency counts for each identified term are calculated, with source or speaker also identified. [. . .] It allows for interpretation of the context associated with the use of the word or phrase" [Hsieh and Shannon, 2005, p. 8]. An example study about sustainability in retailing research and industry is given by [Wiese et al., 2012]. They use the following terms as keywords: sustainability, environment, carbon footprint, fair trade, eco-friendly, green, organic. The occurrences of these keywords are counted in the research and industry magazines from 1980 till 2010. From their various sources, the authors observe that, among others, the retail industry *"has already paid more attention to sustainability"* than the retail research community.

"A summative approach to qualitative content analysis has certain advantages. It is an unobtrusive and nonreactive way to study the phenomenon of interest (Babbie, 1992). It can provide basic insights into how words are actually used" [Hsieh and Shannon, 2005,

Coding approach	Study	Code derivation
Summative	Keywords	Keywords identified before and during analysis
Inductive	Observation	Categories developed during analysis. Unsupervised algorithms: Topic Modeling (i.e., NMF, LDA) and clustering algorithms such as k-means
Directed	Theory	Categories derived from pre-existing theory prior to analysis. & Supervised classification algorithms: Support Vector Machines, Decision Trees and Deep Neural Networks

Table 2.3 Coding differences between the three approaches to content analysis; Taken from Bakharia [2019].

p. 10]. However, the utilized keywords might have ambiguous meanings and varying usages and forms. By default, the method does not take this into account.

2.3 Synthesis: Contributions on Mixed Methods

Are qualitative research procedures and opinion mining compatible to each other?

First, both methodologies are applied to explore textual datasets, i.e., both are obvious candidates to be compared with each other. But also in the way *how* a domain expert or a text miner *uses* these methodologies to explore corpora, there are too many similarities to be ignored. According to Ho Yu et al. [2011, p. 1], the workflow of text mining and grounded theory *"encourage open-mindedness and discourages preconceptions"*. It means both encourage an agile operating method to discover and explore a dataset and its concept space. In both, it is natural to *"add, delete, and revise"* [Ho Yu et al., 2011, p. 1] the explored concepts, i.e., keywords, codes, categories, et cetera, iteratively. In that manner, both methods enable to apply and adapt theory and models to new datasets to change their scope and to validate them. Regarding the scientific criteria, both are comparable with respect to consistency and replicability, too [Ho Yu et al., 2011]. This means that, as long as the focus stays limited to the same dataset, the outcome is determined to a high degree for both methods. All these methodological similarities show that grounded theory as well as opinion mining have many aspects in common, i.e., both are applied on textual data in a similar manner, both give a similar output, and both serve a similar purpose.

Moreover, the recent literature suggests that quantitative and qualitative text analysis methodologies are converging towards each other, mainly due to the gap between big data and the limit of what can be processed by hand. The rapidly growing availability of social media texts has lead to *"an unprecedented proliferation of large unstructured collections of text corpora"* [Eickhoff and Wieneke, 2018, p. 1]. However, as classical qualitative analysis approaches are performed manually on a small and unrepresentative number of samples, the gap between big data and explanatory theories is widening. Thus, explaining human opinions on large-scale social media data is a challenge, which can only be solved with computer-aided procedures. For that purpose, an *"alternative to qualitative coding for textual analysis is given by quantitative text mining methods"* [Eickhoff and Wieneke, 2018, p. 1]. This has the advantage of explaining data which are more representative and less biased, e.g., by drawing *"a statistical representative subset from the population of all documents, or even the full corpus"* [Wiedemann, 2013, p. 339].

But how can quantitative outcomes, e.g., by classification or clustering, be plausible for the manual grounded theory process? Its requirements in regard to depth, meaning, and coherence are high, and thus keeping the human out of the loop hardly appears conceivable. A domain expert is necessary to interpret quantitative findings *"in the light of existing theories, and lead to their adaptation, or the formulation of new ones"* [Baker et al., 2008, p. 296]. Algorithms ought to provide reasonable means to support that process when there is too much data to be investigated manually. So-called *mixed method* approaches, see for instance Figure 2.4, are necessary to connect a domain expert with the outcomes from text mining, such that one finds sufficient qualitative evidence to trust and interpret the quantitative findings [Palinkas et al., 2019]. The notion of "meaning" is crucial in this process. First and foremost, texts and text elements should be explained based on their *meaning*, not based on mere word frequency counts [Wiedemann, 2013].

The new generation of natural language processing (NLP) algorithms are, more than ever before, supposed to provide exactly this — contextualized, higher-level semantic rep-

resentations of texts. These representations are embeddings from sophisticated deep neural network architectures pre-trained on massive amounts of data, such that textual elements are grouped based on their context and meaning. This has high potential to *"bridge the gap between qualitative and quantitative data analysis"*, since they are *"able to dig into "latent" meaning rather than counting surface observations"* [Wiedemann, 2013, p. 354]. As we aim to show in this thesis, *"they are able to keep the link between the qualitative input data and their quantified results"* [Wiedemann, 2013, p. 354]. The ultimate goal hereby is to *"enable the researcher"* and the well-disposed reader *"to build confidence in this approach"*. It shall be shown that *"distant and close reading may interact fruitfully and quantitative text analysis may keep a qualitative quality"* [Wiedemann, 2013, p. 354].

The following paragraphs give an outline of how opinion mining complements qualitative content analyses. For each type of content analysis, i.e., summative, inductive, and directed, there are different technical methods used in practice. Bakharia [2019] gives an overview of it, which is shown in Table 2.3, and we elaborate on it in the following paragraphs. It is described how this is helpful for domain experts. We give an outlook on how these practices can be even more improved in the future with state-of-the-art NLP.

2.3.1 Complementing Inductive Content Analysis

Bakharia [2019] shows that the categories from unsupervised topic modeling correspond to the ones from human coders in inductive content studies. Specifically coarse, abstract categories are similar, whereas there are differences on a more fine-grained level. The author concludes that the derivation of high-level categories can be automatized for inductive content studies. This has high potential on corpora which are too large to be processed manually. Big data also leads to a better representation for data-driven algorithms, which improves topic coherence, interpretability, and statistical significance. Thus, representative categories are found effectively and efficiently. Unsupervised text mining generates them automatically without human intervention. It is already widely used in practice to inform, accelerate, and extend the grounded theory process of domain experts on their domains of interest.

The examples illustrate how traditional topic modeling is used for inductive content analyses in many domains of interest. However, they utilized traditional NLP methods, which does not represent the state-of-the-art in modern, unsupervised NLP. There are limitations in regard to coherence on short texts, memory efficiency, data size, domain knowledge transfer, and translatability, which is outlined in the previous section about Traditional Unsupervised Methods. Modern NLP methods for opinion mining remedy these drawbacks. Thus, they offer high potential to improve the opinion mining process of domain experts during inductive content analysis. For illustration, we give the following examples from our own research, which are also attached to this thesis.

In Danner et al. [2020], we start our journey by correlating fine-grained expert codes with features generated from a deep, pre-trained neural network for natural language understanding called Universal Sentence Encoder (USE). We find significant correlations, and by including unlabeled data as well, we conclude that these semantic features also *"have the potential to serve for the analysis of larger data sets"*. We exemplify that sample size can be scaled up *"while maintaining the detail of class systems provided by qualitative content analyses"*. Since the pre-training data stems from various tasks and domains,

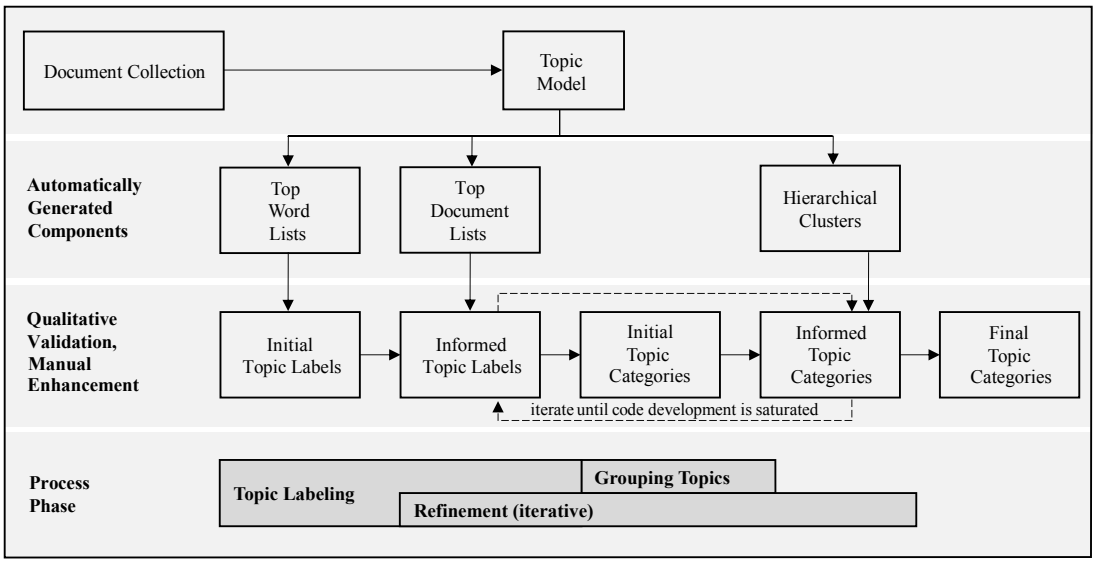


Figure 2.4 Mixed method approach similar to the one proposed by Eickhoff and Wieneke, 2018, p. 6. A topic model is applied on a document collection. For each topic, a top word list and a top document list is created automatically. The top words are used for manual initial topic labeling, and the top documents are used to refine and contextualize the manually assigned topic labels. Topics are then grouped into categories. Hierarchical clustering of topics validates the topic categories, leading to further refinement of the categories. Refinement of topic labels and topic categories can be repeated until there is no information gain.

it carries the same potential on other data than our own specific use case about organic food consumption behavior.

In Hagerer et al. [2021d], we elaborate on that technique. We derive fixed-dimensional, explainable representations for newspaper articles and readers' comments jointly in multiple languages using a single, cross-lingually pre-trained, deep neural network for sentence embeddings. By doing so, the distribution of themes in articles and comments and their mutual influence can be shown and compared between two different cultural areas on the same domain of interest using one single, integrated model. The technique reduces programming, translation, and pre-processing efforts to a minimum compared to traditional topic modeling. At the same time, it provides a high number of consistent and interpretable multi-lingual themes. We enrich the themes with topic-related multi-lingual sentiment analysis. In the bilingual (German English) case study, exemplary newspaper articles and readers' comments are chosen, and their topic-related sentiments are depicted. The distributions appear to be consistent with the comprising opinions. Furthermore, the topics reveal a consistent, hierarchical, and semantic structure for an increasing number of topics. Thus, the technique can be considered as a viable cross-lingual topic model for the mixed method approach formulated by Eickhoff and Wieneke [2018], see Figure 2.4.

In Danner et al. [2022], we leverage the same cross-lingual topic model to support an actual domain expert in a content study. Its research question is *"what are the agenda-setting effects between the news media and its audience regarding organic food"*. The text mining study shows how the newspaper articles influence the comments with regard to the discussed topics, but not the other way round. The method, i.e., cross-lagged correlations based on topic saliency, also depicts events at which media and public attention are di-

verging. It is concluded that *"the news media drives public opinion on organic food in the US and Germany by determining the discussion topics"* about organic food consumption and ought to be *"considered by marketers and policymakers"*.

The previous works are based on clustering pre-trained sentence representations, assigning each sentence to a topic and a related sentiment. This follows the intuition that one sentence contains one dominant theme and, potentially, one sentiment related to it. The previous findings support this intuition by displaying the chosen textual contents correctly. Even though both case studies provide convincing results and findings, we did not perform thorough, objective NLP measurements to evaluate our approach in these studies. Therefore, we aim to show that semantic sentence similarity based on pre-trained sentence embeddings matches the standard of existing NLP metrics. So, in Hagerer et al. [2020a], we use these embeddings to evaluate state-of-the-art abstractive product review summarization. The results show advantages of our proposed sentence embedding technique over the commonly used ROUGE scores, such as, higher correlation with human judgement and correct modeling of synonyms due to semantic similarity. This is supposed to underline the technical feasibility of the approach.

Elaborating further on sentence clustering, in Hagerer et al. [2021a], we demonstrate a prototype of the SocialVisTUM interactive visualization tool for opinion mining. The underlying model provides the state-of-the-art coherence for mining opinion targets on sentence and sub-sentence level [He et al., 2017a, Angelidis and Lapata, 2018a, Luo et al., 2019a]. It is combined with correlated topic modeling, automated hyperparameter optimization, sentiment analysis, graph clustering, and graph visualization. The resulting toolkit depicts the distribution of target-specific sentiments on a dataset of choice. It can be used by text miners to implement a mixed method approach, such as the one depicted in Figure 2.4. In the paper, we use it to confirm results of a separate content study, which was carried out by a domain expert.

In Hagerer et al. [2021b], we applied sentiment-related topic modeling on students' comments of evaluation questionnaires from several teaching courses at our university. The qualitative study derives theories about how the introduction of autograding into several programming courses is consistent with significant changes of course satisfaction according to the opinions of students. The opinion mining results show a close, meaningful relation with the actual polling numbers, substantiating the findings. These results and the attached literature review show how opinion mining has the potential to generate representative and standardized insights to expressed experiences. We conducted the study to highlight the relevance of opinion mining for mixed method approaches as a complementary means for polling-based opinion research.

In Stappen et al. [2021], we propose a new topic modeling method for large-scale video transcripts. The transcripts are taken from the MuSe-CaR corpus provided by Stappen et al. [2021c], containing YouTube videos from drivers about car reviews. In the literature, there are few successful studies about topic modeling methods on that type of data. The presented method, which is based on word2vec, semantic similarities, and graph clustering, outperforms several other approaches. It enables the exploration of discussed themes in large-scale video databases. Inductive content studies can benefit from that considering that this type of transcribed video material is increasingly available for a plethora of domains of interest.

2.3.2 Complementing Directed Content Analysis

Technically, applying a directed content analysis (DCA) means that a human expert applies pre-existing codes, e.g., from a previous study or theory, on a new textual corpus. When this is performed automatically with algorithms by a computer, it is equivalent to document or sentence classification — see Table 2.3. When the algorithm, in turn, is informed by the codings and word statistics of a previously coded corpus, this is what it means to train a machine learning (ML) model on a training and development set for supervised classification. Well-known algorithms are support vector machines, logistic regression, decision trees, and artificial neural networks; commonly used features are bag-of-words, term-frequency inverse-document-frequency (tf-idf), or pre-trained embeddings, such as, word2vec.

Regarding the labels, it is common practice in qualitative research disciplines, including mixed methods, that coding is performed by humans adhering strictly to the grounded theory protocol. Accurate cognizance of the protocol and thorough investigation of the textual data are important requirements to establish consistent qualitative descriptions and correct explanations of the data in terms of annotations. We refer to a person who gains and implements that expertise as domain expert. The idea of domain expertise is that a domain expert provides high quality annotations. Utilizing these annotations to train ML algorithms is wide-spread in mixed methods DCA research.

There are, however, several technical limitations to expert annotations, which are not rigorously examined by the DCA mixed method literature. First of all, it takes a long time to produce expert annotations, because the grounded theory process is complex. A deep understanding of the hidden structure in the given texts needs to be established, which can usually only accomplished by one or two persons accumulating that knowledge — see Figure 2.5 for instance. This limits the number of how many samples can be annotated. At the same time, the domain expert draws many detailed observations from the data, leading to many fine-grained classes. See, for example, the exemplified belief statements in Table 2.2: There are more than 60 belief statements annotated to merely $1,000$ text samples. Also in the related literature, a general tendency towards many classes versus a few annotated samples can be observed. This is opposed to the technical requirements of machine learning, where many samples per class are highly beneficial if not mandatory to achieve meaningful classification accuracy. This means, expert annotations are conflicting with machine learning requirements, which introduces many problems for mixed methods in DCA, first and foremost that only coarse high-level document classes can be predicted reliably. However, the fine-grained annotations lie idle, and there is a lack of thorough analysis of how these can be leveraged. Reasonable trade-off strategies need to be found and validated systematically.

Thus, in Hagerer et al. [2021c], we investigate how fine-grained expert labels from content studies are suitable for text classification using state-of-the-art ML. Two labeling dimensions are used for training and compared with each other: many fine-grained versus few coarse-grained expert classes; expert-defined categories versus hierarchically clustered classes versus completely unsupervised clusters as categories. The study shows that fine-grained expert annotations of belief statements are problematic for classification, whereas automatically generated high-level classes based on clustering improve classification accuracy significantly. Furthermore, the high number of fine-grained expert classes on a small number of samples leads to very low accuracy, which makes machine learning

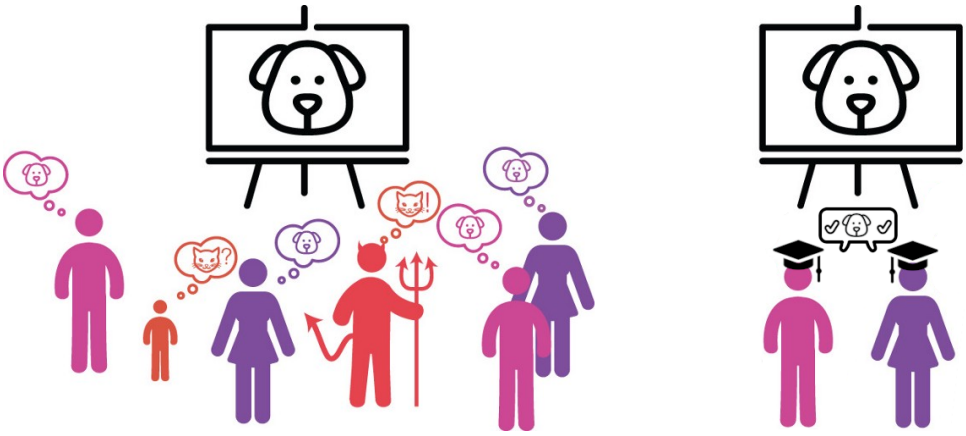


Figure 2.5 Left: Illustration of a crowdsourcing task. Crowdworkers are recognizing the content of the shown image, i.e., a dog. Most subjects would label it correctly. Some of them, however, give a wrong estimate due to being inexperienced or neglectful. Overcoming labeling noise can be achieved, e.g., by modelling each annotator separately.
Right: Expert-based coding. Two domain experts are agreeing on a correct label to describe the sample image. This can be done to agree on a common position for the overall annotation task or for specific samples only. An adjustment of a common understanding can be made before, in between, during, or after the experts perform their coding. The process is based on grounded theory and ensures high consistence and inter-rater agreement.
Image inspired by Verhaar [2020].

unsuitable. Fewer, rather abstract categories are suitable, and informing them with expert knowledge appears feasible.

So, mining opinions from texts with detailed, qualitative expert annotations remains a crucial task to scale expert domain knowledge, whereas this fine-grained information is difficult to classify. Opinions can be given on sub-sentence, sentence, and paragraph level. Furthermore, they can be formulated using numerous relevant and nuanced variations. A categorization leads to a reduction of the number of opinions, resulting in a loss of valuable details given by the analysis of domain experts. This means that, if the level of detail from opinions should remain high, the number of annotated samples has to be increased. Otherwise, reproducing and scaling the results with predictive machine learning does not appear to be achievable. Therefore, we draw inspiration from crowdsourcing research. Snow et al. [2008] demonstrate in five different NLP tasks that *"using non-expert labels for training machine learning algorithms can be as effective as using gold standard annotations from experts"*. The tasks in that study are complex and semantically rich and include affective text analysis, word similarity and disambiguation, and textual entailment. All of them require careful reading and understanding from a human to be resolved correctly, and linguistic expert knowledge is beneficial to resolve tasks. In that respect, these tasks share similarities with qualitative content studies. In affective text analysis, for instance, it is generally not always immediately apparent to which degree an expressed emotion is positive or negative. Emotion is a multi-layered and subjective feature, for which connotations and contextual information need to be considered for a proper qualitative analysis. For content studies, grounded theory appears to be an appropriate methodology to resolve that task.

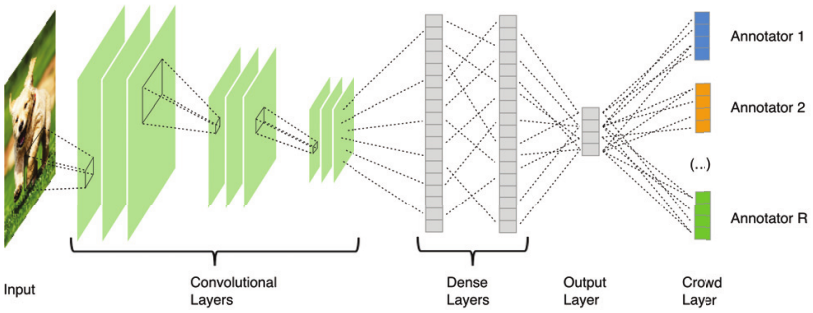


Figure 2.6 Example for a deep learning crowdsourcing convolutional neural network, here for image recognition. Each annotator is modeled separately, but below that there is a shared hidden layer serving as a form of common, latent truth. The concept is analogous to the latent truth network proposed by Zeng et al. [2018]. Taken from Rodrigues and Pereira [2018, Figure 1].

The authors show that annotations for this type of tasks can be equally well executed by non-experts. *Equally well* means that classification accuracy when predicting expert labels is at least equal when training is performed on many non-expert annotations instead of on few expert annotations. The authors state that roughly at least four non-expert labels equal to one expert label for training. We conclude that many non-expert labels have the potential to improve classification on tasks which would presumably require an actual expert to understand the texts. Furthermore, they are cheaper to obtain, and less time is needed for the researcher to implement the annotation process. Thus, the research question appears valid if crowdsourcing is a feasible means, too, for fine-grained opinion mining to support qualitative content studies.

Therefore, in Hagerer et al. [2020b], we gathered this type of non-expert annotations for the task of detailed, domain-specific, aspect-based sentiment analysis. To overcome the problem of too few annotations on too many fine-grained opinion classes, we implement a singly labeled crowdsourcing protocol utilizing many non-expert annotations on as many text samples as possible. For predictive machine learning, we use an artificial, deep, recursive neural network trained on multiple aspect-based sentiment analysis (ABSA) corpora. It is based on state-of-the-art transfer learning techniques to overcome the catastrophic forgetting effect Kirkpatrick et al. [2017], which inheres traditional transfer learning approaches. So, by leveraging complex deep learning, we maximize the leverage of pre-existing knowledge derived from several ABSA datasets. The study shows that the approach called progressive neural networks indeed gives consistent improvements for classification on varying tasks and domains. However, despite all efforts with regard to deep learning and crowdsourcing, classification accuracies on our own dataset are stuck at 17% and are too low to be used for qualitative DCA research in practice.

As the probable cause, we have identified annotator noise and biases, i.e., inconsistent annotations due to the various different backgrounds and attitudes of the crowd workers. The nature of the problem is depicted in Figure 2.5 and is explained by our related work as follows:

> *"The varied personal backgrounds of crowd workers often lead to annotator biases that affect the overall accuracy of the models. Several works have previously ranked crowd workers [Hovy et al., 2013, Whitehill et al., 2009, Yan et al., 2010], clustered annotators [Peldszus and Stede, 2013], captured*

> *sources of bias [Wauthier and Jordan, 2011] or modeled the varying difficulty of the annotation tasks [Carpenter, 2008, Whitehill et al., 2009, Welinder et al., 2010] allowing for the elimination of unreliable labels and the improvement of the model predictions."*
>
> — Hagerer et al. [2021e, p. 2]

These works analyze the behavior of each annotator to discover who improves the annotation quality and who impairs it. By excluding or correcting bad annotators (see Figure 2.5), this is supposed to improve the machine learning model. However, the previously cited works all rely on data of which every sample is annotated at least two or more times, i.e., multi-labeled crowdsourcing annotations. Only then it is possible to see which annotator correlates more with the other annotators and who correlates less. This is an important feature to determine the labeling quality of each annotator and the overall inter-rater agreement — Figure 2.5 illustrates that aspect.

In opinion mining, however, the multi-labeling requirement is oftentimes not met by the naturally available crowdsourcing datasets, since they are often singly labeled by default. This is summarized by one of our works as follows:

> *"However, bias modeling when every data point is annotated by only one person, hereafter called singly labeled crowdsourcing, poses a rather specific and difficult challenge. It is in particular relevant for sentiment analysis, where singly labeled crowdsourced datasets are prevalent. This is due to data from the social web, which is annotated by the data creators themselves (e.g., product ratings given by consumers and reviewers, authors note) [. . .]. While the outlook for such forms of data is promising, end-to-end approaches have not yet been fully explored on these types of crowdsourcing applications."*
>
> — Hagerer et al. [2021e, p. 1]

In Hagerer et al. [2021e], we propose an end-to-end deep learning architecture, which enables unbiased training on single label crowdsourcing datasets. The architecture is similar as the one shown in Figure 2.6, except for the task being image classification instead of opinion mining. End-to-end in crowdsourcing means that there is no ground truth derived from the annotator labels, on which the model is trained afterwards, but the model is trained solely on the annotator labels directly together with the input features — see Figure 2.7. Despite the fact that each sample is only annotated one time by only one annotator, the neural network creates a consistent, reliable, and unbiased ground truth. The shared latent truth layer below each annotator's bias models their smallest common denominator, which is unbiased labeling knowledge from all annotators. The bias here is encapsulated by the bias matrices, which improves classification accuracy significantly. Beyond that, we show theoretically and empirically that the annotator bias is converging to the actual confusion matrix, making it a 100% equivalent to the real bias. This is insofar remarkable, as the related work about crowdsourcing end-to-end learning shows a clear mismatch between bias and confusion matrices — see Rodrigues and Pereira [2018, figure 3] and Zeng et al. [2018, figure 5].

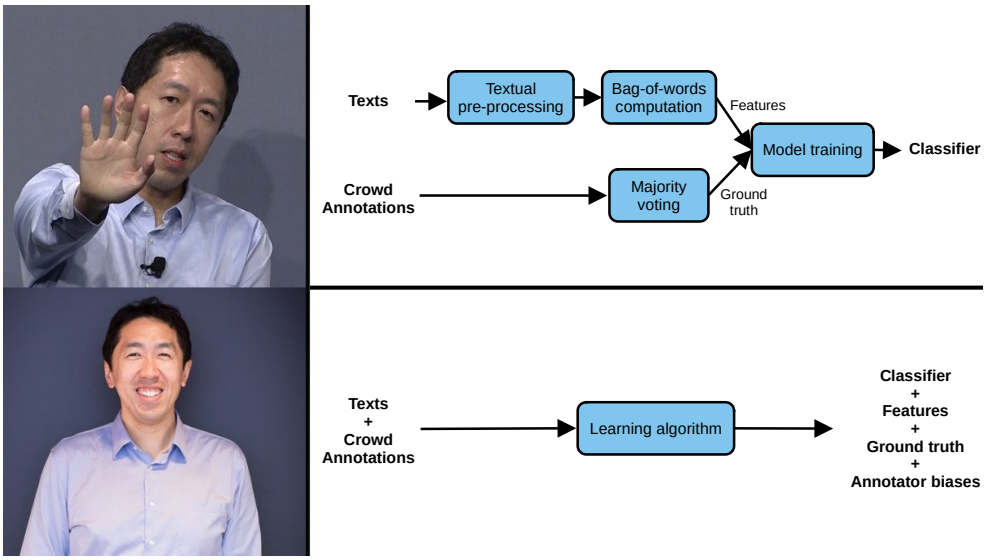


Figure 2.7 Top: Procedure for standard or traditional crowdsourcing modeling. The ground truth is calculated from all crowdsourced annotations prior to model training. Bottom: Procedure for end-to-end learning on crowdsourcing. The ground truth is learned during model training. The annotator biases are calculated as a byproduct. Illustration inspired by DeepLearning.AI [2018].

3 Data Domains

This section aims at presenting the datasets used for the publications of this thesis. Introductory, we establish a system to differentiate and categorize opinion mining datasets. Therefore, the terminology, the commonalities, and the differences of all our datasets are described. We aim to show which kind of data are actually suitable for opinion mining and which aspects they have in common, first and foremost, textual comments in which individuals express their viewpoints about a common theme. Such characteristics are also relevant regarding if opinion mining methodologies can be applied at all.

Secondly, the term domain of interest for opinion mining is defined by explaining the dimensions in which these domains differ among each other, e.g., if opinions are expressed publicly or privately, if the comments are part of a longer discussion or only a single message without any reactions, if there is a single, known recipient of a message or many unknown, and so on. These distinguishing aspects are relevant for a text miner to decide which text mining method can or should be chosen based on the data formats. Further, it is relevant for the domain expert to determine which data source should be chosen for the data collection to answer the given research questions and to match the own domain interest. This step is of high importance, as it is decisive for the success of a research task, to know which type of data fit and can actually answer the formulated research questions. Since domain experts also work qualitatively, coding plays a key role in the respective content-based research, for inductive content analysis as well as for scaling directed content analysis using machine learning. One part of this thesis is focussed on the pitfalls and limitations of these type of annotations for machine learning. In that regard, we differentiate annotation data, too, for instance, if they are given by experts or amateurs, the level of annotation details and abstraction, what is actually being annotated, and so forth. This is helpful to estimate the chances and risks before implementing an elaborate and costly coding scheme for text mining procedures.

Lastly, we introduce our own datasets, classifying them into the denoted domains and annotation types. Concrete aspects, such as, discussed contents, participants, discourse format, et cetera, describe the respective dataset, its specific domain of interest, and its context in detail. Corresponding descriptive statistics give an overview of the distributions of themes and opinions. This information shall build a base to understand the attached research studies based on their application domains. It completes the picture of this thesis by adding the research objectives and challenges from the perspective of the actual opinion mining application and data domains. The overall research is supposed to appear in a different light when it is not merely seen from a methodology perspective as in section 2.3, but also from the viewpoint of how it leads to actual insights into behavior from concrete customers and users. This shall satisfy the interest of domain experts of these specific domain of interests as well as to inspire the application of innovative text mining methods for qualitative research studies in a mixed methods manner.

3.1 Typization of Opinion Mining Data

When differentiating datasets for opinion mining, it becomes immediately clear that the source of the data plays a key role. Thus, we start by categorizing sources according to their type of origin. Subsequently, we discuss annotations in a separate section independently of the sources, since we see various ways and objectives regarding how textual data can be annotated with sentiments and opinion targets.

3.1.1 Differentiation of Sources

The choice of the data source is directly relevant for the research questions of a qualitative content analysis, but also for technical tasks, such as, data crawling, pre-processing, natural language processing (NLP) methods, and many more. Thus, these aspects and constraints should be studied thoroughly for the sake of choosing the right data source *before* implementing steps of actual data collection. There are many ways how data sources can be differentiated. To delimit and focus the scope of possible domains, we start by describing the commonalities of all considered datasets. In short terms, they all contain texts written by independent individuals online in response to themes, products, and services. With respect to differences among datasets, it can be seen that the sources differ among each other as to the type of online medium and environment where individuals express their viewpoints. For the scope of this thesis, three types of sources are investigated: dialog-oriented social media, public product reviews, and evaluation feedback questionnaires. The source types are discriminated based on the following criterions: types of senders and recipients of messages, communication motives, setting, anonymity, privacy, publicity, findability, topic diversity, discourse and interaction, textual structure, data formats and amounts, multilingualism, and language use. The relevance and impact of each of these aspects on the qualitative research questions and the technical implications are analyzed.

Commonalities of Datasets

In principle, opinion mining corpora have in common that individuals express their viewpoints about a common theme. This information is given as texts which are either written or spoken (and transcribed afterwards) by customers (which we use as umbrella term in the following), users, testers, voters, and other subjects reporting about their experience, satisfaction, and attitude towards entities, including but not limited to, products and services.

Motivation for Opinion Mining Data Collection Overall, there is a common reason why such data is collected and analyzed. Institutions providing products and services have a natural interest to maintain and improve their own reputation and the reputation from their offering. This is necessary for them in order to keep their existing and gain new customers [John, 2003]. Therefore, it is crucial to gather evidence of many individual customer experiences and satisfactions, to quantify them, and to derive a representative picture of the overall mindset [Farris et al., 2010]. Only then, it is possible to address the customer needs appropriately in a targeted manner for future products and services design. In the industry, this is achieved, for example, by implementing a voice of the customer process [Gaskin et al., 2010].

While questionnaires with Likert scale answers are a common way to collect that data [Wirtz and Bateson, 1995], there are several reasons to work with textual data. Firstly, textual data already exists naturally in large amounts online in the public sphere, e.g., on social media or product review platforms. Collecting that data is a low-hanging fruit, since the task of downloading it is technically straightforward. Secondly, the literature provides a plethora of evaluated opinion mining methods to depict opinion distributions [Ravi and Ravi, 2015, Yue et al., 2019, Vinodhini and Chandrasekaran, 2012]. It has been shown that these textual features from comments show meaningful and statistically significant correlations with Likert scale answers provided along with open-ended answers [Hujala et al., 2020], [Hagerer et al., 2021b]. Thus, opinion mining is able to reveal a meaningful structure of opinions in an unstructured textual corpus, thus resembling the image in a similar manner as if it would be given by opinion polls. Thirdly, subjects can express concepts and ideas in textual comments which are not being asked for in Likert scale questions. This occurs since the inquirer might not have a preconceived notion when formulating the actual questions. So, the researcher is not unlikely to formulate Likert scale questions which are not targeting aspects and root causes relevant for the overall satisfaction. There can be various reasons for this to happen, such as, errors and problems of an already sold product with problems yet unknown to the provider, or standardized default questionnaires with a generally incomplete set of questions. Individuals, however, can make use of the opportunity to express this information in open-ended texts, such that these aspects become apparent in the data as repeating, detectable patterns and, thus, measurable with according NLP methods.

It can be concluded not only that there is a lot of textual data generated by customers, but also that there is sufficient reason to collect and examine it thoroughly to understand and address customer needs. We propose this to be the general motivation why opinion mining data is collected per se, regardless of if for qualitative, quantitative, or methodological research purposes.

Data Noise in User-Generated Texts Even though opinion mining data are collected from a broad spectrum of different domains, they oftentimes share the same problems hindering the overall practicability of opinion mining. Most notably, opinion mining deals with texts written by ordinary people, who do not act as professionals. In social media, subjects contribute most of the time for their own sake of entertainment and information exchange and are not paid for their contributions. Moreover, user and customer feedback is oftentimes given voluntarily with little or no direct reward except than influencing future product development and maybe receiving attention from other customers. This results in texts containing a relatively high noise level. Frequently, users make many kinds of errors when typing their texts inattentively, of which typographical or spelling errors are the most prevalent ones. Additionally, special characters are a problem, because they are either used inconsistently, e.g., various kinds of apostrophes and quotation marks due to different keyboard layouts, or creatively, e.g., for sophisticated emojis, hashtags, and other kind of internet slang. Speaking of the latter, creative language use is popular in the world wide web, too, leading to a variety of neologisms. Last but not least, automatically generated transcripts from large-scale video data recorded in in-the-wild conditions have a significant word error rate, leading to hard-to-understand texts passages [Stappen et al., 2021]. Considering these circumstances, it is clear that the noise level in user-generated data is a problem which generally needs to be dealt with by one means or another. We opt for two noise removal methods. First, we identify the most frequent spelling errors, con-

tractions, emoticons, and so forth, and replace them with standardized, human-readable expressions [Hagerer et al., 2021e]. Secondly, we apply stop word removal as proposed by Saif et al. [2014], i.e., removing words which seldom occur to reduce the word noise level and improve classification accuracy. Thirdly, we aim at finding and applying noise robust NLP methods, which are inherently able to filter noise at the word and character level [Hagerer et al., 2021c, Stappen et al., 2021].

Unstructured Nature of User-Generated Texts Another common problem of opinion mining data is that there is generally no summary of all the available opinion statements provided. Since such a text corpus is by definition an unstructured collection of numerous individual opinions, there is no structured overview available. Albeit a plethora of technical NLP methods exist, they all require at least some manual, human intervention beyond implementation to tailor the algorithms for the analysis objectives, including tasks, such as, hyperparameter optimization [Hagerer et al., 2021d], cluster labeling [Hagerer et al., 2021b], or manually adding domain knowledge, e.g., by labeling [Hagerer et al., 2021e]. All these manual tasks, in turn, depend on the data itself, which means that the data must be inspected before or after running the algorithm. The inspection typically incorporates two kinds of analyses. One investigates at least some of the following dataset parameters statistically prior or in addition to NLP: Lengths, structure, and number of comments, vocabulary size, word occurrence distribution including n-grams, number and names of opinion targets, themes, and semantic clusters, sentiment dictionary. Alternatively, a qualitative content analysis, e.g., one of the methods explained in section 2.2, can be carried out to build up domain expertise, which in turn helps as a soft factor to reach the analysis objectives manually. A combination of dataset statistics and qualitative inspection can surely be considered as the gold standard for a successful opinion mining process. This conclusion, indeed, is confirmed and consolidated by most of our studies across all domains.

Lack of Data Last but not least, there is an inherent law in data science related disciplines that (semi-) automated data analysis becomes increasingly significant, robust, and representative with increasing amount of data. Although it might appear redundant to highlight it, it has to be stated that also opinion mining datasets need to have a minimum size in order for statistical computer algorithms to yield meaningful results. While this might appear unproblematic especially for social media domains with big, public data sources, the situation quickly changes when a few delimiting factors, such as, time frame, search keywords, or hashtags, are applied [Danner et al., 2022]. Moreover, many opinion mining domains are not public, e.g., questionnaires [Hagerer et al., 2021b], and thus populated sparsely. In conjunction with the previously mentioned problems of opinion mining data being noisy and unstructured, there should always be allowed for a certain surplus to meet a minimum data amount. Concluding, the risks of data sparsity always need to be taken into consideration for the process of data source selection, since it can affect the whole opinion mining process negatively and nullify the outcomes.

Dialog-Oriented Social Media

Speaking about large-scale data availability, public social media platforms are a well-known and rich source for textual comments from individuals. Text-based online communities, such as, Twitter, Quora, Reddit, et cetera, are steeply growing in their user

base. The denoted platforms show a sustainable growth rate over the last 10 years between 30 – 50% in each of the recent years with currently 300 – 500 million monthly active users [Team, 2021, Hallur, 2020, Mowbray-Allen, 2019]. The explosion of user engagement makes these platforms attractive for opinion mining and qualitative research. They are a tool for online users to engage in discussions about products to buy (24.7%) and to share their opinion about them (23.4%) [Chaffey, 2021, Chen, 2021]. The communities are to a high degree dialog-oriented, meaning that users can publish comments with their opinions, which are visible to the general public and can be answered by all other users. This also holds for comment sections under online newspaper articles, which serve as agenda setter for those discussions [Danner et al., 2022].

Communication Format The comments in these kinds of social media discussions are organized in a thread-based structure. The main article or original post sets the overall frame and topic for the series of answering comments posted below it. Based on that, many users can engage in a discourse about the main article, comment, or question, which is aimed to be answered or related to by the answering comments. These, in turn, can also refer to each other, either by literal quotations or by creating a new conversation branch within the existing conversation tree. Mostly, a thread can be unlimited in its length, i.e., as many comments as desired can be posted under an original post. This has implications, since there might be many aspects and facets which want to be discussed by the social media audience. Especially, controversial topics can raise a lot of attention with a high variety of different opinions expressed in the comments below. Due to being open-ended, social media discussions can actually deviate from the original post topic to search for consensus about other issues [Hagerer et al., 2021d]. Thus, we hypothesize that social media is open to unknown concepts and ideas, making it suitable to explore a widespread and rich entity and aspect space of a common theme. It can be helpful for qualitative research to explore as many available concepts as possible within a domain of interest to see which are all the relevant aspects from the perspective of the users. This is a crucial step in getting an overview and understanding the data domain and generally how to structure the research further with regard to potentially more specific focus. As to machine learning, the availability and amount of textual data from specific data domains makes social media texts suitable for pre-training deep neural networks for natural language understanding [Müller et al., 2020]. These are a popular means to improve a variety of NLP tasks, including, sentiment analysis [Azzouza et al., 2020], named entity recognition [Godin et al., 2015], et cetera.

Data Collection Social media data is popular among researchers, since it is simple to collect it. Many of the popular social media platforms are entirely public and provide open access to their databases, including according APIs and software modules. Data can be searched for by using, for instance, content categories, tags annotated to the content by the creators, search queries based on keywords, and a combination of all of them. The usage rights for well-known public social media platforms are oftentimes liberal, granting the right to use the data for research purposes. However, the situation for newspaper articles and comments is oftentimes the opposite, such that data might not be allowed to be collected even for non-commercial or research purposes. Between both worlds of open and restricted access are semi-public social networks, such as Facebook. There, some parts of the networks are public, for example, public groups or pages, and some are private, such as, private groups or personal profiles. It is possible to receive

allowance to crawl public parts of the network. However, the technical setup of the official API registration and implementation is complicated and involved. Every other access to the network via third party software is controlled and blocked. Legally, there might be risks, e.g., due to privacy issues and local data protection laws. Since Facebook is a platform for friends and acquaintances, many people use their real names and share also other private information by mistake. It is not clear how anonymity can be always preserved. As a consequence, a corpus release based on Facebook (or other semi-public social networks) data must be considered difficult. Concluding, data collection is technically and legally recommendable on public social media, such as Twitter, Reddit, or Quora. The other extreme are often but not always newspaper portals, of which the data are mostly public but with restricted legal access. Semi-public social networks, such as Facebook, are similar to the latter, due to privacy issues.

Reviews About Product and Services

Since many years, online shopping platforms, such as, Amazon, eBay, idealo, but also the general online mail order and online booking business have a rapidly growing market share as opposed to the traditional retailing industry. Modern customers increasingly use the internet to buy products remotely online via home delivery, but also booking travels, hotels, restaurants, and other services is done naturally online nowadays. These trend recently became strongly accelerated due to the worldwide COVID-19 pandemic [Deutsche Welle, 2021]. These websites offer the opportunity for customers to state their feedback about the product or service experience below the product descriptions. Additionally, the customer gives Likert scale ratings of his product satisfaction with the bought product. The reviewers use reviews as a means to give public feedback about the product, reporting about deficiencies, voicing gratitude and appreciation, and giving advice to others [Burke, 2021]. These reviews and ratings serve as recommendation for other potential buyers, and they are an important factor for the product decision. More than 90% of all purchasers read online reviews before deciding to buy or not, and they tend to trust them as much as personal recommendations. Conversion rates are significantly higher of products with more reviews, impacting the sales numbers positively, too [Imaad, 2020]. As a result, product reviews are a driving factor for the online shopping and booking industry, containing salient and relevant information about product and service consumption and satisfaction of many individuals.

Communication Format Reviews are mostly posted below a specific product or service description on web portals where these are offered. These web portals can be booking platforms for services, e.g., restaurants, hotels, or flights, or shopping platforms for products, e.g., Amazon, eBay, idealo, but also normal online shops from catalog companies, such as, Otto, MediaMarkt, Zalando, et cetera. On these platforms, it is normally ensured that only real customers who ordered that product or service are requested and authorized by the platform itself afterwards to post a review about the purchased item. This is done to avoid fake reviews for self-promotion and to maintain trustworthiness and credibility of the user-generated content. It is the key to give a plausible and trusted impression of a product or service, which is meant to increase the value of the comments, the platform, and the sold products and services. Thereby, each single review is assumed or at least aimed at being a reliable and trusted information of a real experience with a specific thing. Under this assumption, *truth* is not the result of an interactive and collective search for consensus

in an open debate. Instead, each review stands for itself, which is why only one review is given at a time without any reaction in terms of answering comments onto that review. As a consequence, there is no direct interaction in terms of an observable conversation performed on that platform as opposed to other social media types. We emphasize this by explicitly denoting those as *dialog-oriented* social media in order to distinguish it from reviews, which are non-dialog-oriented and uni-directional. One isolated review is given at a time which is targeted to only one specific, previously known product or service. This very opinion target is known beforehand and is basically the same for all reviews on the same item. This means that, in contrast to dialog-oriented social media, the general entity is determined to a high degree. However, the *aspect* of the reviewed entity is what is unknown a priori. So, the space of possible topics and entities is limited in reviews, but the space of aspects, attributes, and sub-entities is open for exploration. Figuratively speaking, the resolution of opinion *targets* is increased in reviews. Also, the relation of opinion targets to sentiments and, thus, the sentiments themselves have a higher resolution due to the annotated ratings, which are also stated by the reviewers.

Concluding, reviews about products and services are directed towards known entities with annotated sentiments, which helps with statistical inference and concrete product satisfaction analysis. However, the narrow scope on already existing products and services puts a limitation on which other consumer interests might exist *beyond* what can be bought. In qualitative studies, this might be advantageous to investigate incremental product enhancement but potentially unfavorable for the from-scratch design of new products and services. For machine learning, one might consider to use masses of review texts to train sentiment analysis classifiers and fine-grained aspect extraction models. How this can be used for transfer learning onto social media domains depends on the individual case and on how well the data domains fit each other.

Data Collection The review data from shopping and booking platforms is usually publicly accessible, and it is technically feasible to download and parse it using automatized download algorithms. The legal situation, however, is not always clear in advance, such that usage rights need to be checked, and it might be necessary to request permission from the platforms. However, a plethora of recent, large-scale reviews corpora are already provided from many platforms and product and services domains, for instance: Amazon customer reviews dataset with plenty of product categories [Ni et al., 2019]; general TripAdvisor hotel reviews [Alam et al., 2016], European restaurants [Leone, 2021], and hotels and restaurants from UK residents balanced by gender [Thelwall, 2018]; Yelp business reviews [Yelp, 2021]; IMDb movie reviews [Maas et al., 2011]; transcribed car reviews from YouTube [Stappen et al., 2021c], et cetera.

Ultimately, reviews for products and services are available in masses and are a rich and commonly used source for training machine learning (ML) models. Qualitative research can infer relevant, fine-grained consumption aspects of specific product and service categories, but the scope is limited to feedback on existing products only.

Open-Ended Answers From Evaluation Questionnaires

Asking customers directly for their satisfaction with a product or service is an established method in customer satisfaction (CSAT) research. This is traditionally achieved using questionnaires including Likert scale questions, which are answered by subjects with numerical values [Wirtz and Bateson, 1995]. However, features derived from open-ended

textual answers using NLP methods, such as, sentiment analysis and text mining, are found to *"play a key role in understanding how customers feel"* [Gallagher et al., 2019] about their experience. The reason is that *"there is a significant difference between the customer ratings score and the sentiment of their corresponding review of the product"* [Gallagher et al., 2019].

The verbally expressed experiences from individuals are a rich source of information shedding light on new aspect dimensions which are not covered by Likert scale questions, since they are not known to the interested parties. Consequentially, it is meaningful to include textual comments in evaluation questionnaires and to analyze them utilizing opinion mining methods.

"unsolicited comments written by customers in their very own words are deemed to be information-rich, full of dynamic evaluations of the service experienced and having a low extent of response bias."

Communication Format Open-ended comments are written by subjects in response to open-ended questions from evaluation questionnaires. These questionnaires are presented to the participants as part of a user study, survey, feedback form, et cetera. The subjects answer a series of questions on a form, which can be given on a piece of paper or digitally, e.g., on a website. The questionary is delivered to the subjects, for example, after they take part in a product, service, or system related experience, e.g., using a product, making use of a service, or testing a prototype. One might be reminded of well known everyday situations, for instance, receiving an email after a hotel visit containing a link to write a review and answer some questions. This scenario, however, is misleading, since the review might be an ordinary product or service review, which tends to be directed towards other customers and would be publicly available data, see Reviews About Product and Services. Instead, the audience or the addressee of a proper evaluation study are developers, producers, service providers, or other kinds of personnel concerned with actual product, service, or system development. The data is supposed to be read primarily by people which aim to create or improve the investigated item or at least understanding it better from the user perspective. This means that the feedback data is private, as it is not meant to be read by any other people than the denoted personnel. As such, it is supposed to be used internally by an institution to monitor CSAT and user experience for product, service, and system development and advancement, e.g., as part of the voice-of-the-customer methodology [Gaskin et al., 2010] or for usability testing to improve user-centered interaction design and human-computer interaction [Nielsen, 1994]. Thus, the questionnaires have a much higher depth of detail and ask for many more specific information about the user and customer experience compared to ordinary review data. Subjects participate in such user studies, because they might be incentivized, for example, by being paid for it, but also voluntarily when they are not paid, because they want to influence the future development of the item. Therefore, it is assumed that the feedback from evaluation questionnaires is more honest than the feedback given by public Reviews About Product and Services, since there are no external reasons for harmful motivations, such as, being paid for giving fake reviews or attracting attention for its own sake. Moreover, the open-ended answers are a way to canalize feelings, such as, frustration and anger but also gratitude and joy, and mirror them back to the developers after having been exposed to these emotions. These affective aspects make the information particularly rich and fruitful for sentiment analysis, especially because it is rather detailed, raw, and "private" feedback directed to the provider or developer.

Summarized, open-ended answers on evaluation questionnaires from subjects are direct, undistracted, and honest feedback for the developer or provider of the tested entity. Qualitative research profits from aspects which are not covered by Likert scale questions but expressed in textual comments.

Data Collection Data from evaluation questionnaires do not exist at the outset but must be gathered self-initiated, e.g., from a product developer or a service provider himself. To collect answers, questionnaires need to be created and subjects need to be prompted to answer those. Subjects, in turn, need to be exposed to an actual experience of a thing, such as, a test, a purchase, a trial, or any other form of procedure worthy of being investigated. Overall, a whole "experiment" needs to be designed for data collection. If such a user study is ought to measure a specific, previously known effect in a quantifiable and falsifiable manner, the environment of the subjects during the trial needs to be controlled to a high degree to exclude confounding factors. However, oftentimes it is not known a priori which problem exist with a given product or service, such that a controlled study is not even the correct method. Under those circumstances, the collection and analysis of especially the textual data might be particularly usable for qualitative analysis and mixed methods research to investigate grounded theories on the open-ended answers [Hagerer et al., 2021b]. This can be additionally consolidated with the statistical results from Likert scale answers.

In sum, questionnaire data is not publicly available a prior and, thus, needs to be gathered in an own evaluation process, which needs to be implemented manually. If this risk is settled, the resulting textual comments are a valuable source for qualitative content studies leveraging mixed methods text mining approaches. Grounded theories can be found to explain CSAT with respect to user or customer experiences.

3.1.2 Differentiation of Annotation Data

Opinion mining data contain user-generated texts containing expressed opinions. The information, which opinion is actually formulated about which opinion target, needs to be inferred for the analysis. The gold standard, therefore, is to annotate the texts manually with respect to sentiments and opinion targets. There are several ways how this annotation task is implemented in practice. Firstly, there are different things that can actually be annotated, e.g., coarse topic categories or fine-grained entity-attribute-sentiment pairs. This means, the annotations can be given at a high resolution with a high level of detail, i.e., fine-grained, or at a low resolution as rather coarse categories, i.e., coarse-grained. Then, it can be differentiated according to who is performing the labeling task. In that regard, we differentiate between annotations labeled either by amateur crowdworkers or by domain experts. This is relevant for the research question how the annotation resolution and the annotator personnel impact the annotation quality and model prediction performance. Both are important factors regarding how to conduct annotation data collection.

Granularity of Opinion Annotations

A classical and detailed way to mark opinions in texts is by labeling which sentiment is expressed and, optionally, towards which entity and attribute — see notably Kirange and Deshmukh [2014], Pontiki et al. [2015b, 2016]. The annotations are typically assigned on paragraph, sentence, or sub-sentence level. The annotations are in principle explained

in 2.1.1 Sentiment Analysis with an example shown in Figure 2.1. In the following, we differentiate opinion annotations according to their level of detail, if they are fine-grained or coarse-grained, and which are the technical impacts onto processing and dealing with that information.

Fine-Grained Annotations With regard to opinion triplet annotations, there is oftentimes an overall high number of possible triplets, since every possible sentiment can be combined with every possible entity, which in turn can co-occur with every possible attribute. For instance, with three sentiments, five entities, and five attributes, there are $3 \cdot 5 \cdot 5 = 75$ triplets possible, and this number can at times become considerably higher than that [Pontiki et al., 2016, laptop subtask]. Fine-grained annotations (opinion triplets) on texts require multiple of such triplets annotated for each sample, such that it is a multi-class multi-label annotation task due to several possible opinions expressed per text — see Figure 2.1. Beyond opinion triplets, there are other types of fine-grained annotations, such as belief statements from qualitative content studies [Danner and Menapace, 2020b]. These break down statements from consumers regarding if they represent a certain belief, e.g., if organic groceries taste better. This belief can be translated to the opinion triplet (organic groceries, taste, positive), showing that there is a close relation of belief statements to opinion triplets and both can be mapped to each other. Such studies show that fine-grained opinion annotations are a common format in qualitative content studies. Their annotations provide a high level of detail and are a rich, diverse, and high-quality data source for machine learning algorithms, which can be used particularly for Directed Content Analysis. In the thesis at hand we aim at leveraging that data by tackling the challenges of high annotation resolution and limited amount of data for machine learning and NLP in several works [Hagerer et al., 2021c, 2020b, 2021e].

Classification Metric for Fine-Grained Annotations A high annotation resolution means to classify a high number of possible classes. What is a meaningful specification of classification accuracy in that scenario? Accuracy itself is the ratio between true positives and n, and it is unsuitable due to its ignorance about precision and recall[1]. Both precision and recall *together* should be high, since all items of a class should be detected as such and no other items. This combined quality is reflected by the harmonic mean of both, i.e., the F1 score. However, F1 score — and precision and recall respectively — is derived for each class separately in the first place, and it is not immediately clear how to aggregate the number from all classes to give a meaningful summary as to overall classification quality. From the literature, we see a clear trend towards so-called micro-averaging when there are many classes to be recognized [Pontiki et al., 2016]. Micro-averaging means that the multi-label multi-class classification task with n classes is split into n binary classification tasks, of which for each the 2×2 confusion matrix C_i with $i = 1, ..., n$ is derived. C_i by definition contains the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. On that basis, the classification metrics precision P, recall R, F1 score F, and micro and macro F1 score are defined as follows:

[1] Precision: How many recognized samples are correctly recognized? Recall: How many samples of one class are recognized as such?

$$
\begin{aligned}
P(C_i) &= \frac{TP(C_i)}{TP(C_i) + FP(C_i)} \\
R(C_i) &= \frac{TP(C_i)}{TP(C_i) + FN(C_i)} \\
F(C_i) &= \frac{2 \cdot P(C_i) \cdot R(C_i)}{P(C_i) + R(C_i)} \\
F_{micro} &= F\left(\sum_{i=1}^{n} C_i\right)
\end{aligned}
\tag{3.1}
$$

For the micro F1 score, all confusion matrices are added up, and the resulting overall matrix serves as basis for that metric. This computation lays more weight on majority classes with a higher number of samples, whereas the minority classes with few samples are of less consequence. This appears reasonable, since with an n of around 100, there is typically a strong class imbalance with many classes on the edge of any significance, which are deemed problematic and should not distort the classification result by being weighted equally.

Coarse-Grained Annotations At the other extreme of the annotation level of detail are annotations which are coarse-grained, i.e., there is only a small number of classes $m <= 10$, which are rather high-level, abstract, and strongly summarizing or simplifying. In opinion mining, well-known coarse-grained annotations are plain sentiment analysis with two or three classes (positive, negative, and optionally neutral) [Rosenthal et al., 2017], or topic classification [Stappen et al., 2021c, Ganu et al., 2009, MuSe-Topic task] with a small number ($\leq$ 10) of topics as categories summarizing several fine-grained aspects or even a whole entity category. Alternatively, coarse sentiments and opinion targets can also be pre-determined by the data format, e.g., by pre-defined product review categories from shopping platforms [Ni et al., 2019] or pre-defined product ratings which are given by reviewers along with the review [Maas et al., 2011]. Due to being pre-defined, there is no manual a posteriori annotation labor required. In such scenarios, the data availability is comparatively high, which is beneficial for supervised classification performance. In addition, clusters generated by modern unsupervised methods show a high correspondence with coarse-grained topics, aspects, and sentiments, since unsupervised topics show a big overlap with the existing ones from the annotations [Stappen et al., 2021, He et al., 2017a, Hagerer et al., 2021c]. Last but not least, few coarse-grained classes tend to be clearer and more distinguishable than many fine-grained classes, as the latter can be ambiguous and overlapping [Sadegharmaki, 2020]. Apparently, coarse-grained opinion annotations have several advantages appealing for ML and NLP. For qualitative content analysis, coarse classes which summarize many detailed codes can successfully inform a fully automatized analysis for Directed Content Analysis [Hagerer et al., 2021c]. But also for Inductive Content Analysis, coarse themes and theme descriptions can be derived fully automatically and unsupervised [Danner et al., 2022, Hagerer et al., 2021d], since the overlap with manually found classes is significant [Hagerer et al., 2021c, Danner et al., 2020].

Classification Metric for Coarse-Grained Annotations When there are few, coarse-grained classes ($m \leq 10$) to be recognized, it is assumed that each class contains a

significant number of samples and is therefore represented meaningfully. This means that a classifier is expected to recognize all classes as good as possible. In the same way as for fine-grained classification, F1 score ensures that precision and recall are balanced. However, the classification performance metric should incorporate the classification performance of each class equally. In that regard, macro-averaging achieves a fair scoring by unweighted averaging of the class-wise scoring metrics [Rosenthal et al., 2017]:

$$F_{macro} = \frac{\sum_{i=1}^{n} F(C_i)}{n} \tag{3.2}$$

When using unsupervised methods, classification performance can not be calculated immediately, because it is not clear which cluster belongs to which class. In that case, the topic clusters are manually mapped to class labels according to their semantic meaning. While doing so, it becomes apparent which class labels have a cluster counterpart and which not. The ratio between mapable and all classes is the topic coverage [Stappen et al., 2021, He et al., 2017a]. After mapping, classification and evaluation can be conducted in the usual way. If necessary, coherence scoring can quantify the overall meaningfulness of the topics [Rosner et al., 2014].

Crowdsourced vs. Expert-Based

Annotation tasks can be differentiated according to who executes them: Few domain experts or many non-expert crowdworkers. In one of our publications, we define both types as follows:

> *" Experts are needed for complex annotation tasks requiring domain knowledge. Those tasks are not based on crowdsourcing, since the number of annotators is small and fixed. More common are external non-experts. Snow et al. [2008] showed that multi-labeled datasets annotated by non-expert improve performance. Khetan et al. [2017] showed that it also performs well in the singly labeled case. Thus, datasets made of singly labeled non-expert annotations can be cheaper, faster, and obtain performances comparable to those comprised of different types of annotations. Our organic dataset is annotated accordingly."*
>
> — Hagerer et al. [2021e]

It explains that the motivation to leverage crowdworkers stems from the idea of saving costs, ensuring representativeness, and improving performance for machine learning. The cost efficiency and representativeness of crowdsourcing annotations, however, comes at the price of a reduced overall consistency. Inter-rater reliability can be an issue for crowdsourcing, especially when there are many distant, anonymous, amateur crowdworkers. A lack of motivation and control leads to spamming annotators, which needs to be taken into account by modeling each annotator separately [Hagerer et al., 2021e]. The annotation process is distributed, which makes recurring synchronization to maintain a common annotation standard harder. Even if a crowdsourcing protocol is implemented correctly and carefully, crowd annotators are just not as knowledgeable as domain experts, i.e., they are not educated about the qualitative grounded theory methodology and are lacking expertise on the domain of interest.

On the other hand, expert-based annotations are highly detailed, sophisticated, and consistent. Thanks to their quality, they are inevitable for proper qualitative content studies, but their production is also time-consuming and costly, and experts might have their own biases as well. Concluding, crowdsourcing is a valid alternative for predictive modeling, whereas a proper protocol and consistency must be actively maintained. Also, the crowdsourcers' lack of domain expertise needs to be kept in mind.

3.1.3 Conclusions for Data Collection

Opinion mining data is generally user-generated, which introduces noise in terms of spelling errors, creative language use, et cetera. Techniques to filter and tackle that are a necessity. The data is also generally unstructured, so finding a theme-related classification, be it with manual labeling or unsupervised clustering, is generally one of the required opinion mining tasks. Public data from social media or product and service reviews is available in masses, but risks such as legal or privacy problems must be considered before collecting the data. For specific domains and use cases, e.g., product tests or service trials, open-ended answers from questionnaires is a valid option, when a sufficient amount of data can be gathered.

For annotating, crowdsourcing is advisable if the annotations are not too fine-grained and a proper protocol to ensure inter-rater reliability is implemented. Fine-grained annotations should be conducted by a domain expert for qualitative content analysis. The expert labels should be summarized to coarse-grained, synoptic classes in order to be used for predictive machine learning.

3.2 Datasets of This Thesis

Publication	Dataset	Domain	Annotations	NLP Method	Study Design
Hagerer et al. [2021c]	Social media	Organic food	Expert	Classification	Method evaluation
Hagerer et al. [2020b]	Social media	Organic food	Crowdsourced	Classification	Method evaluation
Hagerer et al. [2021e]	Social media	Organic food	Crowdsourced	Classification	Method evaluation
Hagerer et al. [2020a]	Amazon Reviews	Product reviews	Crowdsourced	Text Similarity	Method evaluation
Hagerer et al. [2021d]	Social media	Organic food	None	Clustering	Mixed methods
Danner et al. [2020]	Social media	Organic food	Expert	Clustering	Mixed methods
Danner et al. [2022]	Social media	Organic food	None	Clustering	Mixed methods
Hagerer et al. [2021a]	Social media	Organic food	None	Clustering	Mixed methods
Hagerer et al. [2021b]	Course Evaluations	Autograding	None	Clustering	Mixed methods
Stappen et al. [2021]	YouTube transcripts	Car reviews	None	Clustering	Method evaluation

Table 3.1 List of publications included in this thesis. The respective utilized data, methods, and study designs are shown. *Mixed methods* denote a paradigm, where a qualitative analysis is mixed with an innovative text mining analysis method on a large-scale dataset. *Method evaluations*, on the other hand, are trials for fine-grained sentiment classification, which could not be conveyed into practice for qualitative studies, e.g., directed content analysis (DCA). From our perspective, fine-grained expert and crowdsourcing annotations remain an open and difficult field of study in opinion mining.

3.2.1 Social Media Discourse About Organic Food

The main dataset of this thesis is collected from social media sources, where *organic food* matters are discussed, such as, organic food products, nutrition, sustainability, food safety, health impacts, et cetera. It is our main dataset, since most of our studies are carried out on this data. Various aspects of qualitative content mining and natural language processing (NLP) methods are investigated, for which several subsets are derived from the complete, underlying dataset. We start by describing this underlying dataset, which forms the basic quantity for everything what follows.

Then, we differentiate the subsets and the respective studies regarding if annotations are added. If not, then some kind of clustering or topic modeling method is examined for either a content study, a method evaluation, or a combination of both. Such studies are hereafter called *clustering studies* and seek to implement or enable Inductive Content Analysis. If there are annotations, then supervised classification is conducted to train on and predict the annotations. These studies are referred to as *classification studies* and aim at highlighting the limitations and chances of different annotation strategies and corresponding NLP methods for the purpose of Directed Content Analysis.

Motivation

The topic of organic food attracts increasing attention of consumers around the globe, who deem food safety and environmental issues as relevant for a healthy and sustainable lifestyle. Market research constantly needs to keep up with the needs and requirements of consumers, especially when the domain is complex. We propose social media discourse as particularly suitable, since it is an open-ended discussion format. It is used to search for consensus and, thus, is open for “brainstorming” about new concepts — see Dialog-Oriented Social Media. This development is summarized by our domain expert collaborators as follows:

> *"The organic food sector has grown considerably around the world over the last 20 years, spurred by the development of national organic standards (Sahota, 2018). As such, there is considerable research interest in understanding the respective consumer behavior (Hemmerling, Hamm, & Spiller, 2015; Hughner, McDonagh, Prothero, Shultz, & Stanton, 2007). [...] beliefs play an important role in explaining consumer behavior. [...] consumers hold beliefs about a product and evaluate those beliefs to form an attitude toward the product, which in turn influences purchase intention and behavior. "*
>
> — Danner and Menapace [2020b]

Consumer beliefs are Fine-Grained Annotations for aspect-based sentiment analysis (ABSA) and as such a common annotation format in Inductive Content Analysis. Therefore, opinion mining is relevant, and it has potential to automatize analysis procedures and to give explanations about consumer behavior. However, those Fine-Grained Annotations come at the risk of difficulties:

> *"Exploring these beliefs is complex due to the nature of organic food. [...] This leads to subjective and diverse beliefs about organic food regarding, for example, its safety, healthiness, and environmental friendliness (Fernqvist & Ekelund, 2014). This study analyzed online comments gathered from news websites and forums because they particularly lend themselves to the exploration of consumer beliefs."*
>
> — Danner and Menapace [2020b]

The risk of the organic food topic, which is difficult and diverse, can be addressed by using Dialog-Oriented Social Media, which are open-ended with a wide range of expressed ideas and concepts. Collecting them from "unsupervised" discussions is unobtrusive, maintaining the rawness of the concept space.

> *"The rise of different social media has enabled and sparked users' desires to share opinions publicly on online platforms (Ziegele, 2016), where consumers can write comments stating their beliefs about issues such as organic food. They do so of their own initiative (that is, without being asked by a researcher), revealing what matters to them. This shows which beliefs are on their mind and therefore salient and prevents the influence of the research process and physical/virtual presence of a researcher on stated beliefs (Branthwaite & Patterson, 2011). An additional plus to using online comments as data is that they are abundantly available at little cost."*
>
> — Danner and Menapace [2020b]

Concluding, the openness, availability, and unobtrusiveness of social media discussions are conducive to mine a complete inventory of opinions and attitudes towards complex and contemporary topics, such as, organic food. The goal of our studies is to provide technical methods supporting this process of Inductive Content Analysis as well as showing the limitations and potential of Directed Content Analysis.

Dataset Description

The basic quantity of our organic social media data consists of forum posts as well as newspaper/blog articles and readers' comments, which were written in the time span between 2007–2020. All texts are taken from German-speaking sources (Germany, Austria,

	English	German	Total
# Relevant Articles	5887	5846	11733
# Articles	14661	11992	26653
# Comments	140119	94442	234561
# Comments News Sites	101711	195429	297140
# Comments Forums	413636	374	414010
# Comments Blogs	0	5791	5791
# Sentences	441895	487794	929689
# Tokens	7198582	7752885	14951467
# Vocabulary	141579	262672	404251

Table 3.2 Statistics about the number of texts in the organic social media dataset. The ratio between English an German texts is 60/40. The German data has a focus on news site discussions, whereas the English data contains more forum discussions.

Switzerland) and from the United States. The ratio of German versus English data is roughly 40/60, see Table 3.2. The table also illustrates that for German mainly newspaper articles and comments were found, whereas for English there are more forum comments. The manifold sources are selected as follows:

> *"For German editorial articles, the online outlets of supra-regional print press, national print press (IVW, 2018) and the news sites (AGOF, 2018) with the highest reach were selected. Austrian, Swiss and US editorial sites were determined analogously. For blogs and forums, the snowball technique was applied. As experts in their field the domain experts already know several related blogs and forums. Additionally, other colleagues were consulted to identify further sustainability related blogs and forums in German and English."*
>
> — Widmer [2018]

The data sources were filtered using search engines and specific, organic-related search terms. This makes sure that most of the textual content has relevance for organic food:

> *"To retrieve the articles of a site either the site's internal search engine or Google Search was used. For German sources, the terms 'bio lebensmittel' and 'bio landwirtschaft' were used, and, for English sources, 'organic', 'organic food', 'organic agriculture', and 'organic farming' were used."*
>
> — Widmer [2018]

Articles — forum posts which spawn a discussion thread are also denoted as articles — found by those searches are downloaded together with the corresponding answering comments. After downloading, a relevance flag is assigned to all articles, if they are deemed relevant to the organic food topic:

> *"With the help of a domain expert 1000 random articles for both languages were initially labeled relevant or not and used as training data. Each document in the training and test set was composed by concatenating the article title, text, and the text of the first 100 comments. The output of the classification was shown to the domain expert and revised until satisfactory results were*

Language	Sources
English	New York Times, New York Post, Washington Post, Huffington Post, USA Today, Chicago Tribune, LA Times, US Message Board, Cafe Mom, Disqus, Quora, Reddit, Food Revolution, Food Babe, Organic Authority, Facebook, Organic Consumers
German	Welt, Münchner Merkur, Luzern Zeitung, der Freitag, SRF, Der Standard, Nachrichten.at, Heise, NDR, Zeit, Handelsblatt, RP, Focus, Tagesspiegel, WDR, Huffington Post (de), Salzburg.com, Krone, BR, Tagesanzeiger, Kurier, Spiegel, NZZ, Die Presse, WAZ, Tagesschau, TAZ, Aargauer, SWR, Utopia, Bio-Oeko-Forum.de, SciLogs, Lebe Heute, Greenpeace, Individualisten, Biologisch-Lecker, EAT SMARTER, Nachhaltigleben.ch, KarmaKonsum (Facebook), Campact

Table 3.3 Complete list of data sources of the organic social media dataset in its basic quantity. All studies on the organic dataset work on a respective subset of that data.

> *achieved. Using 10-fold cross validation the accuracy of the classifier was 84.70% for English articles and 78.00% for German articles."*
>
> — Widmer [2018]

The number of articles and relevant articles are depicted in Table 3.3. Roughly one half of all German articles and one third of all English articles are relevant according to the relevance classification.

Clustering Studies

In clustering studies, unsupervised methods are applied on this data. Clustering is an unsupervised technique and is used to automatically find a thematic structure in unstructured social media comments. Each study supposedly shows how domain experts benefit from that for qualitative content studies to find grounded theories using mixed methods approaches.

Topic Visualization For a feasibility study and prototype demonstration, *"we propose the SocialVisTUM toolkit, a new visualization and labeling tool to give users a comprehensible overview of the topics discussed in social media texts. [...] we provide a graph-based visualization showing the topics as labeled nodes and the correlation between them as edges. [...] contextual topic information is provided, such as the number of respective topic occurrences in the social media texts as node diameter, the correlation between the topic occurrences as edge width, example sentences from the data for each topic, a list of representative words for each topic, and the regarding sentiment distribution of a topic"* [Kirchhoff, 2019]. *"To avoid manual labeling [...], topic labels (and optimal hyperparameters, author's note) are generated automatically based on a custom algorithm. [...] we show the results of a case study based on social media texts from online commenters debating about organic food consumption. (According to our domain expert, author's note) the correlated topics give a meaningful graph representation of the social media discussions supporting the understanding of the concerns of consumers"* [Hagerer et al., 2021a]. The visualization study is carried out on the social media comments of the English part of the dataset. Only user-generated comments with the relevance flag are considered. With regard to the research goal, *"the graph-based visualization with topics as nodes and topic correlations as edges reflects the topics and patterns found in a related qualitative content*

analysis [Danner and Menapace, 2020a]. The presentation of additional topic information, such as word lists, representative sentences, topic importance, and meaningful predefined labels, provide a basis for the understanding and interpretation of a topic for domain experts" [Hagerer et al., 2021a]. Concluding, clustering of sentences and aggregating topics and sentiments from sentences yields meaningful results for Inductive Content Analysis. We draw on this evidence for our consecutive studies by choosing sentences as the basis for all expressed opinions.

Media Agenda Setting The following studies are focussed on how newspaper articles influence the readers' comments. As we show in our content analysis, the domain of organic food is especially relevant here, since *"sustainability [Holt and Barkemeyer, 2012] and organic food issues [Lockie, 2006, Meyers and Abrams, 2010, Danner and Thøgersen, 2021] increasingly make it on the agenda of the news media. [. . .] Priming organic food topics [. . .] can nudge consumers toward buying organic food products and voting in favor of policies supporting organic agriculture. Thus, to understand the drivers of organic food consumption, it is crucial to investigate the media's influence on public opinion regarding organic food"* [Danner et al., 2022]. With respect to data collection, *"Comment boards of online newspapers [. . .] provide new opportunities for quantitative analyses and insights into public opinion [Neuman et al., 2014]. [. . .] Such data is exploited by marketing and consumer research to gain insights into consumer thinking [Balducci and Marinova, 2018]. [. . .] The present text-mining research investigates the relationship between [. . .] news articles and comment sections of nytimes.com and spiegel.de as two major online US and German news outlets. [. . .] The dynamics of the media and public agendas between January 2007 and February 2020 are analyzed and compared. [. . .] To detect the discussion topics, the text data from news articles and reader comments are analyzed with a topic modeling approach based on clustered multilingual sentence embeddings. [. . .] [It] enables the analysis and comparison of agendas in different countries and languages, here English and German. Such a comparison across languages is unprecedented in agenda-setting literature to the best of the authors' knowledge"* [Danner et al., 2022].
We draw the data from our basic quantity and take all articles and comments from Der Spiegel (German) and New York Times (English) without considering the relevance flag. Topic modeling is accomplished by chunking all texts to sentences, calculating all sentence embeddings, performing k-means clustering, labeling each cluster, and counting the number of sentences per cluster in each document. Since the sentence embeddings come from a cross-lingual pre-trained deep neural network [Chidambaram et al., 2018], the topic model is coherent across several languages and models them language-agnostically. Thus, we are able to provide fixed-dimensional, cross-lingual, coherent, and explainable vector representations of documents, i.e., articles and comments — see our publications Hagerer et al. [2021d, 2020a] to comprehend how we evaluate the method prior to this media agenda study of Danner et al. [2022]. When each article and comment is represented with such a vector, it can be used compare articles with comments. In that way, the mutual influence is measured by showing if the same topics are discussed. The results of the study demonstrate how multi-lingual Inductive Content Analysis is benefitting from our novel NLP technology. It is especially noteworthy that our method is advantageous compared to *"the classic methodology of agenda-setting research [which] consists of comparing content analyses of (print) media with public opinion surveys [Luo et al., 2019b]. Disadvantages of this approach are costly surveys, the required matching different units of measurement and scales, and possibly biasing time gaps between article publication*

Language	English	German
Main themes	4	4
Themes	21	21
Beliefs	62	60
Documents	1099	789
Sentences	2275	2334

Table 3.4 Statistics of the expert annotations from the organic dataset. Taken from Hagerer et al. [2021c].

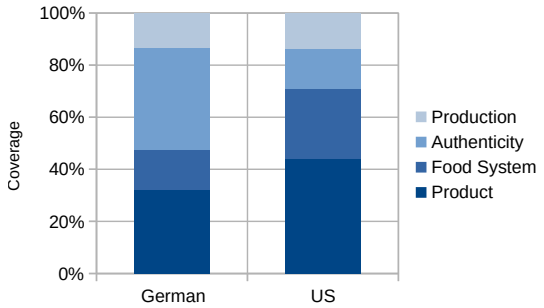


Table 3.5 Distribution of main themes of the expert annotations of the organic social media dataset [Danner and Menapace, 2020b].

Main Theme	Themes
Product	Food safety, Price, Healthiness, Taste, Nutritional value, GMO, Quality, Naturalness, Availability
Food System	System integrity, Food security, Production scale, Farmer welfare
Authenticity	Organic labels, Origin, Retail brand, Product category, Packaging
Production	Environment, Animal welfare, Biodiversity, Working conditions

Table 3.6 Main themes and themes of the expert annotations. Each main theme contains several themes. The themes in turn contain many belief statements, see Table 2.2 for examples.

and survey [Thøgersen, 2006]" [Danner et al., 2022]. With our method, *"agenda-setting research can now directly can now directly compare topics and sentiment in online articles and comments using the same (automatic) text analysis methods"* [Danner et al., 2022].

Classification Studies

In classification studies, we take a subset of our organic social media dataset and annotate the user comments with regard to ABSA. We opt for two annotation strategies: Expert annotations from qualitative content studies and non-expert annotations from crowdsourcing. The annotations are used to train machine learning classifiers, which in turn are supposed to help with Directed Content Analysis. We confine the studies to see how suitable both annotation types are for ABSA classification. Advantages and disadvantages of each are discussed, and problem-solving strategies for the respective pitfalls are laid out. Pitfalls for expert annotations are too little data for too many, detailed annotations; for crowdsourced annotations typical problems are annotation noise.

Expert Annotations The expert annotations in this work are taken from the content study of our collaborators Danner and Menapace [2020b]. ABSA annotations are given as so-called belief statements, of which examples are shown earlier in this thesis in Table 2.2 and explanations in Inductive Content Analysis and Fine-Grained Annotations.
The annotations were assigned to German (DACH) and English (US) comments on newspaper articles and discussion forums, which are enlisted in Table 3.3. The data is drawn for our basic quantity described previously in Dataset Description. The selection of data samples is chosen as follows: *"We randomly sampled threads for the DACH sample (60 threads totaling 1094 comments). The random sampling resulted in approximately 70% of comments stemming from news websites and 30% from forums. This reflected the fact*

Category	%	Label	%	Sub-Label
sentiment			39%	neutral
			32%	positive
			29%	negative
entity	83%	organic	36%	products
			23%	general
			18%	farming
			6%	companies
	11%	conventional	6%	products
			1%	general
			4%	farming
			0%	companies
	5%	GMO		
attribute	33%	general		
	28%	healthiness	14%	pesticides
			10%	healthiness
			4%	food safety
	6%	price		
	12%	trust	7%	certification
			2%	origin
			2%	retailers
			1%	origin
	11%	quality	7%	nutrition
			4%	taste
	10%	environment	5%	environment
			3%	productivity
			2%	animal welfare

Table 3.7 Annotation distribution of all annotated sentences, i.e., 53% or 5561 of all 10441 sentences, of the organic dataset. 668 of the annotated sentences contain two or more opinion triplets.

that in both the DACH and the US population of threads, there were more comments on news websites than on forums. Third, we randomly sampled threads for the US sample (47 threads totaling 1069 comments), subject to the restriction of yielding a similar number of comments and maintaining a constant ratio of news website and forum comments. On average, one thread consisted of 18 comments in the DACH sample and 25 comments in the US sample. The mean length of the comments was 62 words for DACH and 99 words for the US" [Danner and Menapace, 2020b]. The approximately 1000 user comments are annotated with 60 belief statements, which express a positive, neutral, or negative attitude towards an entity and attribute, for instance, the taste of an organic product or the harmfulness of an organic farming practice. The 60 beliefs are grouped to 21 themes, which in turn are grouped to 4 main themes — see Table 3.6.
We aim at predicting the expert annotations with text classification methods based on NLP and machine learning (ML). In a first study, we find that features generated by a pre-trained deep neural network (Universal Sentence Encoder (USE)) are significantly correlated with the themes from the expert annotations [Danner et al., 2020]. However, in a second study [Hagerer et al., 2021c], we evaluate various ML classifiers and find that they perform not good enough for theme prediction in order to be used productively for Directed Content Analysis. Re-grouping the belief statements using hierarchical clustering does not improve the situation. It appears that the amount of data does not suffice to train a well-performing classifier. However, we see that clusterings of deep pre-trained sentence embeddings from Universal Sentence Encoder Cross-Lingual (XLING) are predicted best by all classifiers and yield the most optimal class balance. This supports the conclusion that these kinds of clustering methods have beneficial properties for automatized content analysis.

Crowdsourced Annotations The classification performance of expert annotations appears to be weak due to the small number of annotated samples. However, collecting many annotations is feasible with crowdsourcing, especially since training with many non-expert annotations can outperform training with fewer expert annotations Snow et al. [2008]. Therefore, we chose 1,373 random Quora comments containing 10,441 sentences from our parent dataset to be annotated by 10 non-expert students with some initial instructions given by our domain expert as follows: Each sentence gets sentiment (positive, negative, neutral), entity (organic or non-organic products, farming practices, and companies), and attribute (healthiness, price, trust, quality, environment) annotated — see Table 3.7 for label statistics. *"The data is annotated by each of the 10 coders separately; it is divided into 10 batches of* 1,000 *sentences for each annotator and none of these batches shared any sentences between each other. [...] After annotation, the data splits are 80% training, 10% validation, and 10% test set. The data distribution over sentiments, entities, and attributes remains similar on all splits"* [Hagerer et al., 2021e].
In the first classification study, we implemented a series of experiments to predict ABSA with a state-of-the-art (SOTA) technique called progressive neural networks [Hagerer et al., 2020b]. It can be summarized as an involved combination of transfer learning and deep, recurrent neural networks Rusu et al. [2016]. Our experiments show high accuracies on external challenge corpora and a consistent advantage of the method compared to the baselines. However, predictive F1 scores are not better than 17% on the non-expert ABSA labels. It is apparent that classifying these annotations is problematic and not ready to be used for content studies. The annotation problems were observed to originate from two main causes: too fine-grained annotations and too low inter-rater agreements between different groups of annotators.
Thus, we base our second classification study on a reduced annotation granularity by only considering three-fold sentiment classification on organic-related sentences. The analysis targets to model annotation noise in the form of annotator bias [Hagerer et al., 2021e]. It provides a way to model annotator-specific bias and an overall ground truth in conditions, where each sample is annotated only one single time by exclusively one annotator. The end-to-end approach is proven mathematically, and empirical improvements are established. Still, accuracies do not become higher than 50%, which we do not consider worth of being used for DCA.

Summary of Insights

Overall, we conducted two types of studies with two types of insights. The Clustering Studies are overall successful from an application and from a method perspective. With regard to their domain application, they are able to create concrete insights for implementations of Inductive Content Analysis. At the same time, novel technologies are used successfully for new use cases, e.g., multi-lingual opinion mining or visualizations of correlated neural topic models.

The Classification Studies are successful from a method perspective, since method improvements for transfer learning and annotation noise removal are achieved. However, the classification results are eventually not good enough to be used in practice for Directed Content Analysis, even not for Coarse-Grained Annotations.

As a sidenote to depict the challenges with our classification problems, we give a succinct overview of our related student research projects as follows. We used many forms of — back then — novel deep learning methods for ABSA. Schober [2019] provided an own

from scratch implementation for multi-task modeling. Shouman [2019] utilized the pre-trained ULMFiT deep neural network from Howard and Ruder [2018] for ABSA. Progressive neural networks by Rusu et al. [2016] have been pre-trained and fine-tuned on several tasks by Gupta [2019]. Dugar [2019] leveraged hierarchical long short-term recurrent neural networks. In Datta [2019], multi-instance learning in conjunction with transfomer neural networks were evaluated. With regard to ML based similarity metrics, approaches such as Siamese neural networks Mai [2019], Mendonca [2019] were used in order to increase the number of training samples. For data augmentation, Mosharafa [2019] made use of forth-and-back translation, and Arumugaswamy [2019] employed pre-trained SOTA language models. For causal investigation, semantic features of pre-trained neural networks were used by Mushtaq [2020] to analyze their relation with inter-rater agreement, and by Sadegharmaki [2020] to depict the semantic coherence of annotated classes. Another solution approach to the coherence problems was given by Jiang [2020] by considering only the most informative samples for training. Last but not least, Le [2021] aimed at using hierarchical clustering and various forms of transfer learning to derive more coherent class structures which would be easier to predict.

3.2.2 Comments From Teaching Course Evaluations

The following paragraphs give a short summary of our course evaluation dataset. As a bonus, we provide several data statistics, which are not part of the previous publication about it.

Motivation

The initial idea for this dataset was to investigate *how the introduction of an autograding system impacted student satisfaction*. At Technical University of Munich the number of computer science students is growing rapidly at an annual rate of 20% and reached now 2500. About the students, the paper states that

> *"programming is a crucial skill for their academic and professional careers in engineering and natural and social sciences. Therefore, many instructors apply autograding to programming exercises to provide immediate feedback to students and lower the manual grading effort. However, little is known regarding how autograding relates to student satisfaction with the course and its varied teaching aspects. Therefore, course evaluations are a standard method for course organizers and lecturers to understand which parts of the course the students liked and which did not. Especially open-ended comments can be insightful since they contain opinions about emerging course aspects that are not being asked in Likert scale questions. Here, a two-pronged analysis approach using text mining in addition to basic statistics ensures no information is lost."*
>
> — Hagerer et al. [2021b]

Here, we aim at using evaluation comments and opinion mining as a means to investigate possible reasons why an autograder could be beneficial for student satisfaction and their perception of the course quality. Sentiment-related topics can be related to numerical answers for quantitative evidence. The methodology is based on mixed methods, i.e., a combination of text mining based content analysis and additional statistics, to find theories for the impact of autograding on student satisfaction.

Dataset Description

The dataset consists of questionary answers from teaching course evaluation surveys of computer science lectures, which all include programming homework exercises utilizing autograding. This is explained as follows:

> *"At the Technical University of Munich, students can evaluate the courses they have taken through anonymous feedback. Questionnaires are pre-defined and distributed every semester for every course by the student council. They contain open-ended text fields where students describe in own words aspects of the course they appreciated and which could be improved. In other questions students rate different aspects of the course on a Likert scale, e.g., lecture materials or homework exercises. In the scope of this study, we collected evaluation results from [...] 3 distinct modules of the study plan of informatics, [...] some courses were repeating instances of the same module held in*

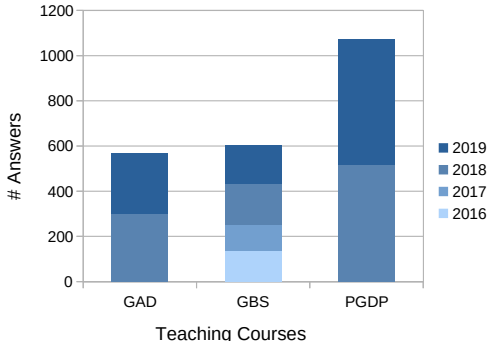


Figure 3.1 Number of submitted answers to the teaching course evaluation questionnaires per course and year. All three courses involve programming aspects.

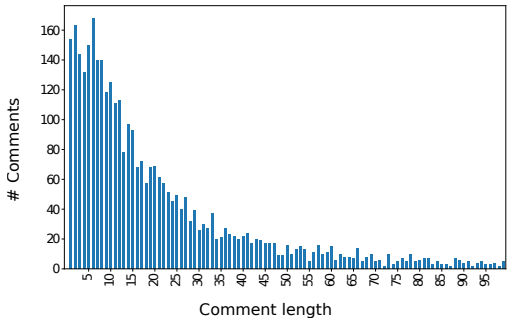


Figure 3.2 Comment length (in number of characters) histogram of the whole course evaluation dataset. Taken from Brauweiler and Neumann [2021].

> *different years. [. . .] Programming tasks are a significant part of the homework in each course."*
>
> — Hagerer et al. [2021b]

In a summary, there are repeating course evaluations of three courses over a timespan of at least two years, i.e., the year before and the year after the autograding system was introduced. In the paper Hagerer et al. [2021b], the difference is shown based on the satisfaction development over the years.

Statistics

Per course year, the number of answers ranges from 100 – 600, and they all sum up to 2245, which is shown in Figure 3.1. Most of the comments are not longer than 50 characters with only few exceptions, see Figure 3.2.
As a bonus unrelated to autograding, we conducted entity and topic related sentiment analysis on all evaluation comments. Regarding entities, only names of online platforms for chatting, q&a, autograding, et cetera are taken. Regarding topics, these are generated with topic modeling based on non-negative matrix factorization (NMF). Sentiments are predicted as follows:

> *"In order to train a sentiment classifier, [. . .] we manually labelled over 600 text comments [. . .]. We chose three categories as our labels: positive, negative and neutral. The first two labels were reserved for comments that clearly expressed positive or negative sentiment. [. . .] we chose a naive Bayes classifier with only two target classes: positive and negative. The reasoning behind this is that we prefer a continuous sentiment scale. [. . .] We trained a multinomial naive Bayes classifier with 0.1 Laplace smoothing and achieved an accuracy of 88% and 2% standard deviation using 10-fold cross validation."*
>
> — Brauweiler and Neumann [2021]

The sentiment distribution of all comments containing one respective entity or topic are displayed in Figure 3.3 and Figure 3.4. It can be seen that Artemis, Slack, and Zulip are popular online tools and Piazza, Matrix, Moodle and RocketChat are rather unpopular. This matches the reported feedback and the overall impression from the organization staff

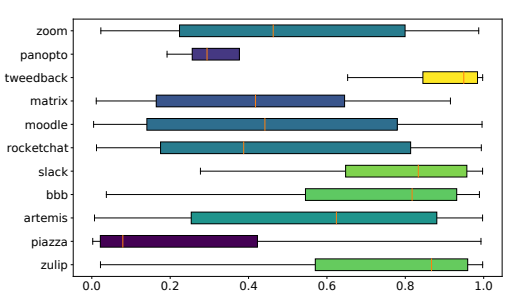


Figure 3.3 Sentiment distribution in evaluation answers grouped by online tool. Taken from Brauweiler and Neumann [2021].

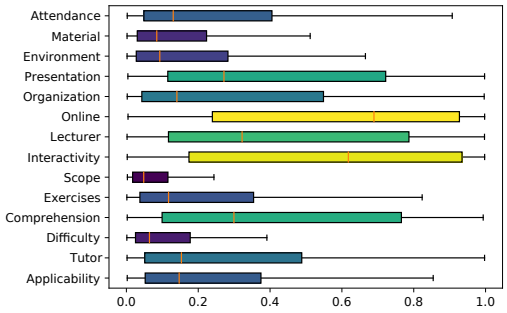


Figure 3.4 Sentiment distribution in evaluation answers grouped by topic. Taken from Brauweiler and Neumann [2021].

and, thus, is deemed as being correct. With regard to topics, online and interactivity aspects are referred to with positive sentiments, since both are seen as an engaging element in learning. Several of the negative topic-related sentiments, e.g., from tutor or applicability, are not immediately clear, since these course aspects are positively evaluated according to the numerical answers, see Hagerer et al. [2021b, sections V-B & V-G]. This shows that sentiment analysis can in some circumstances be orthogonal to pre-defined Likert scale questions. Tutors, for example, could be mentioned in a negative context, as they are the ones alleviating specific problems in learning. Future work should incorporate opinion mining features for creating a more complete picture and establishing well-justified grounded theories in that regard.

Research Questions

The core aspect of this dataset is about the introduction of an autograder in several large-scale programming courses at a university. Using evaluation questionnaires, we aim at answering the following qualitative research questions about the intervention in our paper Hagerer et al. [2021b]:

- How did students report on their learning experience in course evaluations, and how did it change?
- How did the reported interaction between students and tutors change?
- How did the perceived difficulty of the practical programming parts of the courses change?
- How did the perceived overall course quality change?

In a summary, we see a consistent positive impact of our Artemis autograding system on all observed aspects of student satisfaction. We account for two main causes: First, automatic testing improve tutoring sessions and homework assignments, since students can progress independently with meaningful feedback from the autograding system. Second, homework exercises become fairer due to a randomized double-blind correction system.

With regard to the used NLP method, we are also interested in the following technical aspects:

- How do evaluation comments qualitatively relate to Likert scale answers?
- If some Likert questions are biased or missing for some courses, is it compensated by topic models?
- What is the potential of transfer learning for topic modeling on limited datasets?

Qualitatively, some of the results show a close relation between topic distributions and numerical answers. Especially critical comments are in almost all scenarios analogous to Likert scale answers. Hence, the method has the potential to compensate for missing numerical questions. Transfer learning was incorporated using the KG-NMF topic modeling method [Chen et al., 2019], which was successful on our rather small and limited dataset.

Publication and Student Contributions

The following student works were contributions to the research success in Hagerer et al. [2021b] and deserve an honorable mention:

Originally, Professor Stephan Krusche hit on the idea that the introduction of Artemis in several programming courses could be worth investigating. After that, Lahesoo and Anschütz [2020] executed the first student project on the — back then quite limited — dataset. The students showed significant changes in the evaluation after the autograder intervention, demonstrating that the analysis and methodology bears potential. Furthermore, the KG-NMF topic modeling method [Chen et al., 2019] turned out to be useful.

In a follow-up project, Lahesoo [2021] extended the analysis on an increased amount of data with involved hyperparameter optimizations and visualization tuning, leading to many eventual results.

Last but not least, Brauweiler and Neumann [2021] contributed actual sentiment analysis, which was based on NLP instead of pre-defined categories. They showed that sentiments add meaningful information to pre-defined positive/negative categories. This completed the future work as it is stated by Hagerer et al. [2021b].

3.2.3 Transcripts From Video Reviews

As last case study, we examine textual transcriptions from YouTube video reviews. Therefore, the recent *Mu*ltimodal *Se*ntiment analysis in *C*ar *R*eviews (MuSe-CaR) dataset is utilized [Stappen et al., 2021c]. Investigating such large-scale data sources with mixed methods approaches is relevant for modern qualitative research aiming at opinions collected from in-the-wild scenarios. The goal of our study is to develop a simple and robust topic modeling method which can be used for Inductive Content Analysis on this kind of data.

Motivation

The reason why video data is considered to be an important data source is that the amount of available user-generated video data is rapidly growing:

> *"Global video traffic is estimated to grow four-fold in the coming years [1], accounting for 80% of all online traffic in 2019 [2]. On social media, users view eight billion videos daily on Facebook [3] and YouTube has become the second biggest social network with nearly two billion active users and one billion hours watched each day [4]. The internet has undergone a rapid transformation from a largely text-based Web 2.0 to a multimedia, user content-driven net."*
>
> — Stappen et al. [2021c]

As video data is becoming a new default for user-generated content, such data gains relevance for content studies from various fields of domain-specific, content-related research disciplines:

> *"For example, educational information on cancer treatment [Basch et al., 2017] and hearing aids [Manchaiah et al., 2020] are studied in health-care, the influence on election campaigns in social sciences [Gueorguieva, 2008], and large-scale multimodal sentiment in multimodal machine learning [Wöllmer et al., 2013, Morency et al., 2011, Stappen et al., 2021c,a,b]. For these approaches, researchers closely examine the videos for collection, labelling, and analysis, whereby visual patterns and metadata, e.g. authorship, can be exploited. Nowadays, also transcripts — automatically created by YouTube — are available [Harrenstien, 2009]. Since text is the most meaningful modality to understand contextual information, effective computer-assisted text analysis methods are needed."*
>
> — Stappen et al. [2021]

Video transcripts are provided by default from YouTube and can be used straight away for textual data analysis. However, there is a lack of technical NLP methods to find semantic structure in that data, which thus need to be developed:

> *"Video transcripts are an emerging data domain, however, the explicit use for topic modeling is understudied [Morchid and Linarès, 2013, Basu et al., 2016, Das et al., 2019]. To broaden the perspective on this medium more evaluation and new approaches are needed."*
>
> — Stappen et al. [2021]

%	Annotated Topic	Aspects
16	General	series, weight, sales, warranty, models, brands, competitors
13	Driving Experience	braking, steering, gear shifting, centroid, chassis, suspension
13	Performance	electric, hybrid, combustion, horsepower, RPM, acceleration
7	Feature exterior	headlight, foglight, taillight, locks, handle
7	User Experience	screen, bluetooth, realtime traffic, interface, iDrive system, gestures
7	Quality & Aesthetic	interior, exterior, style (sporty, etc.), material quality, clearance
6	Comfort	leather, touch, leg room, head room, luggage
6	Feature interior	radio, speaker, belt, split folding backs
3	Costs	retail price, base price, feature price, insurance, maintenance, resale
2	Safety	anti-lock brakes, traction control

Table 3.8 Annotated topics statistics from the MuSe-CaR dataset [Stappen et al., 2021c]. The dataset consists of text transcripts of video reviews about cars. The topics in the table are manually annotated topics, and they summarize several aspects listed in the right column.

This forms the motivation for our research to provide an according topic modeling method able to deal with video transcripts data, such that it can be analyzed for mixed methods driven Inductive Content Analysis.

Dataset Description

The MuSe-CaR dataset provides YouTube videos containing car reviews together with the transcripts from the spoken texts. Furthermore, manual annotations are provided for emotion recognition, sentiment analysis, and discussed topics. It is due to the multimodal features and manifold annotations that it is a multimodal sentiment analysis dataset. It also has a strong focus on everyday conditions, which is explained as follows:

> *"[It] has been gathered under real world conditions with the intention of developing appropriate methods and further understanding of multimodal sentiment analysis 'in-the-wild'. [...] the dominant focus for MuSe-CaR is to aid in machine understanding of how positive and negative sentiment as well as emotional arousal is linked to an entity and aspects in a review (and other user-generated content in general). In doing so, MuSe-CaR aims to bridge fields within affective computing, which currently utilise a variety of emotionally annotated signals (dimensional and categorical)."*
>
> — Stappen et al. [2021c]

As to our study, we are primarily interested in topic modeling on video transcripts, which we find to be understudied. In that regard, Table 3.8 show all annotated topics and which aspects or entities these contain. The topics cover themes related to car consumption, such as, car driving experience (braking, steering, gear shifting), interior features (radio, speaker, belt), safety (anti-lock brakes, traction control), and costs (retail price, feature price, maintenance).

Statistics

The dataset contains *"over 40 hours of user-generated video material with more than 350 reviews and 70 host speakers (as well as 20 overdubbed narrators) from YouTube"* [Stappen et al., 2021c]. With that amount of data *"MuSe-CaR is one of the largest state-of-the-*

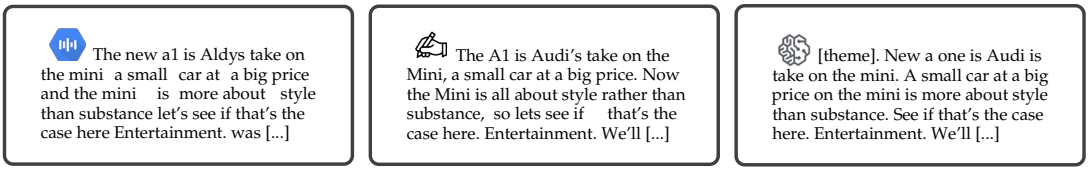


Figure 3.5 The video transcript with the worst speech-to-text result according to word error rate from the MuSe-CaR dataset. Left: Google-Transcribe with 39.44%; middle: manually transcribed; right: AWS with 37.85%. A core challenge is noise robust topic modeling with high coherence despite this type of mistakes — see our GraphTMT approach from Stappen et al. [2021]. Taken from Stappen et al. [2021c].

art video datasets for multimodal sentiment analysis research [. . .]. The reported word error rate of the automatic transcript is estimated around 28%" [Stappen et al., 2021c]. Figure 3.5 shows a particular problematic video transcript to demonstrate the challenge for coherent topic modeling. The transcripts have been chunked via sliding window in paragraphs, and they were manually annotated with the topics in Table 3.8. Driving experience and performance aspects are discussed most, whereas costs and safety issues occur seldom. Sentiment, affect, and trustworthiness is also labeled and can be related with opinion targets.

Research Contributions

For our own study, we are interested in methods to support domain experts in their qualitative content studies. Therefore, user-generated video reviews are an important resource. Automatic transcripts plus topic modeling have the potential to scale the analysis to large-scale data in order to become representative. However, methods yet need to be found which are able to deal with word errors from automatic transcripts:

> *"We propose a novel graph-based approach for topic modeling for the emerging use case of video transcripts. It is the first time, an unsupervised extraction model is applied to the large-scale, noisy MuSe-CaR dataset packed with typical mistakes of automatic speech-to-text. The performance is extensively benchmarked on this dataset against conventional methods. Here, the semantic consistency of the topics is evaluated by assessing a common coherence measure. Furthermore, for a more human-centred evaluation approach of the results and to determine the semantic validity, we conduct a structured word intrusion user study with 31 subjects. Finally, we evaluate the coherence of our approach on a standard topic modeling dataset of product reviews to assess the potential for other use cases. Our results show that GraphTMT outperforms conventional methods on the MuSe-CaR datasets."*
>
> — Stappen et al. [2021]

We conclude that our GraphTMT method is a means to enable Inductive Content Analysis on large-scale video transcripts. This is shown by coherence measurements, user studies, and the fact hyperparameter optimization is unnecessary.

Publication and Student Contributions

The GraphTMT topic modeling approach was evaluated on this MuSe-CaR dataset. This research project is based on the master thesis of Thies [2021], where the approach was firstly formulated and the user study and coherence scorings were carried out. Lukas Stappen, the main supervisor, gathered the MuSe-CaR dataset, which was published as a separate challenge prior to our GraphTMT study under the title “First International Multimodal Sentiment Analysis in Real-life Media Challenge” (MuSe 2020) at the ACM Multimedia 2020 conference [Stappen et al., 2020].

4 Conclusion

4.1 Findings

The main question of this thesis was initially formulated how a domain expert can be supported in her qualitative content analysis with opinion mining methods. In the following paragraphs, we answer several aspects of that question which are formulated in Research Questions.

When discussing perspective widening, one question is if opinion mining methods are able to automatically discover similar themes and insights which would also be found manually through qualitative content analysis by a domain expert. Our studies and related work show that state-of-the-art (SOTA) natural language processing (NLP) methods unveil similar opinion aspects and distributions on large amounts of data [Hagerer et al., 2021a,d] when compared to a qualitative study [Danner et al., 2020]. So, automatically mined themes tend to correspond with the themes found by domain experts and go even beyond that by displaying themes and aspects from very large amounts of data. This principle scales up even to multimodal multimedia data, leading to new opportunities to mine opinions from data domains in even more natural and unbiased in-the-wild scenarios [Stappen et al., 2021]. Since multimedia data is more difficult to be analyzed manually than textual data, this opens another door to the idea of perspective widening.

From the quantitative perspective, large-scale social media text mining methods clearly show potential to enhance and augment representative polling surveys, if not even to replace them. Evaluation questionnaire information from numerical and textual answers are similar and correlated [Hagerer et al., 2021b], and the topics of newspaper articles influence their commenters across multiple languages, making a complicated cross-cultural survey redundant [Danner et al., 2022]. We consider these results as strong evidence for additional value for quantitative depth.

The question if a domain expert benefits more from perspective widening than from quantitative depth cannot be answered unambiguously. Unsupervised clustering methods always provide both insights, i.e., themes with their semantic meanings and statistical distributions. All proposed methods are able to provide meaningful results by one means or the other. Which of both is more utilizable depends on the actual problem definition, since basically both tend to give feasible outcomes. We showed that statistical topic distributions can complement or substitute polling in some cases [Hagerer et al., 2021b, Danner et al., 2022]. On the other hand, modern clustering techniques reflect the big datasets completely with cross-lingually coherent themes, which are to a high degree pre-determined and hierarchically structured [Hagerer et al., 2021d]. They can be leveraged for domain exploration and to derive possible explanations and grounded theories [Hagerer et al., 2021a].

With regard to recent technological trends, there is a significant additional value of SOTA NLP for text miners. Deep pre-trained artificial neural networks for natural language understanding improve the SOTA of semantic similarity [Hagerer et al., 2020a]. These features show significant correlations to expert annotations [Danner et al., 2020]. Due to the inte-

gration of multiple languages into one aligned model, especially multi-lingual content studies are now easily possible [Hagerer et al., 2021d, Danner et al., 2022] without fine-tuning. However, when fine-tuning is applied, aspect-based sentiment analysis (ABSA) can profit from transfer learning by overcoming the forgetting effect, which improves classification on crowdsourced fine-grained sentiment annotations [Hagerer et al., 2020b]. Classification performance can further be improved by removing annotator biases [Hagerer et al., 2021e], which we showed is even feasible in a singly labeled crowdsourcing setting thanks to novel end-to-end learning methods.

Concluding, modern opinion mining technology offers a lot of potential for domain experts. Themes from expert-driven qualitative content studies are found autonomously in social media texts by deep learning models [Danner et al., 2020, Hagerer et al., 2021a]. We also show two instances, where questionnaire-based polls are complemented or replaced by opinion mining [Danner et al., 2022, Hagerer et al., 2021b]. The SOTA methods introduce strong innovations, such as, multi-lingualism [Hagerer et al., 2021d], transfer learning [Hagerer et al., 2020b], and end-to-end learning [Hagerer et al., 2021e], from which expert-driven opinion mining can benefit in future work for classification.

4.2 Limitations

The previously mentioned findings show how expert-guided opinion mining is benefitting from modern natural language processing (NLP). However, there are several limitations with regard to what extent this is the case.

First and foremost, *fine-grained information about opinions in texts* is very difficult to analyze using automatic methods, no matter if they are supervised or unsupervised and also independent of the recent state-of-the-art (SOTA). To the author's best knowledge, there is simply no way for technical methods to match the high resolution and level of detail with which a domain expert can label and structure smallest fractions and aspects on subsentence level of a textual corpus, also not when transfer learning is used [Sadegharmaki, 2020, Kasischke, 2019].

Unsupervised methods tend to become incoherent with increasing number of clusters, which has been shown in the related work [Bakharia, 2019] and corresponds to our own observations [Hagerer et al., 2021c]. Thus, they are incapable to create a highly fine-grained and consistent class labeling hierarchy. For supervised methods, we basically show in our work that predicting highly fine-grained aspect-based sentiment analysis (ABSA) is difficult as well [Hagerer et al., 2020b]. The core problem of predicting *expert annotations* is that there are too few data samples and a comparatively very high level of detail in the annotations. The mismatch is too strong, leading to bad prediction performance, which is too low to be used productively by domain experts [Hagerer et al., 2021c]. Using many crowdsourced non-expert annotations is supposed by the literature [Snow et al., 2008], but it eventually fails because of the complexity and level of detail of the annotation task [Hagerer et al., 2021c, 2020b]. We conclude that robust opinion mining classification for qualitative content studies is only feasible with coarse-grained class labels — see also [Snow et al., 2008].

4.3 Future Work

With regard to the annotation granularity, future work should contribute to locating the break-even point of the coding level of detail, i.e., at which number of opinion classes does the inter-rater reliability of non-experts diverge and collapse. The follow-up question then is if that break-even point, i.e., that level of detail in the annotations, is meaningful enough for directed content analysis. Even though related work already used crowdsourcing for such content studies, these works do not inform about the actual quality and consistency of the predicted opinion classes. We do not see a systematic review and methodology at the overlap of opinion mining and directed content analysis, which would set a standard or purity and consistency in order to conciliate qualitative research with machine learning (ML). This connection, however, is overdue due to the quickly developing field of natural language processing (NLP).

In that regard, there are methods specifically worth of being mentioned, since they come from a rather new branch of NLP research which to our best knowledge is missing any application-related evaluation. Recently, *weakly supervised learning* gained a lot of attention [Ratner et al., 2019]. Its idea is to release the text miner from the burden of collecting annotations by instead using just a small bit of her domain expertise to inform the algorithm for clustering and classification [Angelidis and Lapata, 2018b]. In short terms, providing a small number of words for each class should suffice to provide a meaningful classifier for topic, entity, or aspect classification of an opinion expression [Karamanolakis et al., 2019]. This has an application in qualitative content analysis and particularly in summative content analysis, where pre-defined keywords are being counted in a target data domain. Modern weakly supervised algorithms include universal world knowledge via transfer learning and add contextual meaning and coherence to the concepts expressed by keywords. Here, we see possible contributions in implementing expert-guided qualitative content studies with a more coherent outcome — also with applications to multiple languages at once.

Bibliography

Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yaser Jararweh, and Omar Qawasmeh. Enhancing aspect-based sentiment analysis of arabic hotels' reviews using morphological, syntactic and semantic features. *Information Processing & Management*, 56(2):308–319, 2019. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2018.01.006. URL https://www.sciencedirect.com/science/article/pii/S0306457316305623. Advance Arabic Natural Language Processing (ANLP) and its Applications.

Md. Hijbul Alam, Woo-Jong Ryu, and SangKeun Lee. Joint multi-grain topic sentiment: modeling semantic aspects for online reviews. *Information Sciences*, 339:206–223, 2016. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2016.01.013. URL https://www.sciencedirect.com/science/article/pii/S0020025516000153.

Stefanos Angelidis and Mirella Lapata. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. *arXiv preprint arXiv:1808.08858*, 2018a.

Stefanos Angelidis and Mirella Lapata. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. *CoRR*, abs/1808.08858, 2018b. doi: 10.18653/v1/D18-1403. URL https://arxiv.org/pdf/1808.08858.pdf.

Noureddine Azzouza, Karima Akli-Astouati, and Roliana Ibrahim. Twitterbert: Framework for twitter sentiment analysis based on pre-trained language model representations. In Faisal Saeed, Fathey Mohammed, and Nadhmi Gazem, editors, *Emerging Trends in Intelligent Computing and Informatics*, pages 428–437, Cham, 2020. Springer International Publishing. ISBN 978-3-030-33582-3.

Paul Baker, Costas Gabrielatos, Majid KhosraviNik, Michał Krzyżanowski, Tony McEnery, and Ruth Wodak. A useful methodological synergy? combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the uk press. *Discourse & Society*, 19(3):273–306, 2008. doi: 10.1177/0957926508088962. URL https://doi.org/10.1177/0957926508088962.

Aneesha Bakharia. On the equivalence of inductive content analysis and topic modeling. In Brendan Eagan, Morten Misfeldt, and Amanda Siebert-Evenstone, editors, *Advances in Quantitative Ethnography*, pages 291–298, Cham, 2019. Springer International Publishing. ISBN 978-3-030-33232-7.

Bitty Balducci and Detelina Marinova. Unstructured data in marketing. *J. Acad. Mark. Sci.*, 46(4): 557–590, 2018. ISSN 1552-7824. doi: 10.1007/s11747-018-0581-x.

Corey Basch, Anthony Menafro, Jen Mongiovi, Grace Clarke Hillyer, and Charles Basch. A content analysis of youtube videos related to prostate cancer. *American Journal of Men's Health (AJMH)*, 2017.

Subhasree Basu, Yi Yu, and Roger Zimmermann. Fuzzy clustering of lecture videos based on topic modeling. In *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2016.

Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, 2006.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Molly Burke. Online review stats for 2020: The top 7 reasons consumers leave reviews, Mar 2021. URL https://www.junglescout.com/blog/online-review-statistics/.

Bob Carpenter. Multilevel bayesian models of categorical data annotation, 2008.

Dave Chaffey. Global social media statistics research summary 2022, Oct 2021. URL https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/.

Jenn Chen. 36 essential social media marketing statistics to know for 2021, Feb 2021. URL https://sproutsocial.com/insights/social-media-statistics/.

Yong Chen, Hui Zhang, Rui Liu, Zhiwen Ye, and Jianying Lin. Experimental explorations on short text topic mining between lda and nmf based schemes. *Knowledge-Based Systems*, 163:1–13, 2019.

Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Learning cross-lingual sentence representations via a multi-task dual-encoder model. *arXiv preprint arXiv:1810.12836*, 2018.

S. Cunningham-Nelson, M. Baktashmotlagh, and W. Boles. Visualizing student opinion through text analysis. *IEEE Transactions on Education*, 62(4):305–311, 2019.

Hannah Danner and Luisa Menapace. Using online comments to explore consumer beliefs regarding organic food in german-speaking countries and the united states. *Food Quality and Preference*, 83:103912, 2020a.

Hannah Danner and Luisa Menapace. Using online comments to explore consumer beliefs regarding organic food in german-speaking countries and the united states. *Food Quality and Preference*, 83:103912, 2020b.

Hannah Danner and John Thøgersen. Does online chatter matter for consumer behaviour? A priming experiment on organic food. *Int. J. Consum. Stud.*, 2021. ISSN 14706423. doi: 10.1111/ijcs.12732.

Priyanka Das, Asit Kumar Das, Janmenjoy Nayak, Danilo Pelusi, and Weiping Ding. A graph based clustering approach for relation extraction from crime data. *IEEE Access*, 2019.

DeepLearning.AI. When should you use an end-to-end learning system, and when should you not?, Dec 2018. URL https://twitter.com/deeplearningai_/status/1070414567467937792.

Kerstin Denecke, Mikalai Tsytsarau, Themis Palpanas, and Marko Brosowski. Topic-related sentiment analysis for discovering contradicting opinions in weblogs. Technical report, University of Trento, 2009.

DW Deutsche Welle. Corona sorgt für beispiellosen boom beim onlinehandel: Dw: 26.01.2021, January 2021. URL https://www.dw.com/de/corona-sorgt-f%C3%BCr-beispiellosen-boom-beim-onlinehandel/a-56348180.

Matthias Eickhoff and Runhild Wieneke. Understanding topic models in context: a mixed-methods approach to the meaningful analysis of large document collections. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

Florian Eyben, Matthias Unfried, Gerhard Hagerer, and Björn W. Schuller. Automatic multi-lingual arousal detection from voice applied to real product testing applications. In *ICASSP*, pages 5155–5159. IEEE, 2017. ISBN 978-1-5090-4117-6. URL http://dblp.uni-trier.de/db/conf/icassp/icassp2017.html#EybenUHS17.

Paul W Farris, Neil Bendle, Phillip E Pfeifer, and David Reibstein. *Marketing metrics: The definitive guide to measuring marketing performance*. Pearson Education, 2010.

Conor Gallagher, Eoghan Furey, and Kevin Curran. The application of sentiment analysis and text analytics to customer experience reviews to understand what customers are really saying. *International Journal of Data Warehousing and Mining (IJDWM)*, 15(4):21–47, 2019.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. Beyond the stars: Improving rating predictions using review text content. In *12th International Workshop on the Web and Databases, WebDB 2009, Providence, Rhode Island, USA, June 28, 2009*, 2009. URL http://webdb09.cse.buffalo.edu/papers/Paper9/WebDB.pdf.

Steven P. Gaskin, Abbie Griffin, John R. Hauser, Gerald M. Katz, and Robert L. Klein. *Voice of the Customer*, pages 1–27. American Cancer Society, 2010. ISBN 9781444316568. doi: https://doi.org/10.1002/9781444316568.wiem05020. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444316568.wiem05020.

Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. Multimedia lab @ ACL WNUT NER shared task: Named entity recognition for Twitter microposts using distributed word representations. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 146–153, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4322. URL https://aclanthology.org/W15-4322.

Andrea Gorra. *An analysis of the relationship between individuals? perceptions of privacy and mobile phone location data - a grounded theory study*. PhD thesis, Leeds Metropolitan University, April 2007. URL https://eprints.leedsbeckett.ac.uk/id/eprint/1554/.

Swapna Gottipati, Venky Shankararaman, and Sandy Gan. A conceptual framework for analyzing students' feedback. In *2017 IEEE Frontiers in Education Conference (FIE)*, pages 1–8. IEEE, 2017.

Niku Grönberg, Antti Knutas, Timo Hynninen, and Maija Hujala. An online tool for analyzing written student feedback. In *Proceedings of the 20th Koli Calling International Conference on Computing Education Research*, pages 1–2, 2020.

Niku Grönberg. Palaute : an online tool for text mining course feedback using topic modeling and emotion analysis. Master's thesis, LUT University, School of Engineering Science, Computer Science, 2020.

Vassia Gueorguieva. Voters, myspace, and youtube: The impact of alternative communication channels on the 2006 election cycle and beyond. *Social Science Computer Review*, 2008.

Akshay Hallur. 35+ surprising quora statistics that are hard to ignore, Aug 2020. URL https://bloggingx.com/quora-statistics/.

Ken Harrenstien. Automatic captions in youtube, Nov 2009. URL https://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html. accessed on 29. April 2021.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397, Vancouver, Canada, July 2017a. Association for Computational Linguistics. doi: 10.18653/v1/P17-1036.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397, 2017b.

Chong Ho Yu, Angel Jannasch-Pennell, and Samuel DiGangi. Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. *The Qualitative Report*, 16(3):730–744, 2011. URL https:/suworks.nova.edu/tqr/vol16/iss3/6.

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

Diane Holt and Ralf Barkemeyer. Media coverage of sustainable development issues - attention cycles or punctuated equilibrium? *Sustain. Dev.*, 20(1):1–17, 2012. ISSN 09680802. doi: 10.1002/sd.460.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, 2013. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

Hsiu-Fang Hsieh and Sarah E. Shannon. Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9):1277–1288, 2005. doi: 10.1177/1049732305276687. URL https://doi.org/10.1177/1049732305276687. PMID: 16204405.

Maija Hujala, Antti Knutas, Timo Hynninen, and Heli Arminen. Improving the quality of teaching by utilising written student feedback: A streamlined process. *Computers & Education*, 157:103965, 2020.

Sana Imaad. The 'how' and 'why' of getting product reviews on amazon in 2020, Aug 2020. URL https://www.visualistan.com/2020/08/the-how-and-why-of-getting-product-reviews-on-amazon-in-2020.html.

Joby John. *Fundamentals of customer-focused management: Competing through service*. Penn State Press, 2003.

Immanuel Kant. *Allgemeine Naturgeschichte und Theorie des Himmels, oder Versuch von der Verfassung und dem mechanischen Ursprunge des ganzen Weltgebäudes: nach Newtonischen Grundsätzen abgehandelt*. Petersen, Johann Friederich, 1755.

Immanuel Kant. *Critik der praktischen Vernunft*, volume 1. Hartknoch, Johann Friedrich, Riga, 1788.

Immanuel Kant. *Grundlegung zur metaphysik der sitten*, volume 28. L. Heimann, 1870.

Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. Leveraging just a few keywords for fine-grained aspect detection through weakly supervised co-training. *CoRR*, abs/1909.00415, 2019. URL https://arxiv.org/abs/1909.00415.

Ashish Khetan, Zachary C. Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data, 2017.

Hyunsoo Kim, Haesun Park, and Barry L. Drake. Extracting unrecognized gene relationships from the biomedical literature via matrix factorizations. *BMC Bioinformatics*, 8(9):S6, Nov 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-S9-S6. URL https://doi.org/10.1186/1471-2105-8-S9-S6.

DK Kirange and Ratnadeep R Deshmukh. Aspect based sentiment analysis semeval-2014 task 4. *Asian Journal of Computer Science And Information Technology*, pages 72–75, 2014.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1611835114.

Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Björn Schuller, Kam Star, Elnar Hajiyev, and Maja Pantic. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1022–1040, 2021. doi: 10.1109/TPAMI.2019.2944808.

Stefano Leone. Tripadvisor european restaurants, May 2021. URL https://www.kaggle.com/stefanoleone992/tripadvisor-european-restaurants.

Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

Stewart Lockie. Capturing the sustainability agenda: Organic foods and media discourses on food scares, environment, genetic engineering, and health. *Agric. Hum. Values*, 23(3):313–323, 2006. doi: 10.1007/s10460-006-9007-3.

Ling Luo, Xiang Ao, Yan Song, Jinyao Li, Xiaopeng Yang, Qing He, and Dong Yu. Unsupervised neural aspect extraction with sememes. In Sarit Kraus, editor, *IJCAI*, pages 5123–5129. ijcai.org, 2019a. URL http://dblp.uni-trier.de/db/conf/ijcai/ijcai2019.html#LuoASLYHY19.

Yunjuan Luo, Hansel Burley, Alexander Moe, and Mingxiao Sui. A meta-analysis of news media's public agenda-setting effects, 1972-2015. *Journal. Mass Comm. Q.*, 96(1):150–172, 2019b. ISSN 1077-6990. doi: 10.1177/1077699018804500.

Dehong Ma, Sujian Li, Fangzhao Wu, Xing Xie, and Houfeng Wang. Exploring sequence-to-sequence learning in aspect term extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3538–3547, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1344. URL https://aclanthology.org/P19-1344.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015.

Vinaya Manchaiah, Monica L Bellon-Harn, Marcella Michaels, and Eldré W Beukes. A content analysis of youtube videos related to hearing aids. *Journal of the American Academy of Audiology*, 2020.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, page 171–180, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595936547. doi: 10.1145/1242572.1242596. URL https://doi.org/10.1145/1242572.1242596.

Courtney Meyers and Katie Abrams. Feeding the debate: a qualitative framing analysis of organic food news media coverage. *J. Appl. Commun.*, 94(3-4):22–37, 2010. ISSN 10510834.

Mohamed Morchid and Georges Linarès. A lda-based method for automatic tagging of youtube videos. In *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*. IEEE, 2013.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*. ACM, 2011.

Kylie Mowbray-Allen. Social media statistics 2019, Oct 2019. URL https://www.hellomedia.team/blogs/blog/social-media-statistics-infographic.

Martin Müller, Marcel Salathé, and Per Kummervold. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter, 2020. URL https://europepmc.org/article/PPR/PPR268640.

Zarmeen Nasim, Quratulain Rajput, and Sajjad Haider. Sentiment analysis of student feedback using machine learning and lexicon based approaches. In *International conference on research and innovation in information systems*, pages 1–6. IEEE, 2017.

W. Russell Neuman, Lauren Guggenheim, S. Mo Jang, and Soo Young Bae. The dynamics of public attention: Agenda-setting theory meets big data. *J. Commun.*, 64(2):193–214, 2014. ISSN 00219916. doi: 10.1111/jcom.12088.

Hy Nguyen and Kiyoaki Shirai. A joint model of term extraction and polarity classification for aspect-based sentiment analysis. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 323–328, 2018. doi: 10.1109/KSE.2018.8573340.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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 188–197, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1018. URL https://aclanthology.org/D19-1018.

Jakob Nielsen. *Usability engineering*. Morgan Kaufmann, 1994.

Gokarn Ila Nitin, Swapna Gottipati, and Venky Shankararaman. Analyzing educational comments for topics and sentiments: A text analytics approach. In *2015 IEEE Frontiers in Education Conference (FIE)*, pages 1–9. IEEE Computer Society, 2015. ISBN 978-1-4799-8454-1. doi: 10.1109/FIE.2015.7344296. URL https://ieeexplore.ieee.org/xpl/conhome/7344010/proceeding.

Lawrence A. Palinkas, Sapna J. Mendon, and Alison B. Hamilton. Innovations in mixed methods evaluations. *Annual Review of Public Health*, 40(1):423–442, 2019. doi: 10.1146/annurev-publhealth-040218-044215. URL https://doi.org/10.1146/annurev-publhealth-040218-044215. PMID: 30633710.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*, 2004.

Ioannis Pavlopoulos. Aspect based sentiment analysis. *Athens University of Economics and Business*, 2014.

Andreas Peldszus and Manfred Stede. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 196–204, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 486–495, 2015a.

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 486–495, 2015b.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30, 2016.

Alex Ratner, Stephen Bach, Paroma Varma, and Chris Ré. Weak supervision: the new programming paradigm for machine learning. *Hazy Research. Accessed 4 Dec 2021.*, 4 2019. URL http://ai.stanford.edu/blog/weak-supervision/.

Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46, 2015. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2015.06.015. URL https://www.sciencedirect.com/science/article/pii/S0950705115002336.

Filipe Rodrigues and Francisco C. Pereira. Deep learning from crowds. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):1611–1618, Apr. 2018.

Sara Rosenthal, Noura Farra, and Preslav Nakov. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2088. URL https://aclanthology.org/S17-2088.

Frank Rosner, Alexander Hinneburg, Michael Röder, Martin Nettling, and Andreas Both. Evaluating topic coherence measures. *arXiv preprint arXiv:1403.6397*, 2014.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Shihab Elbagir Saad and Jing Yang. Twitter sentiment analysis based on ordinal regression. *IEEE Access*, 7:163677–163685, 2019. doi: 10.1109/ACCESS.2019.2952127.

Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. On stopwords, filtering and data sparsity for sentiment analysis of twitter. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, and et al, editors, *LREC 2014, Ninth International Conference on Language Resources and Evaluation. Proceedings*, pages 810–817, 2014. ISBN 978-2-9517408-8-4. The LREC 2014 Proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License; 9th International Conference on Language Resources and Evaluation, LREC 2014 ; Conference date: 26-05-2014 Through 31-05-2014.

Vivek Kumar Singh, Rajesh Piryani, Ashraf Uddin, and Pranav Waila. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, pages 712–717. IEEE, 2013.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Morristown, NJ, USA, 2008. Association for Computational Linguistics.

Sharan Srinivas and Suchithra Rajendran. Topic-based knowledge mining of online student reviews for strategic planning in universities. *Computers & Industrial Engineering*, 128:974–984, 2019. ISSN 0360-8352.

Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchen Du, Felix Hafner, Lea Schumann, Adria Mallol-Ragolta, Bjoern W. Schuller, Iulia Lefter, Erik Cambria, and Ioannis Kompatsiaris. Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: Emotional car reviews in-the-wild. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop*. ACM, 2020.

Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva-Maria Meßner, Erik Cambria, Guoying Zhao, and Björn W. Schuller. The muse 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, page 5–14, New York, NY, USA, 2021a. Association for Computing Machinery. ISBN 9781450386784. doi: 10.1145/3475957.3484450. URL https://doi.org/10.1145/3475957.3484450.

Lukas Stappen, Alice Baird, Michelle Lienhart, Annalena Bätz, and Björn Schuller. An estimation of online video user engagement from features of continuous emotions. *arXiv preprint arXiv:2105.01633*, 2021b.

Lukas Stappen, Alice Baird, Lea Schumann, and Björn Schuller. The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements. *IEEE Transactions on Affective Computing*, 2021c.

Foundation Team. 21 quora statistics marketers need to know for 2021, Jan 2021. URL https://foundationinc.co/lab/quora-statistics/.

Mike Thelwall. Gender bias in sentiment analysis. *Online Information Review*, 42(3):343–354, 2018.

John Thøgersen. Media attention and the market for 'green' consumer products. *Bus. Strategy Environ.*, 15(3):145–156, 2006. ISSN 0964-4733. doi: 10.1002/bse.521.

Paul Verhaar. Crowdsourcing Label Aggregation: Modeling task and worker correlation. https://labs.sogeti.com/crowdsourcing-label-aggregation-modeling-task-and-worker-correlation/, 2020. [Accessed 18-Oct-2021].

G Vinodhini and RM Chandrasekaran. Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6):282–292, 2012.

Fabian L. Wauthier and Michael I. Jordan. Bayesian bias mitigation for crowdsourcing. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *NIPS*, volume 24, pages 1800–1808, San Diego, CA, 2011. Curran Associates, Inc.

Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. The multidimensional wisdom of crowds. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, Red Hook, NY, USA, 2010. Curran Associates Inc.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *NIPS*, pages 2035–2043, San Diego, CA, 2009. Curran Associates, Inc. ISBN 9781615679119.

Gregor Wiedemann. Opening up to big data: Computer-assisted analysis of textual data in social sciences. *Historical Social Research / Historische Sozialforschung*, 38(4 (146)):332–357, 2013. ISSN 01726404. URL http://www.jstor.org/stable/24142701.

Anne Wiese, Julian Kellner, Britta Lietke, Waldemar Toporowski, and Stephan Zielke. Sustainability in retailing – a summative content analysis. *International Journal of Retail & Distribution Management*, 40(4):318–335, Jan 2012. ISSN 0959-0552. doi: 10.1108/09590551211211792. URL https://doi.org/10.1108/09590551211211792.

Jochen Wirtz and John E.G. Bateson. An experimental investigation of halo effects in satisfaction measures of service attributes. *International Journal of Service Industry Management*, 6(3): 84–102, Jan 1995. ISSN 0956-4233. doi: 10.1108/09564239510091358. URL https://doi.org/10.1108/09564239510091358.

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 2013.

Yuanbin Wu, Qi Zhang, Xuan-Jing Huang, and Lide Wu. Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1533–1541, 2009.

Zhen Wu, Chengcan Ying, Xinyu Dai, Shujian Huang, and Jiajun Chen. Transformer-based multi-aspect modeling for multi-aspect multi-sentiment analysis. In Xiaodan Zhu, Min Zhang, Yu Hong, and Ruifang He, editors, *Natural Language Processing and Chinese Computing*, pages 546–557, Cham, 2020. Springer International Publishing. ISBN 978-3-030-60457-8.

Wei Xue, Wubai Zhou, Tao Li, and Qing Wang. MTNA: A neural multi-task model for aspect category classification and aspect term extraction on restaurant reviews. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 151–156, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-2026.

Yan Yan, Rómer Rosales, Glenn Fung, Mark W. Schmidt, Gerardo Hermosillo Valadez, Luca Bogoni, Linda Moy, and Jennifer G. Dy. Modeling annotator expertise: Learning when everybody knows a bit of something. In Yee Whye Teh and D. Mike Titterington, editors, *AISTATS*, volume 9 of *JMLR Proceedings*, pages 932–939, Chia Laguna Resort, Sardinia, Italy, 2010. JMLR.org.

Inc. Yelp. Yelp dataset, Mar 2021. URL https://www.kaggle.com/yelp-dataset/yelp-dataset.

Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60(2):617–663, Aug 2019. ISSN 0219-3116. doi: 10.1007/s10115-018-1236-4. URL https://doi.org/10.1007/s10115-018-1236-4.

Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV (13)*, volume 11217 of *Lecture Notes in Computer Science*, pages 227–243, Red Hook, NY, USA, 2018. Springer. ISBN 978-3-030-01261-8.

Part II

Publications Relevant for Examination

5 Dataset: Social Media Discourse About Organic Food

5.1 A Case Study and Qualitative Analysis of Simple Cross-Lingual Opinion Mining

The publication on the consecutive pages is relevant to the examination. It was accepted after peer-review as full paper at the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, where it received the best student paper award. Gerhard Johann Hagerer, the author of the present thesis, is the first author of that paper. His author role, the reprinting permission, and his following contributions are acknowledged by all authors Hagerer, Leung, Danner, and Groh [2021d]:

> *"Gerhard Johann Hagerer headed the research project. He developed the research idea, the concept, and the methodology of the paper. Furthermore, he directed the implementation process and reviewed the source code deeply. Regarding the writing of the paper, he created the outline, directed the drafting, and wrote most of the paper, i.e., he wrote large textual parts, incorporated extensive reviewer feedback, and paraphrased, corrected, combined, and otherwise improved drafted material."*

Publication Summary

The following publication accounts for our organic social media dataset being available in multiple languages. Thus, cross-lingual pre-trained deep neural network has been utilized successfully for unsupervised topic modeling on that data. German and English texts are clustered coherently in a joint, single, integrated topic model. The study shows based on examples that the underlying data is correctly summarized. Further, multi-lingual sentiment analysis is applied, which makes the model suitable for opinion mining purposes. While being applied on sentence level, the topic and sentiment distributions of whole documents, comment sections, and even corpora can be depicted and compared with one another. The model stays consistent with varying number of topics and is thus flexible and adaptive in its resolution. In the context of the present thesis, this is a groundwork for our content study investigating the influence of the media agenda onto opinion formation [Danner et al., 2022].

A Case Study and Qualitative Analysis of Simple Cross-lingual Opinion Mining

Gerhard Hagerer[1][a], Wing Sheung Leung[1], Qiaoxi Liu[1], Hannah Danner[2][b] and Georg Groh[1][c]
[1]*Social Computing Research Group, Department of Informatics, Technical University of Munich, Germany*
[2]*Chair of Marketing and Consumer Research, TUM School of Management, Technical University of Munich, Germany*

Keywords: Opinion Mining, Topic Modeling, Sentiment Analysis, Cross-lingual, Multi-lingual, Market Research.

Abstract: User-generated content from social media is produced in many languages, making it technically challenging to compare the discussed themes from one domain across different cultures and regions. It is relevant for domains in a globalized world, such as market research, where people from two nations and markets might have different requirements for a product. We propose a simple, modern, and effective method for building a single topic model with sentiment analysis capable of covering multiple languages simultaneously, based on a pre-trained state-of-the-art deep neural network for natural language understanding. To demonstrate its feasibility, we apply the model to newspaper articles and user comments of a specific domain, i.e., organic food products and related consumption behavior. The themes match across languages. Additionally, we obtain an high proportion of stable and domain-relevant topics, a meaningful relation between topics and their respective textual contents, and an interpretable representation for social media documents. Marketing can potentially benefit from our method, since it provides an easy-to-use means of addressing specific customer interests from different market regions around the globe. For reproducibility, we provide the code, data, and results of our study[a].

[a]https://github.com/apairofbigwings/cross-lingual-opinion-mining

1 INTRODUCTION

Topic modeling on social media texts is difficult, since lack of data as well as spelling and grammatical errors can make the approach unfeasible. Dealing with multiple languages at the same time adds more complexity to the problem which oftentimes makes the approach unusable for domain experts. Thus, we propose a cross-lingually pre-trained deep neural network as a black box with very little textual preprocessing necessary before embedding the texts and forming their clustering and topic distributions.

For our method, we leverage current research regarding multi-lingual topic modeling, see Section 2. We provide an extensive description of a simple method to support domain experts from specific social media domains in its application in Section 3. We qualitatively demonstrate our topic model, its feasibility, and its cross-lingual semantic characteristics on English and German newspaper and social media texts in Section 4. We aim at inspiring pragmatic ideas to explore the potential for comparative, intercultural market research and agenda setting studies. Unsolved problems and future potential are given in Section 5.

2 RELATED WORK

Topic modeling is meant to learn thematic structure from text corpora. With probabilistic topic modeling methods, such as latent semantic indexing (LSI) (Deerwester et al., 1990) or latent Dirichlet allocation (LDA) (Blei et al., 2003), researchers try to extend the capabilities of topic modeling for application from a single language to multiple languages. Using multilingual dictionaries and translated corpora is an intuitive way to tackle cross-lingual topic modeling problems (Zhang et al., 2010; Vulić et al., 2013). Further examples exist for topic modeling with either dictionaries or translation text collections (Gutiérrez et al., 2016; Boyd-Graber and Blei, 2012; Jagarlamudi and

[a] https://orcid.org/0000-0002-2292-0399
[b] https://orcid.org/0000-0001-8387-0818
[c] https://orcid.org/0000-0002-5942-2297

Hagerer, G., Leung, W., Liu, Q., Danner, H. and Groh, G.
A Case Study and Qualitative Analysis of Simple Cross-lingual Opinion Mining.
DOI: 10.5220/0010649500003064
In *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2021) - Volume 1: KDIR*, pages 17-26
ISBN: 978-989-758-533-3; ISSN: 2184-3228

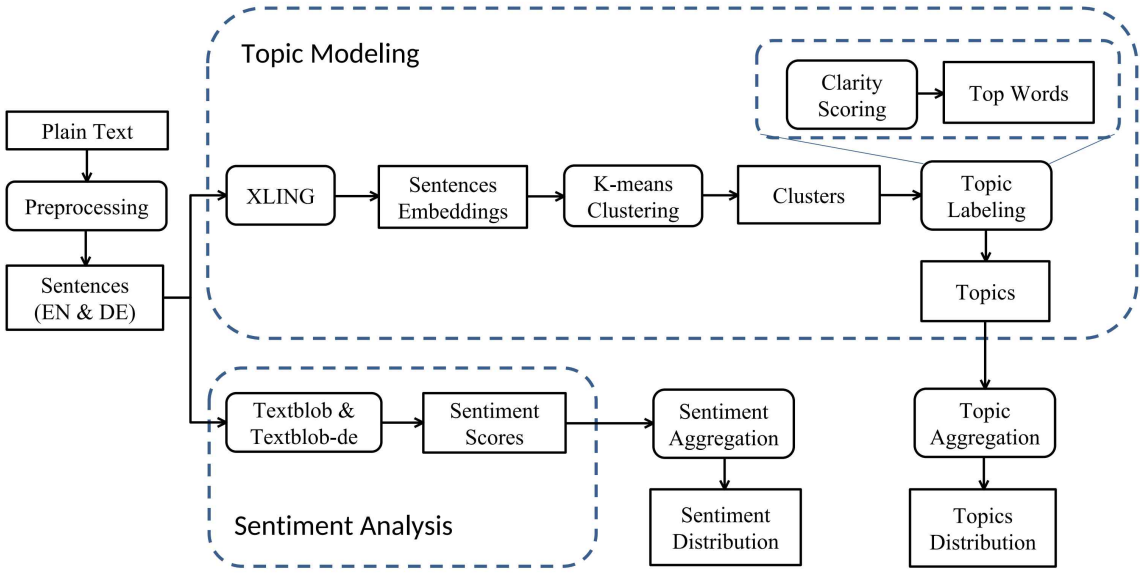


Figure 1: Plain text is first tokenized into sentences and passed to topic modeling and sentiment analysis. Topic modeling involves (1) converting sentences of both languages into embeddings with XLING, (2) clustering all embeddings with K-means and (3) deriving a topic label of each cluster. Sentiment analysis is performed using Textblob. Topic and sentiment scores are aggregated for the analysis.

Daumé, 2010). However, this puts dependence on the availability of dictionary or good quality of translations. Significant manual labor and verification are required to prevent deteriorating noise.

Recently, methods converting words to vectors according to their semantics are widely adopted (Mikolov et al., 2013). Several studies showed text embeddings improve topic coherence (Bianchi et al., 2020; Srivastava and Sutton, 2017). Regarding multilinguality, embeddings in word level and sentence level enable text in different languages to be projected to the same vector space (Cer et al., 2018) such that semantically similar texts are clustered together independently of their languages. This favors studies on multi-lingual topic modeling without relying on dictionaries and translation (Xie et al., 2020; Chang et al., 2018). Although providing highly coherent topics, a recreation of word spaces is required when new text corpora are introduced. In our scenario, these limitations are not present.

Regarding the application of topic modeling, various social media corpora are studied by domain experts (Tsur et al., 2015; Ko et al., 2018). They covered on different domains, such as politics, marketing, and public health. Regarding media agenda setting, (Field et al., 2018) studied on how much degree a Russian newspaper related to economic downturn. They also *"introduced embedding-based methods for cross-lingually projecting English frame to Russian"* based on Media Frames Corpus. In contrast, we propose a straightforward topic modeling method without fine-tuning but only clustering necessary on a social media corpus. This enables further investigation on media agenda setting cross-lingually and cross-culturally.

3 TOPIC MODELING METHOD

Figure 1 shows the overall workflow of our topic modeling approach. We aim to conduct simple, cross-lingual topic modeling on user-generated content with no translation, dictionary, and parallel corpus required for aligning the semantic meanings across languages. Our approach solely depends on clustering sentence embeddings for topic modeling. Ready-made sentence representations simplify the approach, since these suppress too frequent, meaningless, and unimportant words automatically without the need to model that part explicitly (Kim et al., 2017).

3.1 Preprocessing

The raw texts of articles and comments are first tokenized into sentences with Natural Language Toolkit (NLTK). Then, URLs, specially for those enclosed with HTML `<a>` tag, are replaced with string 'url'. After that, sentences with character length smaller than 15 are omitted to minimize noise, since they appear inscrutable and they are only 6.6% out of all sen-

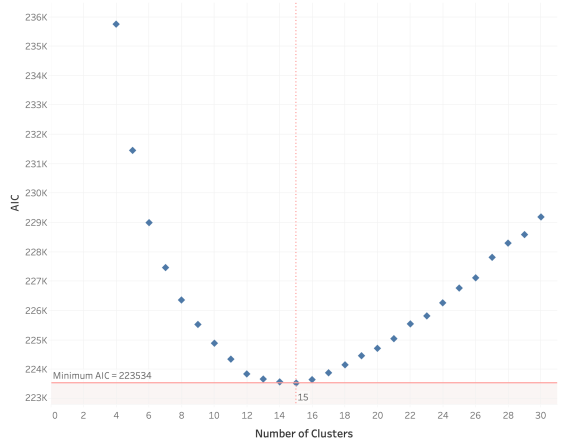


Figure 2: AIC plot indicates k = 15 is the global minimum.

tences which is a small portion. After preprocessing, there are 127,464 English sentences and 200,627 German sentences, i.e., total 328,091.

3.2 Cross-lingual Embeddings

In the following paragraph, we provide an explanation of the pre-trained XLING model, which we use for the present work, based on the words of the authors (Chidambaram et al., 2018). XLING calculates *"sentence embeddings that map text written in different languages, but with similar meanings, to nearby embedding space representations"*. Similarity is calculated mathematically as dot product between two sentence embeddings. In order to train the model, the authors *"employ four unique task types for each language pair in order to learn a function g"*, i.e., the eventual sentence-to-vector model. The architecture is based on a Transformer neural network (Vaswani et al., 2017) tailored for modeling multiple languages at once. The tasks on which the model is eventually trained are *"(i) conversational response prediction, (ii) quick thought, (iii) a natural language inference, and (iv) translation ranking as the bridge task"*. The data for training *"is composed of Reddit, Wikipedia, Stanford Natural Language Inference (SNLI), and web mined translation pairs"*.

3.3 Sentence Clustering

K-means clustering algorithm is implemented on both English and German sentence embeddings at the same time. Since XLING provides semantically aligned sentence embeddings of both languages, this joint clustering step helps to establish one topic model for two disjunct datasets irrespective of their language. Clustering is established for a varying number of clusters$k$, ranging from 1 to 30. Elbow method is first used for choosing the optimal k but the inertia (sum of squared distances of samples to their closest cluster center) of increasing k decreases rapidly at the beginning and then gently without a significant elbow point. Therefore, Akaike Information Criterion (AIC) is adopted and k = 15 is chosen as optimal value as it is the global minimum, see Figure 2. In Section 4.2, further discussion on topic coherence is conducted proving the fact that k = 15 resulted semantically coherent topics.

3.4 Topic Labeling

To be able to derive a meaningful topic label for each sentence cluster, the respective top words of each cluster are required. In order to get the top word list, the clarity score is adopted (Cronen-Townsend et al., 2002). According to (Angelidis and Lapata, 2018), it ranks terms with respect to their importance of each cluster c and language l, such that

$$\text{score}_{l,c}(w) = t_{l,c}(w) \log_2 \frac{t_{l,c}(w)}{t_l(w)}, \qquad (1)$$

where $t_{l,c}(w)$ and $t_l(w)$ are the l1-normalized tf-idf scores of the word w in the sentences within cluster c and in all sentences, respectively, for a certain language.

Additionally, stopword removal from the top word lists is also a concern when calculating the clarity score. Generally, stopwords are the most frequent words in the documents and sometimes they are too dominant such that they interfere with the result from clarity scoring. Thus, we remove domain-specific high frequency words for each language from corresponding topic top word lists.

Topics are labeled manually based on the English and German top word lists. The results are shown in Table 1 and will be discussed further in Section 4.2 evaluating topic coherence across languages.

3.5 Sentiment Analysis

In addition to topic modeling, we conduct sentiment analysis to investigate the feasibility and meaning of cross-lingual topic-related sentiments in articles and respective comment sections. We make use of Textblob[1] and Textblob-de[2] to assign each of the English and German pre-processed sentences a polarity score. The polarity assignment is first proposed by (Pang and Lee, 2004) and reimplemented by

[1]https://github.com/sloria/textblob
[2]https://github.com/markuskiller/textblob-de

Table 1: Top words for all meaningfull topics with k = 15 of English and German data.

Topic	English top words	German top words
Environment	pesticide, plant, soil, use, crop, fertilize, pesticide, garden, herbicide, grow	pflanze, pestizid, dunger, boden, gulle, garten, anbau, gemuse, tomate, feld
Retailers	store, whole, shop, groceries, supermarket, local, market, amazon, price, online, discount	aldi, supermarkt, lidl, kauf, laden, lebensmittel, cent, einkauf, wochenmarkt
GMO & organic	gmo, label, gmos, monsanto, product, certificate, usda, genetic, product	produkt, bioprodukt, lebensmittel, gesund, konventionell, biodiesel, herstellung, enthalt, monsanto, pestizid
Food products & taste	taste, milk, sugar, cook, eat, fresh, flavor, fruit, potato, sweet	kase, schmeckt, gurke, essen, analogkase, schmeckt, tomate, milch, geschmack, kochen
Food safety	chemical, cancer, body, acid, effect, cause, toxic, toxin, glyphosat, disease	dioxin, gift, grenzwert, ddt, menge, giftig, toxisch, substanz, chemisch, antibiotika
Research	science, study, scientific, research, gene, scientist, genetic, human, stanford, nature	gentechnik, natur, mensch, wissenschaft, lebenserwartung, genetisch, studie, menschlich, planet
Health & nutrition	eat, diet, healthy, nutritious, health, fat, calory, obesity, junk	lebensmittel, essen, ernahrung, gesund, nahrungsmittel, lebensmittel, nahrung, fett, billig
Politics & policy	govern, public, politic, corporate, regulation, law, obama, vote	politik, skandal, verantwortung, bundestag, schaltet, bestraft, strafe, kontrolle, kriminell
Animals & meat	meat, chicken, anim, cow, beef, egg, fed, raise, pig, grass	tier, fleisch, eier, huhn, schwein, futter, kuh, verunsichert, vergiftet, deutsch
Farming	farm, farmer, agriculture, land, sustain, crop, yield, acre, grow, local	landwirtschaft, landwirt, bau, flache, okologisch, nachhaltig, konventionell, landbau, produktion, ertrag
Prices & profit	price, consume, market, company, profit, product, cost, amazon, money	verbrauch, preis, produkt, billig, qualitat, kunde, kauf, geld, unternehmen, kosten

(De Smedt and Daelemans, 2012). Since the subjectivity assignment is not well-developed in Textblobde, we filter out sentences with polarity equals to 0 for both English and German sentences in order to derive comparable results.

3.6 Topic and Sentiment Distributions

After assigning a labeled cluster, i.e., a topic, and a sentiment score to each sentence of the corpus, we derive the corresponding distributions.

For topic distributions, all sentences from a document are counted per topic. The distribution is then normalized to be comparable. For sentiment distributions, all sentences from a document are grouped per topic. Topic-wise sentiment distribution is derived based on the sentence-wise polarity scores and the respective median and quartiles. A document in that regard is either an article or all of its comments, i.e., its comment section.

4 TOPIC COHERENCE

In this section, we evaluate the feasibility and semantic coherence of our cross-lingual topic modeling qualitatively. Instead of providing quantitative coherence scores, we aim at a detailed, qualitative analysis of textual examples. We depict representative sentences and words of each topic in subsection 4.2. We investigate to what extent these are semantically coherent, also across languages. We expose the ratio of coherent and incoherent topics and how it develops with increasing number of topics in in subsection 4.3. Eventually, we show the distribution of topics in selected newspaper articles and their respective comment sections to relate the discussed content with our actual topic model on English and German texts.

4.1 Data

The collection of the data used in this study is described in another publication (Danner et al., 2021) as follows. For the analysis we downloaded *"news arti-*

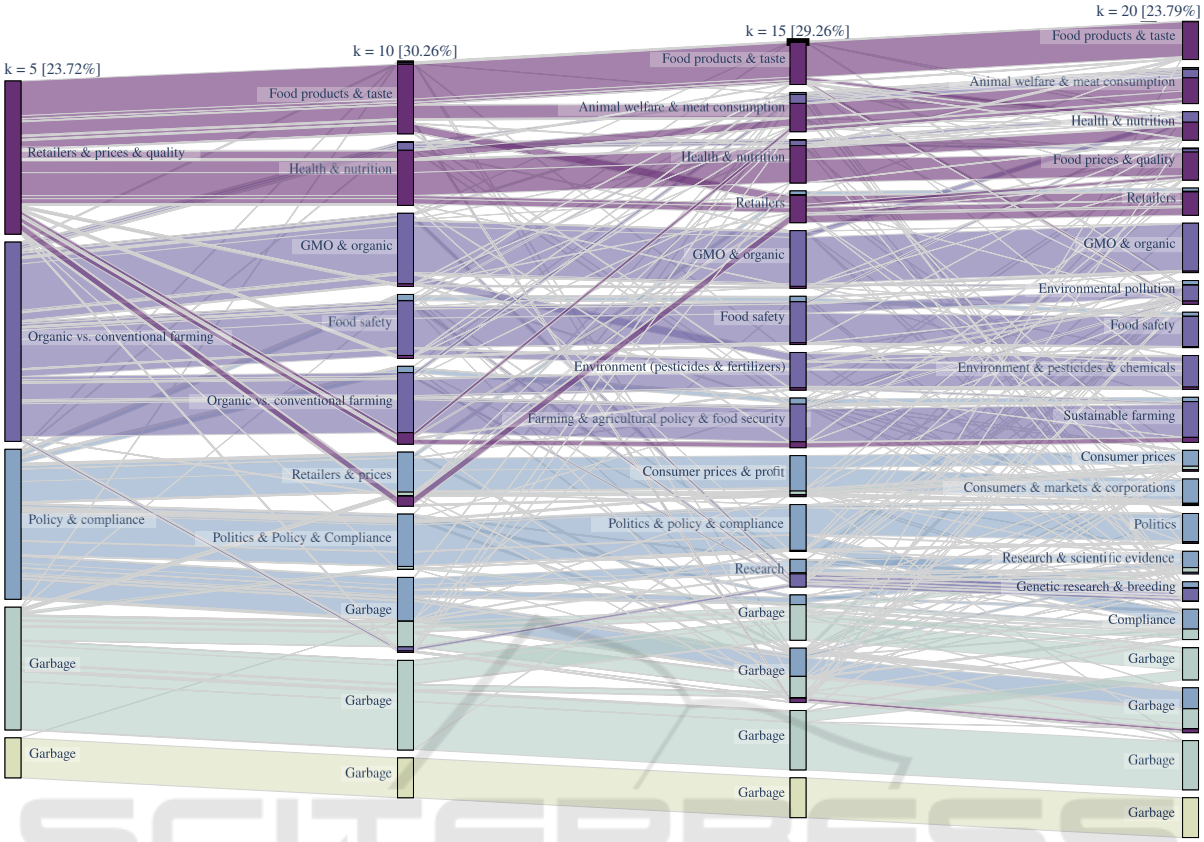


Figure 3: Topic distributions with increasing number of topics k. The percentage is the amount of sentences in garbage topics.

cles and reader comments of two major news outlets representative of the German and the United States (US) context", i.e., *spiegel.de* and *nytimes.com*. The creation dates of the texts are *"spanning from January 2007 to February 2020"*. *"Articles and related comments on the issue of organic food were identified using the search terms organic food and organic farming and the German equivalents*. For topic modeling, we utilized *"534 articles and 41,320 comments from the US for the years 2007 to 2020, and 568 articles and 63,379 comments from Germany for the years 2007 to 2017 and the year 2020"*.

4.2 Multi-linguality of Topics

In this section, we evaluate semantic coherence of our cross-lingual topic modeling by depicting the representative sentences and words for each topic and showing the semantic relation. Table 1 shows the first 10 English and German words having the highest clarity scores (see Section 3.4) in each cluster for $k = 15$. Table 2 shows the first 3 English and German sentences whose embeddings have the largest cosine similarity to their corresponding cluster centroids. Both top words and top sentences indicate that the clusters are grouped reasonably in terms of semantics. For example, this is the case for the topic *Environment (pesticides & fertilizers)* which is indeed related to use of pesticides in planting. Even though this also appears to be the case for the sentences in *GMO & organic* on the first glance, those are actually about organic food and how aspects such as GMO and pesticides relates to the food itself. This and the other representative top words and sentences indicate that clustering on cross-lingual sentence embeddings yield semantically coherent topics.

According to our analysis, top sentences from garbage clusters are always short in length with slightly more than 15 characters. Together with top words (Table 1), these hardly contribute to the organic food domain and corresponding entities. Thus, it is feasible in our case to ignore them.

4.3 Amount of Meaningful Topics

Besides providing coherent cross-lingual topics, our method performs well to distinguish usable from unusable topics, and it provides a constantly high number of relevant topics independently of the number of overall topics. Figure 3 is a Sankey diagram showing

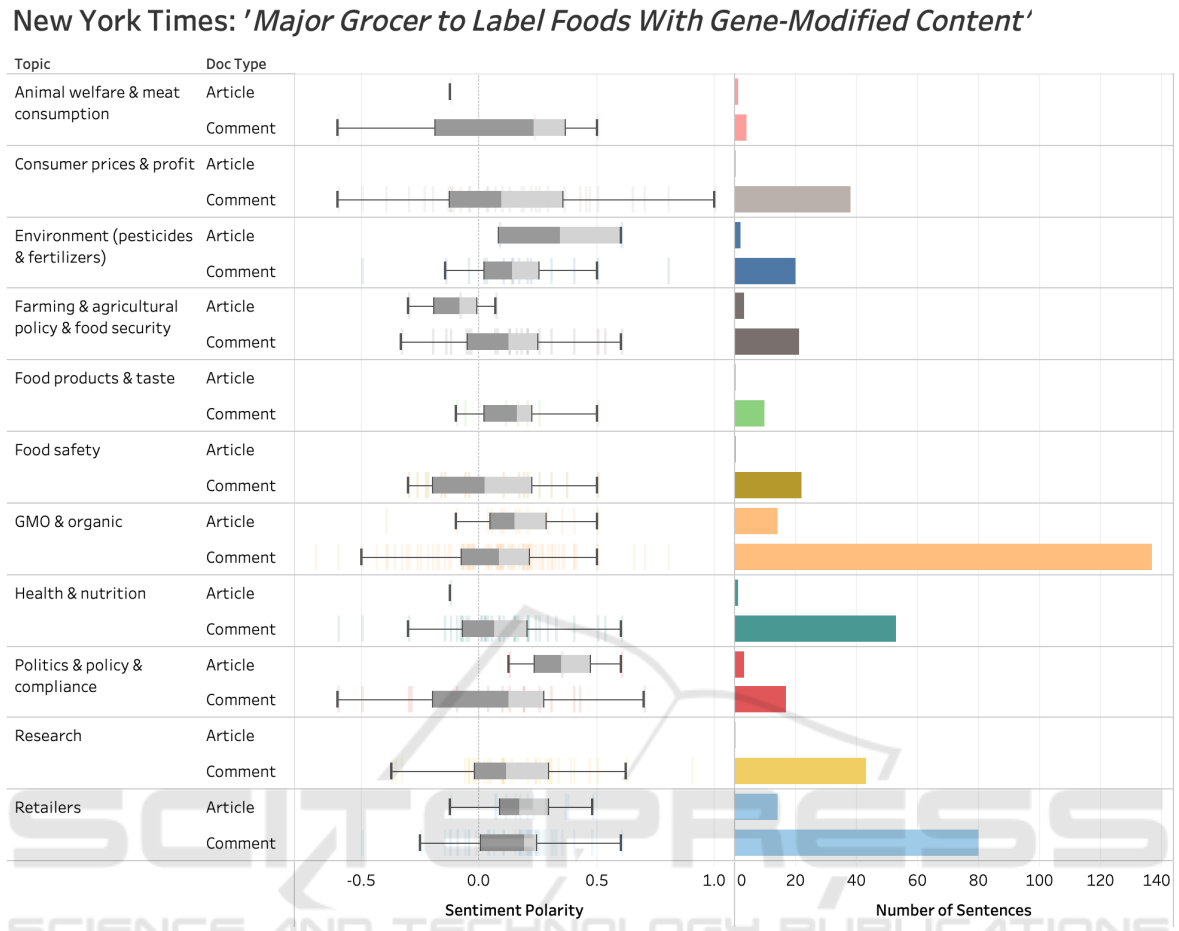


Figure 4: Topic and sentiment distribution for *Grocer*.

the flow of topic assignments for all English and German sentences with increasing number of clusters k. Topic modeling is performed for each k with all preprocessed sentences independently. It can be seen that more specific topics descend from general but related topics as indicated by the colors.

For instance, *GMO & organic*, *Food safety*, *Environment (pesticides & fertilizers)* and *Farming & agricultural policy & food security* for $k = 15$ are derived from *Organic vs. conventional farming* for $k = 5$. *Organic vs. conventional farming* in $k = 5$ generally focuses on advantages brought by organic farming when comparing to conventional farming, such as reducing persistent toxin chemicals from entering to environment, food, and bodies; thus, bio-products are recommended. For $k = 15$, the children topics are more specific. For example, *GMO & organic* shows the aims for having organic food, i.e., avoidance of GMO and poisoning with pesticides. Moreover, *Food safety* in $k = 15$ is further split into *Food safety* and *Environmental pollution*.

To see how the topics relate to their actual sentences, we try to observe the top sentences of each topic, i.e., those sentences of which the embeddings are closest to the centroid. Both English and German sentences are similar and share strong semantic similarity. The *food safety* topic focuses on the toxicity issue of dioxin and other chemicals towards consumers. *Environmental pollution*, which is further splitted from it, for $k = 20$ indeed tells contamination of water resources by chemicals. This shows that fine-grained topics and the way they develop with increasing k have a meaningful relation to ancestor and sister topics.

Sentences without contribution to the organic food domain always remain in garbage clusters in a way that the proportion of usable and unusable clusters does not fluctuate. Thus, the topic model maintains its coherence independent to the number of topics and the despite the fact that k-means is not deterministic in its clustering. This property is helpful, since the number of topics can be chosen as high as necessary to provide a sufficient level of detail for the domain of interest. Moreover, this highlights the meaningfulness

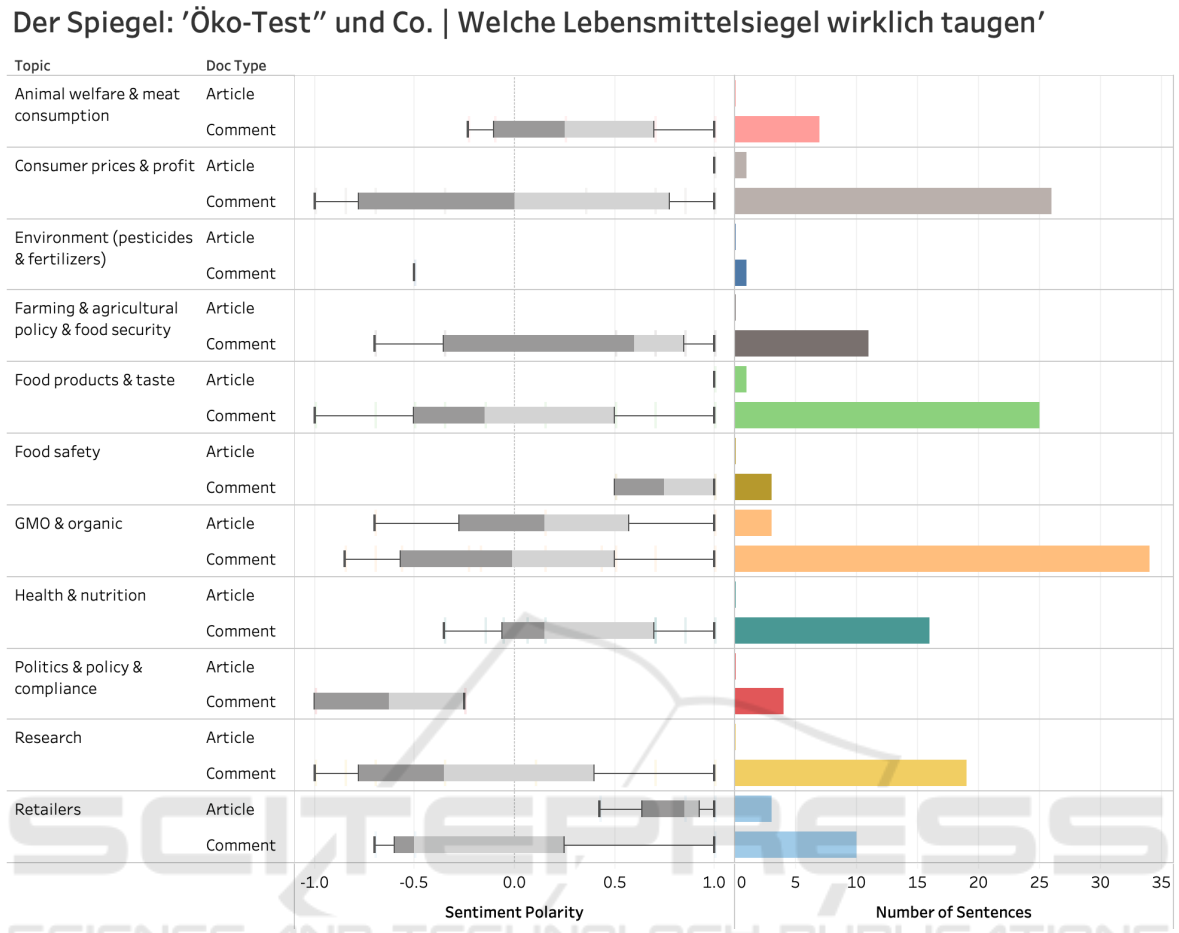


Figure 5: Topic and sentiment distribution for *Öko-Test*.

and robustness of the given sentence representations being able to separate noise from informative data in an unsupervised fashion.

4.4 Validation of Opinion Distributions

In this section, two real text examples are given to evaluate our method qualitatively. The first one is an article from New York Times, titled '*Major Grocer to Label Foods With Gene-Modified Content*'[3], hereafter referred to as *Grocer*. It reported that the first retailer in the United States announced to label all of its genetically modified food sold in its stores. Advocating and opposing stakeholders stated their arguments regarding different aspects. The second example is from Der Spiegel, titled '*"Öko-Test" und Co. – Welche Lebensmittelsiegel wirklich taugen*'[4], below denoted as *Öko-Test*. It reported that number of food claims, certifications, and seals in Germany were growing as organic labeling was a good promotional strategy indicating high food quality. However, consumers knew little about the details even when the tests for each label were transparent and well-documented. Based on these two summaries, it would be expected that topics related to supermarkets, retailers, and GMO labels are shown to be present in those articles. The *Grocer* article, however, expresses concerns about the consumption of genetically modified food, whereas *Öko-Test* discusses organic food labeling issues from various point of views, among others fair trading and organic fishing.

Topic Distribution. Figures 4 and 5 show the distribution of topics in the overall article sentences. It can be seen that the two topics *Retailers* and *GMO & organic* are mentioned the most in both articles, supporting our hypothesis. The comment section of the *Grocer* article corresponds to the article itself such that most of its sentences also talk about *GMO & organic* and the second most for *Retailers*. How-

[3]https://www.nytimes.com/2013/03/09/business/grocery-chain-to-require-labels-for-genetically-modified-food.html

[4]https://tinyurl.com/spiegel-lebensmittelsiegel

ever, the commenters of the *Öko-Test* articles commented more about *GMO & organic* followed with *Consumer prices & profit* and *Food products & taste*. Even though the dominating topics in German differ between article and comments, it can be stated that the topic distribution overall still refers to the actual topics of the given texts and domains. At the same time, differences in the distribution not only between article and comments but also between languages and thus cultures are directly visible, providing a means for clear comparability in several respects.

Sentiment Distribution. Figures 4 and 5 also show the sentiment distribution. Generally, the sentiment of the *Grocer* article spreads out less than that of the *Öko-Test* article. It is observed that, in topic *GMO & organic*, comments score sentiment polarity ranging between 0.50 to -0.70 in *Grocer* and between 1 and -0.85 in *Öko-Test*. This means sentences from *Grocer* show weaker sentiment compared to those from *Öko-Test*. The actual texts indicate that sentiment on our German data indeed has more variance than on English. Thus, the proposed multi-lingual sentiment analysis, Textblob and Textblob-de, appears to represent the data adequately in the given use case. However, it cannot be excluded that the sentiment distribution could be affected given the fact that two different but methodically similar frameworks are used. Different biases and variances could be caused by different models which have differences in the sentiment dictionary size and the subjectivity of human-assigned sentiment scores based on different cultures. Further studies should examine this problem for more robust, domain-independent multi-lingual sentiment prediction.

5 CONCLUSION

This case study shows that our technically simple approach successfully generates an high proportion of relevant and coherent topics for our domain, i.e., organic food products and related consumption behavior based on English and German social media texts. Moreover, the topics display the text contents correctly and support a domain expert in the content analysis of social media texts written in multiple languages.

However, the presented paper did not provide quantitative measurements of topic coherences and comparisons with the state-of-the-art. For mono-language topic modeling, it would be LDA (Blei et al., 2003); for advanced cross-lingual topic modeling, it could be attention-based aspect extraction (He et al., 2017) utilizing aligned multi-lingual word vectors (Conneau et al., 2017). Several multi-lingual datasets would need to be included for a representative comparison. Since pre-trained models trained on external data are used for the proposed method, it might be relevant for coherence score calculation to include intrinsic coherence scoring methods based on train test splits, such as, UMass coherence score (Mimno et al., 2011), and explore extrinsic methods calculated on external validation corpora, e.g., Wikipedia (Röder et al., 2015).

Regarding multi-lingual sentiment analysis, the difference in the sentiment analysis frameworks for different languages must be considered. For example, since two independent but similar sentiment analysis models are applied for English and German, the sentiment distribution could be affected. Therefore, future studies on developing and evaluating comparable sentiment models should be conducted.

REFERENCES

Angelidis, S. and Lapata, M. (2018). Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. *ArXiv*, abs/1808.08858.

Bianchi, F., Terragni, S., and Hovy, D. (2020). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Boyd-Graber, J. and Blei, D. (2012). Multilingual topic models for unaligned text.

Cer, D., Yang, Y., yi Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder.

Chang, C.-H., Hwang, S.-Y., and Xui, T.-H. (2018). Incorporating word embedding into cross-lingual topic modeling. *2018 IEEE International Congress on Big Data (BigData Congress)*, pages 17–24.

Chidambaram, M., Yang, Y., Cer, D., Yuan, S., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Learning cross-lingual sentence representations via a multi-task dual-encoder model. *arXiv preprint arXiv:1810.12836*.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Cronen-Townsend, S., Zhou, Y., and Croft, W. B. (2002). Predicting query performance.

Danner, H., Hagerer, G., Pan, Y., and Groh, G. (2021). The news media and its audience: Agenda-setting on organic food in the united states and germany.

De Smedt, T. and Daelemans, W. (2012). Pattern for

python. *The Journal of Machine Learning Research*, 13(1):2063–2067.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Field, A., Kliger, D., Wintner, S., Pan, J., Jurafsky, D., and Tsvetkov, Y. (2018). Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

Gutiérrez, E., Shutova, E., Lichtenstein, P., de Melo, G., and Gilardi, L. (2016). Detecting cross-cultural differences using a multilingual topic model. *Transactions of the Association for Computational Linguistics*, 4:47–60.

He, R., Lee, W. S., Ng, H. T., and Dahlmeier, D. (2017). An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397, Vancouver, Canada. Association for Computational Linguistics.

Jagarlamudi, J. and Daumé, H. (2010). Extracting multilingual topics from unaligned comparable corpora. In Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., and van Rijsbergen, K., editors, *Advances in Information Retrieval*, pages 444–456, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kim, H. K., Kim, H., and Cho, S. (2017). Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing*, 266:336–352.

Ko, N., Jeong, B., Choi, S., and Yoon, J. (2018). Identifying product opportunities using social media mining: Application of topic modeling and chance discovery theory. *IEEE Access*, 6:1680–1693.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Srivastava, A. and Sutton, C. (2017). Autoencoding variational inference for topic models.

Tsur, O., Calacci, D., and Lazer, D. (2015). A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1629–1638, Beijing, China. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Vulić, I., De Smet, W., and Moens, M.-F. (2013). Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Information Retrieval*, 16.

Xie, Q., Zhang, X., Ding, Y., and Song, M. (2020). Monolingual and multilingual topic analysis using lda and bert embeddings. *Journal of Informetrics*, 14(3):101055.

Zhang, D., Mei, Q., and Zhai, C. (2010). Cross-lingual latent topic extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1128–1137, Uppsala, Sweden. Association for Computational Linguistics.

APPENDIX

Table 2: Top sentences of meaningful topics from the whole dataset for k = 15 in English and German.

Topic	Top 3 sentences for English and German
Environment (pesticides, fertilizers)	Usually, the plant which uses conventional farming will produce the residue of the pesticides. – Some pesticides used in conventional farming, however, may reduce the level of resveratrol in plants. – Also, there is the question of naturally occurring pesticides produced by the plant itself.
	Viele Biopflanzen werden zwar nicht mit Pestiziden behandelt, Ihnen wird jedoch sehr viel mehr Wachstumsfläche zugestanden. – Nicht nur Biobauern benutzen Gülle, und Herbizide und Pestizide werden vor allem in der konventionellen Landwirtschaft eingesetzt. – Zum einen bauen sich die Pestizide und Herbizide relativ schnell ab, nicht zu verwechseln mit Überdüngung durch Gülle oder Belastung mit Schwermetallen.
Retailers	Whole Foods also sells a lot of high quality grocery items that aren't available elsewhere in a lot of places. – Whole Foods executives, however, say their supermarkets can be high quality, organic and natural but also inexpensive. – Larger competitors like Safeway and Kroger have vastly expanded their store-brand offerings of natural and organic products, and they are often cheaper than those at Whole Foods.
	Auch bei ALDI und CO lassen sich hochwertige Lebensmittel erwerben. – Wobei ich feststellen muss, dass andere Supermärkte - zumindest die, die ich frequentiere - auch Wert darauf legen, das gewisse Produkte aus der Region stammen, auch wenn sie konventionell hergestellt wurden. – Der Trend zur Feinkost beschert dem Handel vor allem in den Großstädten steigende Umsätze, wo Bio-Läden hip sind und die kaufkräftigen Kunden beim Einkaufen nicht auf jeden Cent schauen.
GMO & organic	In a nutshell, though, organic means the product meets a number of requirements, such as no GMOs, no non-organic pesticides, etc. – When consumers buy organic, they are guaranteed little more than food that is (in theory at least) produced without synthetic chemicals or G.M.O.'s (genetically modified organisms), and with some attention (again, in theory) to the health of the soil. – Organic food includes products that are grown without the use of synthetic fertilisers, sludge, irradiation, GMOs, or drugs, which already shows how much better it is for health.
	Jeder weiß doch, daß der Vorteil von bio nicht in der erhöhten Aufnahme von Nährstoffen gegenüber konventionellen Produkten liegt, sondern in der Vermeidung, sich mit Pestiziden zu vergiften. – Bio-Lebensmittel genießen einen guten Ruf, weil sie wesentlich weniger Schadstoffe enthalten als konventionell hergestellte Lebensmittel. – Bioprodukte sind kaum gesünder als konventionelle Lebensmittel
Food products & taste	It tastes totally different from your normal vegetables. – It can also be mixed in with the other foods (milk and fruit, oats/rice cooked in milk). – The increased flavor is the result of the food containing more micronutrients.
	Die meisten frischen Zutaten müssen etwas aufbereitet werden, damit sie gegessen werden können. – Es braucht wirklich nur Mehl, Wasser und etwas Salz, keinerlei andere Inhaltsstoffe, Punkt. – Nichts geht über frisch zubereitete Speisen aus gesunden Zutaten.
Food safety	Dioxins are extremely toxic chemicals, and their bioaccumulation in the food chain may potentially lead to dangerous levels of exposure. – Many of the toxins found in non-organic foods are toxins that have a cumulative effect on our bodies. – As proved by various researchers, these chemicals have harmful effects not only on the consumers but also on the environment and farmers.
	Tatsache ist die Dosen um die es bei Nahrungsmittelkontaminationen durch Dioxine geht sind derart gering dass sie auf keinem Wege zu eine signifikanten Gesundheitsgefahr führen. – Dioxine sind UBIQUITÄR und entstehen in nicht unerheblichen Mengen durch natürliche Vorgänge. – Es bedarf erheblicher Dosen um gefährliche Effekte von Dioxin und Lindan nachzweisen.
Research	Science is not always applied in benign ways, even when we know as much - growth hormones and indiscriminate use of antibiotics in livestock, for example. – The bottom line is that genetically modified organisms have not been proven in any way to be safe, and most of the studies are actually leaning the other direction, which is why many of the world's countries have banned genetically engineered foods. – However, evolution and adaptation, especially to unknown and unnatural substances, takes many generations for humans to achieve.
	Zu wenig verstehen die Wissenschaftler noch von ökologischen und evolutionären Prozessen. – Hinzu kommt dass die Natur ständig neue genetisch veränderte Organismen hervorbringt. – Es geht nicht um die Behauptung von Gentechnik sondern um die Behauptung ihrer Resultate.
Health & nutrition	While the government wants us to eat healthy, it is very true that organic foods are outrageously priced for the small amount of food we recieve. – The health issue with foods lies in our collective wisdom that insists on making foods as cheap as possible. – Of course there are health benefits to eating "organic" food.
	Für die Bevölkerungsschichten die auf günstige Lebensmittel angewiesen sind es komplizierter sich gesund zu ernähren. – Im Vordergrund unserer Lebensmittelwirtschaft steht eben der Profit und nicht die gesunde Ernährung. – Es ist sowieso viel gesünder Lebensmittel zu essen, die einen möglichst geringen Verarbeitungsgrad aufweisen.
Politics & policy & compliance	There are more that "government regulators" involved. – We full well know that the industry in all of its glory takes precedence over the concerns or welfare of the people of this country. – This is all being decided in PRIVATE, There is no involvement by the political or judicial processes that normally make laws in this country.
	Lobbyismus müsste als Straftatbestand angesehen werden und ähnlich schwerwiegend behandelt werden wie Landesverrat. – Das ist das Ergebnis der Lobbyarbeit und unsere Volksvertreter verabschieden solche Strafrahmen nicht versehentlich, sondern ganz bewußt. – Das und ähnliches, ändert nichts an der kriminellen Energie der durch die Politik und Gesetzgebung Vorschub geleistet wird.
Animal welfare & meat consumption	Organically raised animals used for meat must be given organic feed and be free of steroids, growth hormone and antibiotics. – When it comes to meat, again organic is the better option as animals are often treated cruelly and inhumanely to increase production. – Manure produced by organically raised animals wreaks less havoc on the environment, but the meat may still wreak havoc on arteries.
	Diejenigen die noch Fleisch essen nehmen Bio weil diese Tiere etwas weniger gequält werden als wie im konventionellen Bereich. – Im Falle von Fleisch geht das nicht anders, als dass man Tiere quält und dazu mit Dingen füttert, die man kaum noch als Futter bezeichnen kann. – Rinder fressen natürlicherweise kein Zucht-Getreide wegen des hohen Stärke und Fettgehalts.
Farming & agricultural policy & food security	Regardless of capacity the modern crops still need more and more land to feed the more and more people, even if the inefficiencies and failures of corporate agriculture are overcome. – The main problem in organic farming is the availability of adequate organic sources of nutrients (crop residues, composts, manures) to supply crops with all the required nutrients and to maintain soil health. – Without this, the organic farming industry can't sustain economically as most of the Organic food that is produced is bought by the big packaged food companies.
	Daß der Dünger für Bio-Acker nämlich von Nutztieren hergestellt wird und mehr Bedarf daran auch mehr Nutztiere zur Folge hat, gehört nicht zu den Notwendigkeiten, mit denen die Bio-Lobby hausieren geht. – Wegen der Förderung von Biogasanlagen und dem dadurch entstandenen Bedarf etwa an Mais sei Ackerland inzwischen vielerorts so teuer, dass die Bio-Bauern nicht mehr konkurrieren könnten. – Hinzu kommt ein Trend, der immer mehr Landwirte zu Energiewirten werden lässt: Nachwachsende Rohstoffe sind gefragt wie nie.
Consumer prices & profit	Acquisitions such as this takes away consumers' prerogative on where to spend our hard-earned dollars. – The consumer gains nothing from this. – But final cost to consumer is based on supply and demand.
	Beim Verbraucher bleibt so gut wie kein Preisvorteil. – Das man für "bessere" Erzeugnisse mehr zahlen muss, liegt auf der Hand. – Nur müssen diese Billigartikel erst mal produziert werden, bevor der Verbraucher zugreifen kann.

5.2 Evaluation Metrics for Headline Generation Using Deep Pre-trained Embeddings

The publication on the consecutive pages is relevant to the examination. It was accepted after peer-review as full paper at the 2020 12th Language Resources and Evaluation Conference. Gerhard Johann Hagerer, the author of the present thesis, is one of three first authors of that paper. His author role, the reprinting permission, and his following contributions are acknowledged by all authors Hagerer, Moeed, An, and Groh [2020a]:

> *"Gerhard Johann Hagerer headed the research project. He developed the research idea, the concept, and parts of the methodology of the paper. Furthermore, he supervised the implementation process and reviewed the source code. Regarding the writing of the paper, he created the outline, directed the drafting, and wrote significant parts of the paper, i.e., he paraphrased, corrected, combined, and otherwise improved drafted material."*

Publication Summary

The following publication is part of the groundwork for Hagerer et al. [2021d] and Danner et al. [2022]. We aimed at evaluating the semantic coherence of the Universal Sentence Encoder (USE), which forms the basis of the Universal Sentence Encoder Cross-Lingual (XLING), on which in turn our multi-lingual studies are based. Based on the USE, we propose a new metric to evaluate abstractive summarization models. Since traditional metrics for abstractive summarization, such as, ROUGE, are based on word counting, they do not consider features like semantic similarity between words and texts. USE works differently in that regards, such that it improves correlation with human judgement when being used to score abstractive summarization capabilities. Among others, this is tested on product reviews, showing that the approach is feasible to summarize and compare these kinds of user-generated texts as well. These insights were important in the course of this dissertation to justify and develop the topic modeling used in Hagerer et al. [2021d] and Danner et al. [2022].

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1796–1802
Marseille, 11–16 May 2020

Evaluation Metrics for Headline Generation Using Deep Pre-Trained Embeddings

Abdul Moeed[†], Yang An[†], Gerhard Hagerer[†], Georg Groh
Research Group Social Computing, Department of Informatics
Technical University of Munich, Germany
{abd.moeed, yang.an, gerhard.hagerer}@tum.de, grohg@in.tum.de
[†] These are equally first authors and appear in random order.

Abstract

With the explosive growth in textual data, it is becoming increasingly important to summarize text automatically. Recently, generative language models have shown promise in abstractive text summarization tasks. Since these models rephrase text and thus use similar but different words as found in the summarized text, existing metrics such as ROUGE that use n-gram overlap may not be optimal. Therefore we evaluate two embedding-based evaluation metrics that are applicable to abstractive summarization: Fréchet embedding distance, which has been introduced recently, and angular embedding similarity, which is our proposed metric. To demonstrate the utility of both metrics, we analyze the headline generation capacity of two state-of-the-art language models: GPT-2 and ULMFiT. In particular AES shows close relation with human judgments in our experiments and has overall better correlations with them compared to ROUGE. To provide reproducibility, the source code plus human assessments of our experiments is available on GitHub[1].

Keywords: Evaluation Methodologies, Language Modelling, Natural Language Generation, Summarization, Textual Entailment and Paraphrasing, Statistical and Machine Learning Methods

1. Introduction

The recent development of generative language models (LMs) is leading to new capabilities regarding the quality of text generation (Radford et al., 2019). This also holds true for tasks such as abstractive summarization, which is related to the language model generating summaries de nouveau and paraphrasing the text in its own words (Moratanch and Chitrakala, 2016). This is of high importance considering the large and always increasing amount of available texts and their relevance for humans.

An advantage of abstractive summarization is its superior readability (Hsu et al., 2018) compared to extractive summarization where keywords from the text are extracted and rearranged (Lin and Hovy, 2003). This benefit can be used for generating realistic headlines (Takase et al., 2016; See et al., 2017). However, it remains a challenge to find a faithful evaluation metric. ROUGE (Lin, 2004) - a standard performance metric for extractive summarization - is not always ideal for abstractive summarization, since readability is not taken into account (Paulus et al., 2017). Instead, it only accounts for n-gram overlap which is a problem for use cases when summaries rephrase the respective content using different but similar words.

To address this problem, pre-trained semantic similarity embeddings such as InferSent (Conneau et al., 2017) have been used successfully to evaluate the quality of GAN-based text generation (Semeniuta et al., 2018). Therefore, the concept of the Fréchet distance (Heusel et al., 2017), which is a well-known procedure for computer vision, is successfully applied for text generation as well. Due to the novelty of the approach, it appears unclear *how this method relates to human judgment on the task of abstractive summarization*. Further, it stays unclear how the concept works with more recent pre-trained embeddings than InferSent, and based on which language models these research questions could be solved.

Since most recent pre-trained embedding models are trained on sentences, we perform headline generation as an instance for abstractive summarization in order to evaluate the general feasibility of the approach. More specifically, we generate headlines for user product reviews and news stories using OpenAI's GPT-2 (Radford et al., 2019) after comparing its performance to fastai's ULMFiT (Howard and Ruder, 2018). For a comprehensive analysis, GPT-2 is trained on four datasets: a sub-dataset of the Amazon Product Dataset (He and McAuley, 2016; McAuley et al., 2015), CNN/Daily Mail (Hermann et al., 2015), Newsroom (Grusky et al., 2018) and Gigaword (Napoles et al., 2012) [2].

In order to generate headlines, we fine-tune and condition the language models. Based on the Universal Sentence Encoder (USE) (Cer et al., 2018) we derive Fréchet distances and depict their relation with human judgments. Additionally, we show another measurement based on angular similarity (Cer et al., 2018) with similar properties as Fréchet distance for the evaluation of generated headlines.

2. Methodology

2.1. Language Models

To generate summaries, we use two autoregressive language models: GPT-2 and ULMFiT. BERT (Devlin et al., 2018) is another powerful language model, and has been previously modified and used for extractive summarization (Liu, 2019). However, owing to its bi-directional nature, BERT expectedly performs poorly with masked input for text generation, and is thus not further considered. Released in February 2019, OpenAI's GPT-2 achieved state-of-the-

[1] https://github.com/Abdul-Moeed/headline-gen-metrics

[2] Gigaword corpus is taken from `https://github.com/harvardnlp/sent-summary`

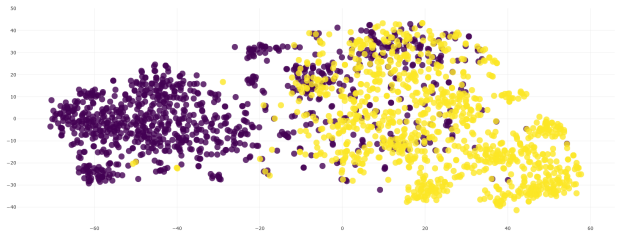


Figure 1: Dimension reduction of embeddings for Musical instruments (purple) and patio, lawn and garden (yellow) using PCA and t-SNE.

art performance on a variety of tasks in the zero-shot setting. Furthermore, it is trained in an unsupervised regime, with no domain-specific knowledge. The model is trained on OpenAI's custom "WebText" dataset. For our task, we fine-tune the smallest model available, with about 117 million parameters. Universal Language Model Fine-tuning for Text Classification (ULMFiT) was introduced by fastai in 2018, and is still used for many NLP tasks, including text generation. It uses cyclical learning rates (Smith, 2015) to converge faster compared to other models.

2.2. Automatic Evaluation Metrics

2.2.1. ROUGE

The current standard for evaluating summaries is ROUGE. Most authors report their ROUGE-1, ROUGE-2 and ROUGE-L scores as the only automatic evaluation score besides human evaluation (Liu, 2019; Nallapati et al., 2016; Nallapati et al., 2017; Paulus et al., 2017). Because they only compare n-gram overlap, ROUGE scores are agnostic of semantic similarity between reference and hypothesis summaries. This property is magnified in abstractive summarization, as the models try to paraphrase the original content. Thus they might use different wording compared to a reference text whereas the semantic meaning stays identical. The present paper addresses the issue by using embedding-based metrics. BLEU (Papineni et al., 2002) – a standard metric for machine translation – is also sometimes used for text summarization (Graham, 2015) though it is far from being as ubiquitous as ROUGE, and is known to not perform any better (Graham, 2015).

2.2.2. Proposed Embedding-Based Metrics

In this paper we use metrics based on sentence embeddings. The intuitive idea is to embed the generated and reference headline and the corresponding article text, then compare the semantic similarity between each of them.

Embedding Model As features for semantic similarity of sentences, we use the pre-trained embedding module Universal Sentence Encoder (USE) by Google (Cer et al., 2018). The model takes text of variable length as input, i.e. word sequences, sentences or small paragraphs, and encodes it as a 512 dimensional vector which can be used for text classification, semantic analysis, or other natural language processing tasks. For this study, we use the USE version based on the deep averaging network architecture (DAN) (Iyyer et al., 2015), which is available for download[3]. Each word in the sentence is first mapped to a word2vec (Mikolov et al., 2013) embedding before the USE averages them. This vector representation is then pushed through a feed forward neural network which produces a normalized sentence embedding. The DAN was pre-trained by the authors on English data from Wikipedia articles, question answering web pages, web news and discussion forums. The corresponding training tasks included conversation response prediction, quick thought, natural language inference (NLI) and translation ranking. The model is thus pre-trained to resemble semantically meaningful feature vectors which are suitable for a wide range of tasks. Consequentially, we utilize the pre-trained model as it is provided to calculate feature vectors for headlines and respective stories.

Review: My Weber grill cover couldn't make it 2 years in the Chicago winter and summer. [...] It has survived a few normal rain/thunderstorms without blowing off and protected the grill.I'll update my review in the future.[Update April, 2014]The cover survived a harsh cold winter, and more importantly my grill survived.
Reference: A replacement for a weber genesis
Generated 1: My grill survived the Midwest's toughest winter.
Generated 2: Still works today because of superior quality and durability.
Generated 3: Nice cover but nothing survives frost.

Table 1: Example headlines generated by GPT-2 on the 'patio, lawn and garden' subset.

Embeddings from semantically similar content ideally lie close to each other. We validate this by first applying PCA to reduce the dimensionality of the embedding space and then feeding the result to t-SNE to project the vector onto two dimensions for visualization (Figure 1). An interactive version can be found in our repository. Generated headlines from the 'musical instruments' sub-dataset (purple) on the one hand and 'patio, lawn, and garden' sub-dataset (yellow) on the other hand are separated adequately. Digging further, we observe that headlines from the same product lie close together, e.g. headlines of effect pedal reviews lie in the upper left corner. Moreover, non-informative headlines, e.g. "great product", "cheaply made", are centered between both clusters, as they do not contain identifying information about the product being reviewed. We conjecture that omitting the centered headlines during training would produce more reasonable headlines.

Angular Embedding Similarity Cer et al. (2018), the authors of USE, propose angular similarity to compare the semantics of two embeddings as

$$\mathrm{sim}(\mathbf{u}, \mathbf{v}) = 1 - \arccos\left(\frac{\mathbf{u} \cdot \mathbf{v}}{\|u\| \, \|v\|}\right) \cdot \frac{1}{\pi} \qquad (1)$$

which is a modification of cosine similarity to perform better on small angles.

To our knowledge, USE with angular similarity between generated and reference samples has never been used as an

[3] https://tfhub.dev/google/universal-sentence-encoder/2

Language Model	AES	Human
ULMFiT	0.575	17.7%
GPT-2	0.623	53.3%

Table 2: Average similarity of headlines generated by our fine-tuned language models compared to the corresponding reviews from Amazon. AES is our proposed metric defined in formula 2. Human is the similarity as perceived by two human annotators. The higher the values, the better is the headline generation capability of the language model.

evaluation metric for headlines or other kinds of abstractive summaries before. Therefore we propose the *angular embedding similarity* (AES) as the average of all angular similarities between the USE embeddings of two related text samples. For instance, when comparing all stories $\in$ R and their corresponding headlines $\in$ H of a corpus, the AES between them is defined as

$$\text{AES}_{S,H} = \frac{1}{n} \cdot \sum_{i=1}^{n} \text{sim}(\hat{s}_i, \hat{h}_i), \tag{2}$$

with $\hat{s}_i$ and $\hat{h}_i$ being the USE embedding of s_i and h_i, i.e., the ith story and corresponding headline, and n the total number of stories in that corpus.
As described in the experiments section, we evaluate AES between reference headlines and stories, generated headlines and stories, and generated headlines and reference headlines.

Fréchet Embedding Distance Fréchet distance (Fréchet, 1957) is a measurement to compare two Gaussian distributions

$$\text{FID}(r, g) = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{0.5}\right) \tag{3}$$

where r refers to the reference sample distribution and g to the generated sample distribution, μ and Σ to their corresponding means and covariance matrices.
The Fréchet distance has already been used successfully in computer vision to evaluate generative models (Heusel et al., 2017). Recently, it has been used in natural language generation as an alternative to ROUGE (Semeniuta et al., 2018). There, InferSent is utilized (Conneau et al., 2017) to compare the output of GANs for language generation. d'Autume et al. (2019) later calculated the Fréchet distance on USE embeddings and called it the Fréchet embedding distance (FED). The authors also noticed a drawback of FED: its sensitivity to length. We also confirm this observation experimentally in section 4. In contrast to AES, FED is a distance and lower scores mean higher similarity. Also worth noting is the fact that FED requires multiple samples for computing the distance between distributions, while AES can compute similarity for pairs of samples.

3. Experiments

3.1. Datasets

Amazon Product Reviews The Amazon product review dataset (He and McAuley, 2016; McAuley et al., 2015) is a collection of domain-specific sub-datasets. Each sub-dataset is of varying size, and contains user product reviews from Amazon.com. The `summary` attribute is used as reference headline (Ma et al., 2018). We train the language models on the 'patio, lawn and garden' dataset.

CNN/Daily Mail The CNN/Daily Mail dataset (Hermann et al., 2015) is composed of news stories collected from *cnn.com* and *dailymail.co.uk*. While explicit headlines are not provided, each story has multiple 'highlights' – key takeaways from the story. For our experiments, we use the first highlight as the ground-truth/reference headline for that story. The dataset, though originally created for the question/answering task, was adapted for summarization by Nallapati et al. (2016).

Gigaword Another standard dataset used in text summarization is the annotated Gigaword coprus (Napoles et al., 2012). The dataset contains 10 million articles, each having a corresponding headline. The headline has been previously used as a summary (Rush et al., 2015). The length of each story in Gigaword is much shorter compared to other datasets used in our experiments.

Newsroom Newsroom (Grusky et al., 2018) is a recently released news-centric dataset specifically aimed at text summarization tasks. It is composed of English news-related articles produced by 38 notable publications. The authors claim that the dataset captures a variety of human summarization styles, making it amenable to abstractive, extractive and mixed summarization strategies.
For our experiments, we take approximately 90,000 story/headline pairs from CNN/DailyMail, Newsroom and Gigaword each. These are then split into train/test sets. As Amazon's 'patio, lawn and garden' is much smaller than the rest, we use the whole dataset for our experiments. Each dataset is split into 90% training data and 10% test data, the latter of which is used to generate headlines.

3.2. Training

For training, we follow a modified version of the approach introduced by Radford et al. (2019). The authors evaluate the quality of summarization using GPT-2 without further fine-tuning on CNN/Daily Mail. We improve the quality of generated headlines, by fine-tuning GPT-2 and ULMFiT on the review/story text and reference headlines. The training data has the following format:

`Review/Story Text + [TL;DR:] + Headline + [End]`

The model learns how a reference headline should look like given the full review. The `[TL;DR:]` token signals to the model the end of the review and start of the headline. During headline generation, the model is then given the input:

`Review/Story Text + [TL;DR:]`

Both ULMFiT and GPT-2 are conditioned and fine-tuned for Amazon reviews, though only GPT-2 is subsequently also trained for CNN/Daily Mail, Gigaword, and Newsroom (discussed further in section 4.1.). The datasets are trained using the same scheme as above. For each dataset, GPT-2 is trained for 5000 counters with *learning rate = 1e-4*, and headlines are generated with *top-k = 40* and *temperature = 1.0*. The randomness of text generation can be

Dataset	Metric	Sample size	Gen-Story		Ref-Story		Gen-Ref	
			r	p-value	r	p-value	r	p-value
Amazon	ROUGE-1	20	0.1199	0.6148	0.2100	0.3743	0.602	0.00498
	ROUGE-2	20	-0.0746	0.75446	0.2417	0.30459	0.5423	0.0135
	ROUGE-L	20	0.0843	0.72379	0.1959	0.40773	0.575	0.008
	AES	20	-0.0330	0.89112	0.1260	0.59791	**0.6540**	**0.00176**
CNN/Daily Mail	ROUGE-1	20	0.3785	0.09985	-0.1565	0.51001	0.4875	0.02923
	ROUGE-2	20	0.5019	0.02414	-0.3061	0.18933	0.4681	0.0374
	ROUGE-L	20	0.4346	0.05549	-0.1802	0.44702	0.4436	0.05007
	AES	20	0.2430	0.30174	0.0000	0.99916	0.2550	0.27735
Newsroom	ROUGE-1	20	-0.0465	0.84576	0.5390	0.01419	**0.7706**	**7e-05**
	ROUGE-2	20	-0.1231	0.60523	0.2858	0.22184	0.6027	0.00491
	ROUGE-L	20	0.0563	0.81366	0.4891	0.02865	**0.7560**	**0.00012**
	AES	20	0.5010	0.02449	0.4110	0.07149	**0.8240**	**1e-05**
Gigaword	ROUGE-1	20	0.5450	0.01295	0.4636	0.0395	**0.6584**	**0.0016**
	ROUGE-2	20	0.5522	0.01158	0.1640	0.48955	0.3722	0.10613
	ROUGE-L	20	0.5056	0.02296	0.4976	0.02559	0.5985	0.00531
	AES	20	0.4070	0.07509	**0.7260**	**0.00029**	0.4480	0.04759
Overall	ROUGE-1	80	0.2829	0.01099	0.2308	0.03941	**0.6749**	**6.6e-12**
	ROUGE-2	80	0.3023	0.00643	0.0797	0.48196	**0.5128**	**1.1e-06**
	ROUGE-L	80	0.2878	0.00963	0.2570	0.02136	**0.6542**	**4.69e-11**
	AES	80	0.2290	0.04128	0.2820	0.01127	**0.5810**	**2e-08**

Table 3: Summary of Pearson correlation of human judgment with automatic metrics. Bold *r* and *p-value* pairs indicate Bonferroni-adjusted statistically significant results (*p-value* less than 0.0042)

adjusted by the latter two hyperparameters, and we observe that using the aforementioned values for each strikes a sufficient balance between relevant and creative summaries in our case.

Pre-processing steps on the dataset include filtering out useless phrases, such as time, place, or author of a story and clipping each story to a maximum of five sentences. In the case of Amazon reviews, shorter headlines (less than 15 characters) are filtered out. This proved to be useful for generating more meaningful headlines as the shorter headlines are generic and not informative of the product.

3.3. Human Evaluation

We perform three manual evaluation tasks in order to compare the proposed automatic evaluation metrics to human judgments. For the first task, the person is asked to read a story text and decide if the generated headline is a reasonable summary of the respective story. The sentiment and content of a story and a correspondingly generated headline are supposed to be related, and the headline text should be understandable and not artificial. Similarly, the second task asks the person to assess the same, only this time for the reference headline instead of the generated one. The third and final task is asking the person to judge the similarity between the reference and generated headlines. The tasks are termed as 'Gen-Story', 'Ref-Story' and 'Gen-Ref' respectively. The testers are kept ignorant of whether a given headline is real or generated. Five testers are asked to score the respective similarities on a scale from 1-5, where 5 means very similar and 1 hardly similar.

All three tasks are performed for each of the four datasets, with 20 samples taken from each. This gives each tester a total of 80 samples with three tasks.

4. Results

4.1. Comparison of Language Models

We initially test which language model is more suitable to generate relevant headlines. This is done by fine-tuning both GPT-2 and ULMFiT on the 'patio, lawn and garden' dataset, generating headlines for said dataset and evaluating them using AES and human judgment. The results can be seen in table 2. GPT-2 clearly outperforms ULMFiT which is consistent with human assessment. Henceforth, we only use GPT-2 for our subsequent experiments to test the validity of AES and FED as automatic metrics.

Table 1 shows a good, average, and bad example of the capabilities of GPT-2. Note that the model still remembers that Chicago lies in the Midwest of the US from its pre-training on WebText. More examples can be found in our repository.

4.2. Metric Evaluation

In this section, we report whether AES and FED relate with human judgment in a significant manner. We also compare AES with the standard metric ROUGE.

AES We use Pearson correlation as the measure to gauge the correlation between two variables, and perform null hypothesis testing using p-values. As AES can be calculated on a per-sample basis, we have AES scores for 20 samples per dataset (80 in total) for each of the 3 tasks. The tasks are listed as 'Gen-Story', 'Ref-Story' and 'Gen-Ref' and their description can be found in section 3.3. This gives us a total of 240 AES scores (80 per task). As we have 5 human evaluators, we calculate the human average for each

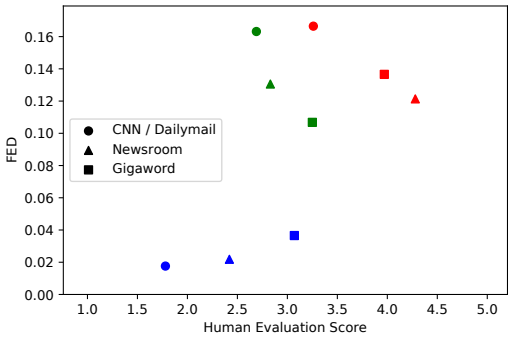


Figure 2: Scatter plot of FED vs human evaluation. The three tasks are color-coded: Green denotes 'Gen-Story', orange denotes 'Ref-Story' and blue denotes 'Gen-Ref.'

task per sample to match the count of 240 AES scores. Finally, the correlation between the AES scores and average human judgment is calculated for each task as well as the significance of the correlation via p-values.

Table 3 shows the results of the metric evaluation per dataset, as well as an overall assessment. ROUGE values are also listed for the sake of comparison. As can be observed, AES correlates positively with human judgment in all but two cases. Many of the positive correlations are also statistically significant.

Table 4 shows how many times a metric was the highest correlated one with human judgment. Note that this is done by looking at each task in table 3 for each dataset and noting which metric has the highest *r* value. Although per convention a p-value below 0.05 is considered statistically significant, we perform the Bonferroni adjustment (Weisstein, 2004) which corrects for the number of experiments done with a dataset. In our case, 12 experiments are done on each dataset. Dividing 0.05 by 12 yields a Bonferroni-corrected statistical significance threshold of 0.0042. The results are reported in table 4. A metric may have a statistically significant correlation with human evaluation while never having the highest *r* value, as in the case of ROUGE-L. The table provides a clear picture of AES when compared to ROUGE; AES is highest correlated with human perception more frequently than any ROUGE metric individually, and the correlations are statistically significant more often than any ROUGE metric.

FED In contrast to AES, FED can only be calculated on a per-corpus basis, rather than a per-sample basis as stated in section 2.2.2. The efficacy of calculating *r* between human judgment and FED is thus diminished due to low sample size (we have 12 FED values in total as a result of all experiments). However, the relation between human perception and FED can still be demonstrated, albeit in a less robust manner compared to AES, by plotting the average human scores against FED values. This can be seen in figure 2. Human comparison with Amazon's 'patio, lawn and garden' is omitted from the figure as the values were considered outliers (more than 10x those of other datasets). We hypothesize that this is due to the vast number of uninformative headlines in that corpus.

A clear trend in the first two tasks (colored green and orange) is that an increase in human scores results in lower

Metric	# best correlations	Bonferroni corrected
ROUGE-1	4	3
ROUGE-2	5	1
ROUGE-L	0	2
AES	6	4

Table 4: Table showing how many times each metric was the highest correlated one with human judgment. Additionally, we can see how many of the correlations were statistically significant after applying the Bonferroni correction.

FED, as hypothesized. The last task shows no strong relation.CNN/Daily Mail have high FED values in task 2 and task 3 even though they follow the expected trend of negative relation with human judgment. We speculate that this is due to the fact that the dataset does not contain headlines for each story, rather containing multiple 'highlights' which emphasize key points of the story. As such, any single highlight may be unable to capture the crux of the story. FED's sensitivity to sentence-length is also demonstrable as the values for task 3 (blue) are visibly smaller than the other two tasks that involve the story/review text.

5. Conclusion

In this paper, we fine-tune and condition language models to generate abstractive summaries in the form of headlines. Qualitatively, many generated headlines appear to be valid for the given text. To further evaluate the headlines, we rely on the recently published FED (Semeniuta et al., 2018; Conneau et al., 2017) as well as on AES, which is our proposed metric. Experimentally, we show that AES corresponds to human perception and performs mostly better than the traditional ROUGE metric, whereas FED does not always relate to human perception not least due to its sensitivity to text length.

All evaluated metrics for abstractive summarization are merely based on the pre-trained Universal Sentence Encoder (Cer et al., 2018). However, recently many other textual embeddings have been published which might be a better choice with respect to computational efficiency, e.g. smooth inverse frequency (Arora et al., 2017), accuracy, e.g. BERT (Devlin et al., 2018), or stability with respect to length of text, e.g. doc2vec (Le and Mikolov, 2014). In particular the latter could potentially lead to the development of metrics which would be more suitable for abstractive summarization tasks other than headline generation.

6. Bibliographical References

Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. *ICLR*.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

d'Autume, C. d. M., Rosca, M., Rae, J., and Mohamed, S. (2019). Training language gans from scratch. *arXiv preprint arXiv:1905.09922*.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Fréchet, M. (1957). Sur la distance de deux lois de probabilité. *COMPTES RENDUS HEBDOMADAIRES DES SEANCES DE L ACADEMIE DES SCIENCES*, 244(6):689–692.

Graham, Y. (2015). Re-evaluating automatic summarization with bleu and 192 shades of rouge. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 128–137.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Hsu, W.-T., Lin, C.-K., Lee, M.-Y., Min, K., Tang, J., and Sun, M. (2018). A unified model for extractive and abstractive summarization using inconsistency loss. *arXiv preprint arXiv:1805.06266*.

Iyyer, M., Manjunatha, V., Boyd-Graber, J. L., and Daumé, H. (2015). Deep unordered composition rivals syntactic methods for text classification. In *ACL*.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Lin, C.-Y. and Hovy, E. (2003). The potential and limitations of automatic sentence extraction for summarization. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5*, pages 73–80. Association for Computational Linguistics.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Liu, Y. (2019). Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moratanch, N. and Chitrakala, S. (2016). A survey on abstractive text summarization. In *2016 International Conference on Circuit, power and computing technologies (ICCPCT)*, pages 1–7. IEEE.

Nallapati, R., Zhou, B., dos Santos, C., Gulcehre, C., and Xiang, B. (2016). 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.

Nallapati, R., Zhai, F., and Zhou, B. (2017). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Paulus, R., Xiong, C., and Socher, R. (2017). A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. *CoRR*, abs/1509.00685.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Semeniuta, S., Severyn, A., and Gelly, S. (2018). On accurate evaluation of gans for language generation. *arXiv preprint arXiv:1806.04936*.

Smith, L. N. (2015). No more pesky learning rate guessing games. *CoRR*, abs/1506.01186.

Takase, S., Suzuki, J., Okazaki, N., Hirao, T., and Nagata, M. (2016). Neural headline generation on abstract meaning representation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1054–1059.

Weisstein, E. W. (2004). Bonferroni correction.

7. Language Resource References

Grusky, M., Naaman, M., and Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

He, R. and McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 507–517, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Ma, S., Sun, X., Lin, J., and Ren, X. (2018). A hierarchical end-to-end model for jointly improving text summarization and sentiment classification. *arXiv preprint arXiv:1805.01089*.

McAuley, J., Targett, C., Shi, Q., and van den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 43–52, New York, NY, USA. ACM.

Napoles, C., Gormley, M., and Van Durme, B. (2012). Annotated gigaword. In *Proceedings of the Joint Workshop*

on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, pages 95–100. Association for Computational Linguistics.

5.3 An Evaluation of Progressive Neural Networks for Transfer Learning in Natural Language Processing

The publication on the consecutive pages is relevant to the examination. It was accepted after peer-review as full paper at the 2020 12th Language Resources and Evaluation Conference. Gerhard Johann Hagerer, the author of the present thesis, is one of five first authors of that paper. His author role, the reprinting permission, and his following contributions are acknowledged by all authors Hagerer, Moeed, Dugar, Gupta, Ghosh, Danner, Mitevski, Nawroth, and Groh [2020b]:

> *"Gerhard Johann Hagerer headed the research project. He developed the research idea, the concept, and the methodology of the paper. He managed the data annotation process of one of the datasets. Furthermore, he directed the implementation process and reviewed the source code deeply. Regarding the writing of the paper, he created the outline, directed the drafting, and wrote most of the paper, i.e., he wrote large textual parts, incorporated extensive reviewer feedback, and paraphrased, corrected, combined, and otherwise improved drafted material."*

Publication Summary

The following paper is leveraging transfer learning in order to improve classification of our crowdsourced aspect-based sentiment analysis (ABSA) annotations on our organic food social media dataset. The technique of progressive neural networks is used to overcome the effect of catastrophic forgetting in neural transfer learning on several natural language processing (NLP) tasks, including, ABSA, sentiment analysis, and named entity recognition. The proposed method gives significant improvements throughout all experiments. However, we find that the prediction accuracy on our organic dataset with crowdsourced annotations is still too low to be used successfully for Directed Content Analysis studies. We conclude that overcoming catastrophic forgetting is not enough, and that annotation noise should be investigated, too.

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1376–1381
Marseille, 11–16 May 2020

An Evaluation of Progressive Neural Networks for Transfer Learning in Natural Language Processing

Gerhard Hagerer[1†], Abdul Moeed[1†], Sumit Dugar[1†], Sarthak Gupta[1†], Mainak Ghosh[1†]
Hannah Danner[1] Oliver Mitevski[2], Andreas Nawroth[2], Georg Groh[1]
[1] Technical University of Munich, [2] Munich Re
[†] These authors contributed equally.
gerhard.hagerer@tum.de

Abstract

A major challenge in modern neural networks is the utilization of previous knowledge for new tasks in an effective manner, otherwise known as transfer learning. Fine-tuning, the most widely used method for achieving this, suffers from catastrophic forgetting. The problem is often exacerbated in natural language processing (NLP). In this work, we assess *progressive neural networks* (PNNs) as an alternative to fine-tuning. The evaluation is based on common NLP tasks such as sequence labeling and text classification. By gauging PNNs across a range of architectures, datasets, and tasks, we observe improvements over the baselines throughout all experiments.

Keywords: Document Classification, Text categorisation, Named Entity Recognition, Opinion Mining / Sentiment Analysis, Statistical and Machine Learning Methods, Other (Transfer Learning)

1. Introduction

Transfer learning is the ability of a model to generalize over previously unseen domains and/or tasks in a competent manner. The intuition is to re-use previously learned knowledge effectively when learning new tasks. The most common approaches to transfer learning include fine-tuning and multi-task learning (MTL). The former, where the weights of the already pre-trained layers are re-trained for a new task, performs well for similar tasks (Min et al., 2017) but fails to transfer over unrelated tasks (Mou et al., 2016). The latter adds terms to the objective function for each new task (Rei, 2017), which means re-training the whole model from scratch each time a new task is added (Chen et al., 2017).

A major problem faced by traditional transfer learning approaches is *catastrophic forgetting* (French, 1999) – a phenomenon where the model loses performance on previously learned tasks when trained on a new task. Catastrophic forgetting is thoroughly documented in artificial neural network literature and a few solutions have been proposed (Kirkpatrick et al., 2016; Awasthi and Sarawagi, 2019). The problem is more prevalent in NLP compared to computer vision; the shallow nature of networks used for NLP has been cited as a possible explanation for this discrepancy (Howard and Ruder, 2018).

Transfer learning approaches other than fine-tuning and MTL have also been explored for neural architectures (Hodas et al., 2017; Riemer et al., 2017). One such example - progressive neural networks (PNNs) (Rusu et al., 2016) - offers a novel solution to catastrophic forgetting. The idea is to train multiple networks - one for each new domain/task - that share information learned from previous tasks with each other through lateral connections. PNNs have gained popularity and have already been used for transfer learning in video summarization (Choi et al., 2018) and emotion recognition (Gideon et al., 2017).

The aim of this work is to evaluate PNNs for transfer learning in the context of various NLP tasks related to sequence labeling and text classification (Aggarwal and Zhai, 2012). Specifically, three tasks are targeted: named entity recognition (NER), sentiment analysis (SA) and aspect-based sentiment analysis (ABSA). We perform a cross-task, cross-architecture comparison of PNNs with traditional transfer learning methods. The architectures differ significantly between these tasks; NER uses bi-directional LSTMs (BiLSTMs) with no convolutional or attention layers, SA uses convolutional layers and ABSA uses BiLSTMs with attention.

2. Background

2.1. Sequence labeling

NER falls under sequence labeling - the task of labeling each word in a sentence as a category (part of speech, entity etc.) Classically, solutions to NER have taken the form of Hidden Markov Models (Luo et al., 2015) and Conditional Random Fields (Lafferty et al., 2001; Passos et al., 2014). More recently, end-to-end neural techniques have been deployed that do not require task-specific knowledge (Ma and Hovy, 2016).

2.2. Text classification

Sentiment analysis is the task of classifying text according to the sentiment it exhibits. The sentiment labels are usually *positive*, *negative* and *neutral*. While sentiment analysis is well-studied in the NLP literature (Dave et al., 2003; Mäntylä et al., 2016) deep networks have nonetheless proved beneficial lately (Dos Santos and Gatti, 2014; Kim, 2014).

Closely related to sentiment analysis is aspect-based sentiment analysis (ABSA), which is more fine-grained. Here, the task is to find relevant aspects (e.g. product) in the text and detect their corresponding sentiments. Traditionally, aspect extraction has been treated as a secondary task for ABSA and the focus has been to classify the sentiment polarities of the aspects
(Schouten and Frasincar, 2015; Lakkaraju et al., 2014). Recent work, including that of this paper, departs from

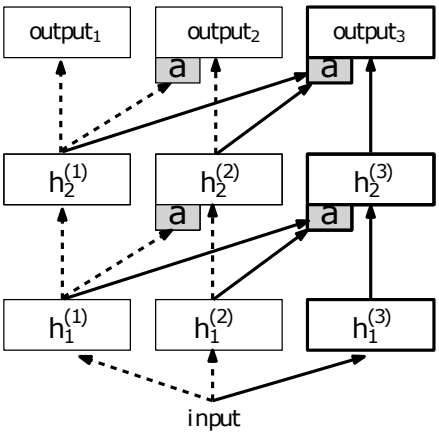


Figure 1: A three column progressive network. The first two columns each are trained on different tasks. The grey box represents the adapter layers. The third column is trained on the target task. Taken from (Rusu et al., 2016).

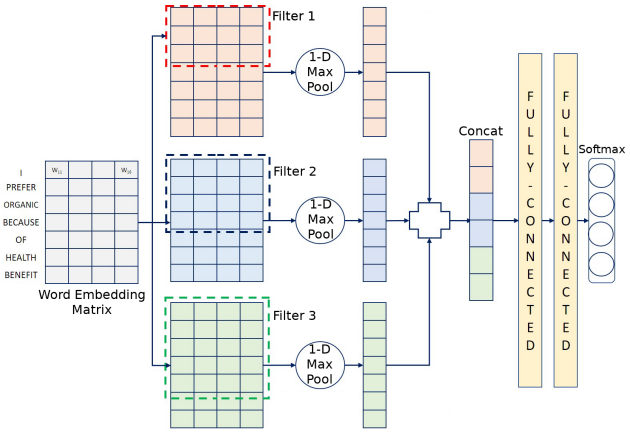


Figure 2: Model architecture for sentiment analysis (Gupta, 2019). The convolutional layer from the source column is passed to the first fully-connected layer (FCL) of the target column. The first FCL is then passed to the second FCL.

this formulation and treats aspect extraction as part of the ABSA task (Wojatzki et al., 2017; Schmitt et al., 2018).

2.3. Progressive Neural Networks

(Rusu et al., 2016) proposed progressive neural networks (PNNs) as a transfer learning technique for both cross-domain and multi-task purposes – see also (Gupta, 2019). The authors showed the effectiveness of PNNs on reinforcement learning tasks, with the technique demonstrating superior performance to pre-training and fine-tuning. The technique consists of adding lateral connections - coming from networks trained for source tasks - to the network being trained for the target task. Only the parameters of the target network are learned while the source weights are frozen. This ensures the immunity of PNNs to catastrophic forgetting.

The first part of a PNN is a neural network which is trained on the source task containing L hidden layers. This is called the first column with activations denoted as h_i^1 of layer i. After the training of the first column is finished, a second so called target column after being initialized randomly is trained on the target task. The activations h_i^2 of the second column are calculated based on the activations from the previous layer of the same column h_{i-1}^2 and from the previous layer of the source column h_{i-1}^1. Therefore lateral connections between the layers of the source and target column are created. These connections are trained, too, whereas *the weights of the source column are not updated.*

The generalized mathematical formulation for multiple columns is

$$h_i^k = \sigma(U_i^{k:j}\sigma(V_i^{k:j}\alpha_i^{<k}h_{i-1}^{<k}) + W_i^k h_{i-1}^k) \qquad (1)$$

where σ is the activation function, K is the number of columns, $U_i^{j:k}$ is the weight matrix representing the lateral connections from column j to k, and W_i^k is the weight matrix of the i^{th} layer in column k^{th}.

In place of connecting the previous column directly by multiplying its activations h_{i-1}^1 with $U_i^{k:j}$, a non-linear downprojection using matrix $V_i^{k:j}$ is added. This concept, termed *adapter*, enhances the lateral connections and reduces the model complexity. The learnable scalar parameter α_i scales the activations of the source task such that their order of magnitude fits to the target task.

3. Experiments

3.1. Transfer Learning

For each task mentioned in the paragraphs below, we evaluate transfer learning using progressive neural networks as they have been introduced above. As a baseline to show improvements using that technique, an appropriate neural network model is trained on that task. For each task there are at least two or more domains or sub-datasets given. Transfer learning is evaluated by training on a source domain and fine-tuning on the respective target domain of the same task. Therefore we apply normal fine-tuning of all layers (FT) and progressive neural networks with one (1PNN) and two (2PNN) source columns.

The latter is only done for the task of named entity recognition (NER), since, as shown in the results section, the increase in performance is small while the increase of the model complexity is big. For NER with 1PNN the best performing source column is chosen to be connected laterally to the target network.

For NER we further investigate the effect of catastrophic forgetting, i.e., the degradation of the prediction performance of a model which is firstly trained on the source task, then fine-tuned to a target task, and then again evaluated on the source task. Due to the fine-tuning on the target domain, a decrease of prediction accuracies is expected to happen due to the modification of the network weights. This is done to shed light upon what happens to the source networks during training on NLP tasks.

3.2. Named Entity Recognition

Modeling For all evaluations on named entity recognition (NER), micro F1 score is used as the metric. GloVe 100-D (Pennington et al., 2014) is used for word embeddings. We make a slight modification to (Ma and Hovy, 2016)'s architecture, having two LSTM layers instead of

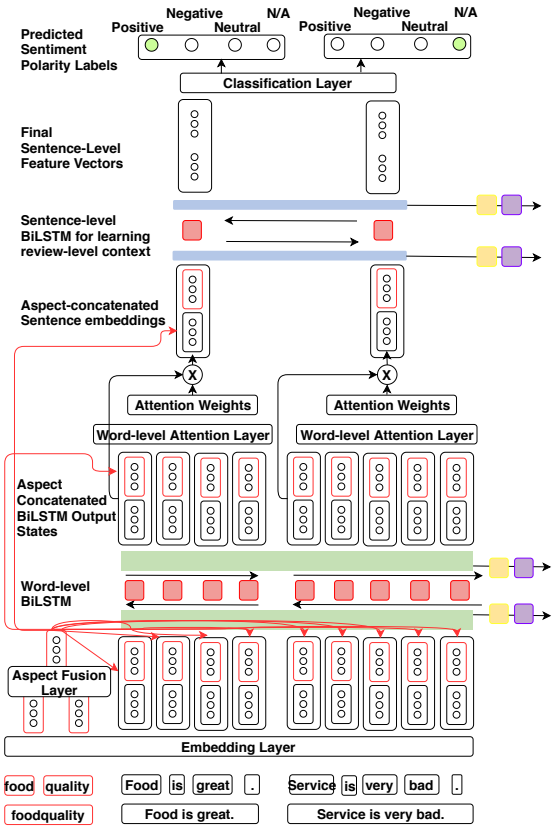


Figure 3: Model architecture for ABSA (Dugar, 2019). Green and blue rectangles denote lateral connections of BiLSTM layers on word and sentence representations. Red squares represent BiLSTM cells, yellow represent learnable scalar, purple represent the adapter layers.

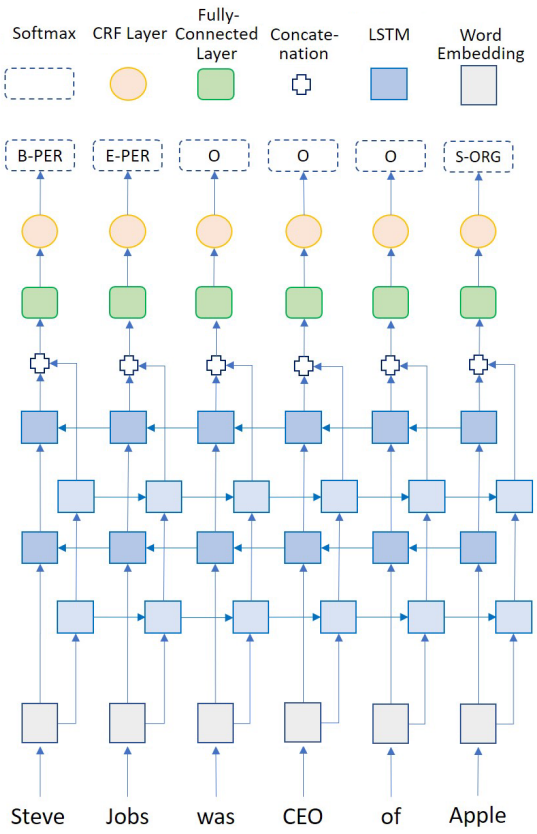


Figure 4: Model architecture for the NER experiments (Gupta, 2019). For the PNN, the first BiLSTM layer of the source column is passed to the second RNN layer of the target column, and the second RNN layer to the fully-connected layer analogously.

the original's one (Gupta, 2019). The architecture is illustrated in illustration 4.

Regarding the data, the experiments for named entity recognitions are executed on three different publicly available biomedical datasets as they are provided by (Crichton et al., 2017).

BC5CDR Dataset The BioCreative V Chemical-Disease Relation dataset (BC5CDR) is released along with the CDR task of the BioCreativ V challenge in 2015 (Li et al., 2016). The overall goal of the challenge is to find relations between chemicals and their associated diseases. Thus, the entity classes *chemical* and *disease* have been annotated manually which is the ground truth for NER in our experiments.

NCBI Dataset The NCBI Disease Corpus from 2014 aims at evaluating the task of disease name recognition. It comes with manual annotations for all mentioned diseases and according classifications based on 793 PubMed abstracts (Doğan et al., 2014). In our experiments, only the target entity *disease* is classified for each word.

JNLPBA Dataset This dataset was published for the JNLPBA challenge of bio-entity recognition in 2004. The data is based on the GENIA v3 named entity corpus of MEDLINE abstracts (Kim et al., 2004). The target classes in this dataset are *DNA*, *RNA*, *cell line*, *cell type*, and *protein*.

3.3. Sentiment Analysis

Modeling For all the evaluations of sentiment analysis, accuracy score is used as the evaluation metric. Similar to NER, the model is based on pre-trained GloVe 100-D embeddings. The model architecture is inspired by (Kim, 2014) which uses three one dimensional convolutional kernels with varying sizes in the first layer to capture local features as can be observed in figure 2.

Amazon Dataset As data the Amazon product review dataset as provided by (Blitzer et al., 2007) is used in the experiments. For transfer learning, the categories 'kitchen houseware' and 'personal healthcare' are considered. The annotated sentiment target classes are positive and negative.

3.4. Aspect-Based Sentiment Analysis

Modeling We also consider aspect-based sentiment analysis (ABSA) as a task for our experiments (Dugar, 2019). The utilized architecture is a hierarchical neural network as shown in figure 3 which is inspired by (Yang et al., 2016). It also uses GloVe word vectors BiLSTMs and attention (Wang et al., 2016), with a joint end-to-end formulation of ABSA similar to (Schmitt et al., 2018).

SemEval Dataset The dataset for ABSA is taken from the SemEval 2016 challenge task 5 subtask 1 (Pontiki et al., 2016). The subtask is defined as aspect extraction and sentiment polarity classification with regard to that aspect.

Category	%	Summary Label
	39%	neutral/ambiguous
sentiment	32%	positive
	29%	negative
	83%	organic
entity	11%	conventional
	5%	genetic engineering
	33%	general
	28%	healthiness
attribute	12%	trustworthiness
	11%	quality
	10%	environment
	6%	price

Table 1: Annotation distribution on the organic of all sentences to which at least one opinion triplet (entity+attribute+sentiment) was assigned, i.e., 53% of all 10,000 sentences. 668 of the annotated sentences contain two or more opinion triplets.

Task-Dataset	Train	Val	Test
NER-JNLPBA	16691	1853	3856
NER-BC5DR	5423	922	939
NER-NCBI	4559	4580	4796
SA-Amazon	2880	320	800
ABSA-SemEval-R	5654	106	106
ABSA-SemEval-L	55136	6892	6892
ABSA-Organic	8824	712	908

Table 2: Data splits of the named entity recognition (NER), sentiment analysis (SA) and aspect-based sentiment analysis (ABSA) tasks. SemEval-R and SemEval-L refer to the restaurant and laptop datasets respectively.

For our experiments, we solve the subtask as a whole by jointly classifying the aspect and its related sentiment. For the evaluation of transfer learning, we utilize both given domains, i.e., *laptops* and *restaurants*.

Organic Dataset One important goal of transfer learning is to improve performance on custom datasets of the respective target domain of interest. In that regard it is not always possible to provide many expert annotations of high reliability on noisy real world data. In that regard aspect-based sentiment analysis can be considered as an interesting use case due to reasons such as high number of classes, multi-labeling classification, and small number of annotated samples.

Therefore, we collected $10,000$ social media comments from the well-known question-and-answer website Quora which contain opinions about *organic food* and related consumer issues. After being thoroughly instructed, each of 10 labelers annotated relevance, entity, attribute, and sentiment for 1000 sentences. Relevance is merely a binary flag to indicate if the sentence contains a relevant opinion. The other classes and their respective distributions are enlisted in table 1.

Transfer learning is evaluated by first training on the laptop and restaurant dataset jointly and then fine-tuning and evaluating on the organic dataset.

Tasks-Dataset	Baseline	FT	1PNN	2PNN
NER-JNLPBA	73.8	73.1	74.0	**74.3**
NER-BC5DR	81.7	83.7	**84.8**	84.0
NER-NCBI	84.9	85.0	85.6	**85.7**
SA-Kitchen	79.0	79.0	82.5	-
SA-Healthcare	80.5	81.1	**82.9**	-
ABSA-SemEval-R	39.4	32.4	**47.1**	-
ABSA-SemEval-L	24.6	22.9	**27.6**	-
ABSA-Organic	12.9	6.2	**17.0**	-

Table 3: Summary of the results across different tasks and datasets. NER and ABSA results are reported using micro F1 scores, while SA results use model accuracy. SemEval-R and SemEval-L correspond to the restaurant and laptop datasets of SemEval respectively. PNN outperforms both the baseline and fine-tuned models across all tasks and datasets. Bold entries indicate the best performing architecture for that row.

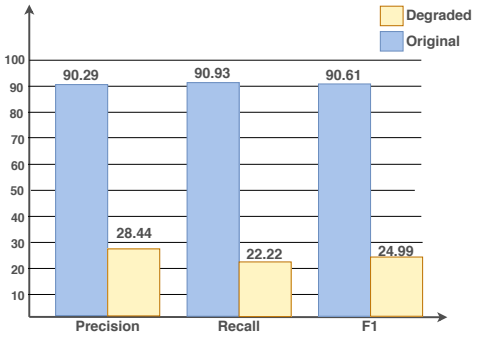


Figure 5: Graph showing catastrophic forgetting on the CoNLL03 domain after fine-tuning on JNLPBA.

4. Results

As a general overview, table 3 provides a summary of the results. It can clearly be seen that PNNs exceed the baseline and the standard fine-tuning approach for transfer learning throughout all tasks, domains, and architectures. This is denoted as bold number in table table 3.

4.1. Named Entity Recognition

For JNLPBA as the target domain, 2PNN fares marginally better than 1PNN. We train the source columns on the NCBI and BC5CDR datasets. Using NCBI as the target domain, 1PNN and 2PNN are comparable. Finally, 1PNN outpeforms all other transfer techniques for BC5CDR as target domain. Varying the source datasets between JNLPBA and NCBI does not change performance in any significant manner (Gupta, 2019).

4.2. Sentiment Analysis

Similar to the results of NER, PNNs outperform fine-tuning the model. Fine-tuning yields results not too dissimilar to the baseline (Gupta, 2019).

4.3. Aspect-based Sentiment Analysis

Results for ABSA are no different; PNN surpasses fine-tuning notably in terms of performance. The performance gain of PNN is varied, however, and depends on the domain. For the 'Restaurant' dataset as target, PNN achieves

a micro F1 score of 47.1% vs. 32.4% achieved by fine-tuning. The second experiment, however, does not show as remarkable a difference, with micro F1 scores of 27.6% vs 22.9% for PNN and fine-tuning respectively.
In our experiments, PNNs with two source columns are not evaluated for ABSA; we hypothesize, however, that similar to NER 2PNN performs at least as well as 1PNN in most cases (Dugar, 2019).

4.4. Catastrophic forgetting

As demonstrated in figure 5, we confirm the occurrence of catastrophic forgetting for NER. Initially, the model is trained on the CoNLL03 (Sang and De Meulder, 2003) dataset. After being subsequently fine-tuned on JNLPBA, the model's performance is crippled on the original domain. A performance degradation of approximately 70% can be observed (Gupta, 2019).

5. Conclusion

Transfer learning ensures a learning algorithm's ability to generalize over new domains and tasks in a competent manner. In this paper, we evaluate progressive neural networks as a transfer learning approach with reference to natural language processing tasks. We observe that progressive networks consistently outperform the conventional transfer technique of fine-tuning the network on named entity recognition, sentiment analysis, and aspect-based sentiment analysis. We further observe that PNNs with two source networks and according lateral connections produce marginally better results than with a single source network.

6. Acknowledgements

We are thanking Munich Re AG for their support of the experiments regarding named entity recognition and sentiment analysis.

7. Bibliographical References

Aggarwal, C. C. and Zhai, C. (2012). A survey of text classification algorithms. In *Mining text data*, pages 163–222. Springer.

Awasthi, A. and Sarawagi, S. (2019). Continual learning with neural networks: A review. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pages 362–365. ACM.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Domain adaptation for sentiment classification. In *45th Annv. Meeting of the Assoc. Computational Linguistics (ACL'07)*.

Chen, Z., Badrinarayanan, V., Lee, C.-Y., and Rabinovich, A. (2017). Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. *arXiv preprint arXiv:1711.02257*.

Choi, J., Oh, T.-H., and Kweon, I. S. (2018). Contextually customized video summaries via natural language. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1718–1726. IEEE.

Crichton, G., Pyysalo, S., Chiu, B., and Korhonen, A. (2017). A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18(1):368.

Dave, K., Lawrence, S., and Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM.

Doğan, R. I., Leaman, R., and Lu, Z. (2014). Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Dos Santos, C. and Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78.

Dugar, S. (2019). Aspect-based sentiment analysis using deep neural networks and transfer learning, 3. Master's thesis under supervision of Gerhard Hagerer, M.Sc, and PD Dr. Georg Groh.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Gideon, J., Khorram, S., Aldeneh, Z., Dimitriadis, D., and Provost, E. M. (2017). Progressive neural networks for transfer learning in emotion recognition. *CoRR*, abs/1706.03256.

Gupta, S. (2019). Neural transfer learning for natural language processing, 4. Master's thesis under supervision of Gerhard Hagerer, M.Sc, and PD Dr. Georg Groh.

Hodas, N. O., Shaffer, K., Yankov, A., Corley, C. D., Anderson, A., and Cheney, W. (2017). Beyond fine tuning: Adding capacity to leverage few labels. Technical report, Pacific Northwest National Lab.(PNNL), Richland, WA (United States).

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y., and Collier, N. (2004). Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 70–75. Citeseer.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N. C., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. (2016). Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796.

Lafferty, J. D., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.

Lakkaraju, H., Socher, R., and Manning, C. (2014). Aspect specific sentiment analysis using hierarchical deep learning. In *NIPS Workshop on deep learning and representation learning*.

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A. P., Mattingly, C. J., Wiegers, T. C., and Lu, Z. (2016). Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Luo, G., Huang, X., Lin, C.-Y., and Nie, Z. (2015). Joint entity recognition and disambiguation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 879–888, Lisbon, Portugal, September. Association for Computational Linguistics.

Ma, X. and Hovy, E. H. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *CoRR*, abs/1603.01354.

Mäntylä, M. V., Graziotin, D., and Kuutila, M. (2016). The evolution of sentiment analysis - A review of research topics, venues, and top cited papers. *CoRR*, abs/1612.01556.

Min, S., Seo, M., and Hajishirzi, H. (2017). Question answering through transfer learning from large fine-grained supervision data. *arXiv preprint arXiv:1702.02171*.

Mou, L., Meng, Z., Yan, R., Li, G., Xu, Y., Zhang, L., and Jin, Z. (2016). How transferable are neural networks in nlp applications? *arXiv preprint arXiv:1603.06111*.

Passos, A., Kumar, V., and McCallum, A. (2014). Lexicon infused phrase embeddings for named entity resolution. *CoRR*, abs/1404.5367.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Mohammad, A.-S., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30.

Rei, M. (2017). Semi-supervised multitask learning for sequence labeling. *arXiv preprint arXiv:1704.07156*.

Riemer, M., Khabiri, E., and Goodwin, R. (2017). Representation stability as a regularizer for improved text analytics transfer learning. *arXiv preprint arXiv:1704.03617*.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. (2016). Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Schmitt, M., Steinheber, S., Schreiber, K., and Roth, B. (2018). Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks. *CoRR*, abs/1808.09238.

Schouten, K. and Frasincar, F. (2015). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830.

Wang, Y., Huang, M., Zhu, X., and Zhao, L. (2016). Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas, November. Association for Computational Linguistics.

Wojatzki, M., Ruppert, E., Holschneider, S., Zesch, T., and Biemann, C. (2017). Germeval 2017: Shared task on aspect-based sentiment in social media customer feedback. *Proceedings of the GermEval*, pages 1–12.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June. Association for Computational Linguistics.

5.4 End-to-End Annotator Bias Approximation on Crowdsourced Single-Label Sentiment Analysis

The publication on the consecutive pages is relevant to the examination. It was accepted after peer-review as full paper at the 2021 4th International Conference on Natural Language and Speech Processing. Gerhard Johann Hagerer, the author of the present thesis, is the first author of that paper. His author role, the reprinting permission, and his following contributions are acknowledged by all authors Hagerer, Szabo, Koch, Ripoll Dominguez, Widmer, Wich, Danner, and Groh [2021e]:

> *"Gerhard Johann Hagerer headed the research project. He developed the research idea, the concept, and the methodology of the paper. He managed the data annotation process of the released dataset. Furthermore, he directed the implementation process and reviewed the source code deeply. Regarding the writing of the paper, he created the outline, directed the drafting, wrote large textual parts, incorporated extensive reviewer feedback, and paraphrased, corrected, combined, rewrote, and otherwise improved drafted material."*

Publication Summary

The following publication investigates the role of annotator bias in crowdsourced sentiment analysis annotations when trying to predict them using artificial neural networks. Every annotator is modeled separately, such that the bias of each annotator is factored out during model training and a common ground truth is found as well. Thus, annotator bias is removed during training, which is shown theoretically and empirically. The work is a contribution to state-of-the-art (SOTA) crowdsourcing and bias modeling as to a completely unbiased training procedure, which we did not find in the related work. Classification accuracy and modeling overall is improved by our approach. However, the accuracy does still not meet our requirements to be used for Directed Content Analysis, even if we reduce the complexity of our task to three class sentiment analysis. While we claim that our proposed method is worth being considered for future crowdsourcing research, we are not able to circumvent the problems imposed by the aspect-based sentiment analysis (ABSA) annotations on our organic dataset up to this point.

End-to-End Annotator Bias Approximation on Crowdsourced Single-Label Sentiment Analysis

Gerhard Hagerer, **David Szabo**, **Andreas Koch**, **Maria Luisa Ripoll Dominguez**, **Christian Widmer**, **Maximilian Wich**, **Hannah Danner**, **Georg Groh**
Technical University of Munich, Germany
{ghagerer, grohg}@mytum.de

Abstract

Sentiment analysis is often a crowdsourcing task prone to subjective labels given by many annotators. It is not yet fully understood how the annotation bias of each annotator can be modeled correctly with state-of-the-art methods. However, resolving annotator bias precisely and reliably is the key to understand annotators' labeling behavior and to successfully resolve corresponding individual misconceptions and wrongdoings regarding the annotation task. Our contribution is an explanation and improvement for precise neural end-to-end bias modeling and ground truth estimation, which reduces an undesired mismatch in that regard of the existing state-of-the-art. Classification experiments show that it has potential to improve accuracy in cases where each sample is annotated only by one single annotator. We provide the whole source code publicly[1] and release an own domain-specific sentiment dataset containing 10,000 sentences discussing organic food products[2]. These are crawled from social media and are singly labeled by 10 non-expert annotators.

1 Introduction

Modeling annotator bias in conditions where each data point is annotated by multiple annotators, below referred to as multi-labeled crowdsourcing, has been investigated thoroughly. However, bias modeling when every data point is annotated by only one person, hereafter called singly labeled crowdsourcing, poses a rather specific and difficult challenge. It is in particular relevant for sentiment analysis, where singly labeled crowdsourced datasets are prevalent. This is due to data from the social web which is annotated by the data creators themselves, e.g., rating reviewers or categorizing image uploaders. This might further include multi-media contents such as audio, video, images, and other forms of texts. While the outlook for such forms of data is promising, end-to-end approaches have not yet been fully explored on these types of crowdsourcing applications.

With these benefits in mind, we propose a neural network model tailored for such data with singly labeled crowdsourced annotations. It computes a latent truth for each sample and the correct bias of every annotator while also considering input feature distribution during training. We modify the loss function such that *the annotator bias converges towards the actual confusion matrix of the regarding annotator and thus models the annotator biases correctly*. This is novel, as previous methods either require a multi-labeled crowdsourcing setting (Dawid and Skene, 1979; Hovy et al., 2013) or do not produce a correct annotator bias during training which would equal the confusion matrix, see Zeng et al. (2018, figure 5) and Rodrigues and Pereira (2018, figure 3). A correct annotator- or annotator-group bias, however, is necessary to derive correct conclusions about the respective annotator behavior. This is especially important for highly unreliable annotators who label a high number of samples randomly – a setting, in which our proposed approach maintains its correctness, too.

Our contributions are as follows. We describe the corresponding state-of-the-art for crowdsourcing algorithms and tasks in section 2. Our neural network model method for end-to-end crowdsourcing modeling is explained in section 3, which includes a mathematical explanation that our linear bias modeling approach yields the actual confusion matrices. The experiments in section 4 underline our proof, show that the model handles annotator bias correctly as opposed to previous models, and demonstrate how the approach impacts classification.

[1] https://github.com/theonlyandreas/end-to-end-crowdsourcing
[2] https://github.com/ghagerer/organic-dataset

2 Related Work

2.1 Crowdsourcing Algorithms

Problem definition. The need for data in the growing research areas of machine learning has given rise to the generalized use of crowdsourcing. This method of data collection increases the amount of data, saves time and money but comes at the potential cost of data quality. One of the key metrics of data quality is annotator reliability, which can be affected by various factors. For instance, the lack of rater accountability can entail spamming. *Spammers* are annotators that assign labels randomly and significantly reduce the quality of the data. Raykar and Yu (2012) and Hovy et al. (2013) addressed this issue by detecting spammers based on rater trustworthiness and the SpEM algorithm. However, spammers are not the only source of label inconsistencies. The varied personal backgrounds of crowd workers often lead to *annotator biases* that affect the overall accuracy of the models. Several works have previously ranked crowd workers (Hovy et al., 2013; Whitehill et al., 2009; Yan et al., 2010), clustered annotators (Peldszus and Stede, 2013), captured sources of bias (Wauthier and Jordan, 2011) or modeled the varying difficulty of the annotation tasks (Carpenter, 2008; Whitehill et al., 2009; Welinder et al., 2010) allowing for the elimination of unreliable labels and the improvement of the model predictions.

Ground truth estimation. One common challenge in crowdsourced datasets is the ground truth estimation. When an instance has been annotated multiple times, a simple yet effective technique is to implement majority voting or an extension thereof (TIAN and Zhu, 2015; Yan et al., 2010). More sophisticated methods focus on modeling label uncertainty (Spiegelhalter and Stovin, 1983) or implementing bias correction (Snow et al., 2008; Camilleri and Williams, 2020). These techniques are commonly used for NLP applications or computer vision tasks (Smyth et al., 1995; Camilleri and Williams, 2020). Most of these methods for inferring the ground truth labels use variations of the EM algorithm by Dawid and Skene (1979), which estimates annotator biases and latent labels in turns. We use its recent extension called the *Fast Dawid-Skene* algorithm (Sinha et al., 2018).

End-to-end approaches. The Dawid-Skene algorithm models the raters' *abilities* as respective bias matrices. Similar examples include GLAD (Whitehill et al., 2009) or MACE (Hovy et al., 2013), which infer true labels as well as labeler expertise and sample difficulty. These approaches infer the ground truth only from the labels and do not consider the input features. *End-to-end approaches* learn a latent truth, annotator information, and feature distribution jointly during actual model training (Zeng et al., 2018; Khetan et al., 2017; Rodrigues and Pereira, 2018). Some works use the EM algorithm (Raykar et al., 2009), e.g., to learn sample difficulties, annotator representations and ground truth estimates (Platanios et al., 2020). However, the EM algorithm has drawbacks, namely that it can be unstable and more expensive to train (Chu et al., 2020). LTNet models imperfect annotations derived from various image datasets using a single latent truth neural network and dataset-specific bias matrices (Zeng et al., 2018). A similar approach is used for crowdsourcing, representing annotator bias by confusion matrix estimates (Rodrigues and Pereira, 2018). Both approaches show a mismatch between the bias and how it is modeled, see Zeng et al. (2018, figure 5) and Rodrigues and Pereira (2018, figure 3). We adapt the LTNet architecture (see section 3), as it can be used to model crowd annotators on singly labeled sentiment analysis, which, to our knowledge, is not done yet in the context of annotator bias modeling. Recent works about noisy labeling in sentiment analysis do not consider annotator bias (Wang et al., 2019).

2.2 Crowdsourced Sentiment Datasets

Sentiment and Emotion. Many works use the terms *sentiment* and *emotion* interchangeably (Demszky et al., 2020; Kossaifi et al., 2021), whereas sentiment is directed towards an entity (Munezero et al., 2014) but emotion not necessarily. Both can be mapped to valence, which is the affective quality of goodness (high) or badness (low). Since emotion recognition often lacks annotated data, crowdsourced sentiment annotations can be beneficial (Snow et al., 2008).

Multi-Labeled Crowdsourced Datasets. Crowdsourced datasets, such as, Google GoEmotion (Demszky et al., 2020) and the SEWA database (Kossaifi et al., 2021), usually contain multiple labels per sample and require their aggregation using ground truth estimation. Multi-labeled datasets are preferable to singly labeled ones on limited data. Snow et al. (2008) proved that many non-expert annotators give a better performance than a few expert annotators and are cheaper in comparison.

Singly Labeled Crowdsourced Datasets. Singly labeled datasets are an option given a fixed budget and unlimited data. Khetan et al. (2017) showed that it is possible to model worker quality with single labels even when the annotations are made by non-experts. Thus, multiple annotations can not only be redundant but come at the expense of fewer labeled samples. For singly labeled data, it can be distinguished between reviewer annotators and external annotators. Reviewer annotators rate samples they created themselves. It is common in forums for product and opinion reviews where a review is accompanied by a rating. As an example of this, we utilized the TripAdvisor dataset (Thelwall, 2018). Further candidates are the Amazon review dataset (Ni et al., 2019), the Large Movie Review Dataset (Maas et al., 2011), and many more comprising sentiment. External annotators annotate samples they have not created. Experts are needed for complex annotation tasks requiring domain knowledge. These are not crowdsourced, since the number of annotators is small and fixed. More common are external non-experts. Snow et al. (2008) showed that multi-labeled datasets annotated by non-expert improve performance. Khetan et al. (2017) showed that it also performs well in the singly labeled case. Thus, datasets made of singly labeled non-expert annotations can be cheaper, faster, and obtain performances comparable to those comprised of different types of annotations. Our organic dataset is annotated accordingly, see section 4.3.

3 Methodology

3.1 Basic Modeling Architecture

The model choice is determined by the fact that some of our datasets are small. Thus, the model should have only few trainable parameters to avoid overfitting. We utilize a simple attention mechanism, as it is comon for NLP applications. The input words w_j are mapped to their word embeddings $e_{w_j} \in \mathbb{R}^D$ with $j = 1, ..., S$, and S being the input sequence length and D the dimensionality of the input word vectors. These are GloVe embeddings of 50 dimensions pre-trained on 6B English tokens of the "Wikipedia 2014 + Gigaword 5" dataset (Pennington et al., 2014). Then, it computes the attention a_i of each word using the trainable attention vector $e \in \mathbb{R}^D$ via $a_j = e \cdot e_{w_j}$. It takes the accordingly weighted average $z_n = \sum_{i=1}^{S} a_i \cdot e_{w_i}$ of the word vectors with n denoting the n-th sample or input text.

Finally, the classification head is the sigmoid of a simple linear layer $p_n = \text{softmax}(W \cdot z_n + b)$, with $W \in \mathbb{R}^{L \times D}$ and $b \in \mathbb{R}$ as the weights of the model. We refer to this last layer and to p_n as *latent truth layer* or *latent truth*.

3.2 End-to-End Crowdsourcing Model

On top of the basic modeling architecture, the biases of the annotators are modeled as seen in figure 1. The theory is explained by Zeng et al. (2018) as follows:

"*The labeling preference bias of different annotators cause inconsistent annotations. Each annotator has a coder-specific bias in assigning the samples to some categories. Mathematically speaking, let* $\mathcal{X} = \{x_1, \ldots, x_N\}$ *denote the data,* $y^c = [y_1^c, \ldots, y_N^c]$ *the regarding annotations by coder* c*. Inconsistent annotations assume that* $P(y_n^c|x_n) \neq P(y_n^{\hat{c}}|x_n), \forall x_n \in \mathcal{X}, c \neq \hat{c}$*, where* $P(y_n^i|x_n)$ *denotes the probability distribution that coder* c *annotates sample* x_n*.*

LTNet assumes that each sample x_n *has a latent truth* y_n*. Without the loss of generality, let us suppose that LTNet classifies* x_n *into the category* i *with probability* $P(y_n = i|x_n; \Theta)$*, where* Θ *denotes the network parameters. If* x_n *has a ground truth of* i*, coder* c *has an opportunity of* $\tau_{ij}^c = P(y_n^c = j|y_n = i)$ *to annotate* x_n *as* j*, where* y_n^c *is the annotation of sample* x_n *by coder* c*. Then, the sample* x_n *is annotated as label* j *by coder* c *with a probability of* $P(y_n^c = j|x_n; \Theta) = \sum_{i=1}^{L} P(y_n^c = j|y_n = i)P(y_n = i|x_n; \Theta)$*, where* L *is the number of categories and* $\sum_{j=1}^{L} P(y_n^c = j|y_n = i) = \sum_{j=1}^{L} \tau_{ij}^c = 1$*.*
$T^c = [\tau_{ij}^c]_{L \times L}$ *denotes the transition matrix (also referred to as annotator bias) with rows summed to* 1 *while* $[p_n]_i = P(y_n = i|x_n; \Theta)$ *is modeled by the base network (Zeng et al., 2018). We define* $[p_n^c]_j = P(y_n^c = j|x_n; \Theta)$*. Given the annotations from* C *different coders on the data, LTNet aims to maximize the log-likelihood of the observed annotations. Therefore, parameters in LTNet are learned by minimizing the cross entropy loss of the predicted and observed annotations for each coder* c*."*

We represent the annotations and predictions as vectors of dimensionality L such that y_n^c is one-hot encoded and p_n^c contains the probabilities for all class predictions of sample n. The

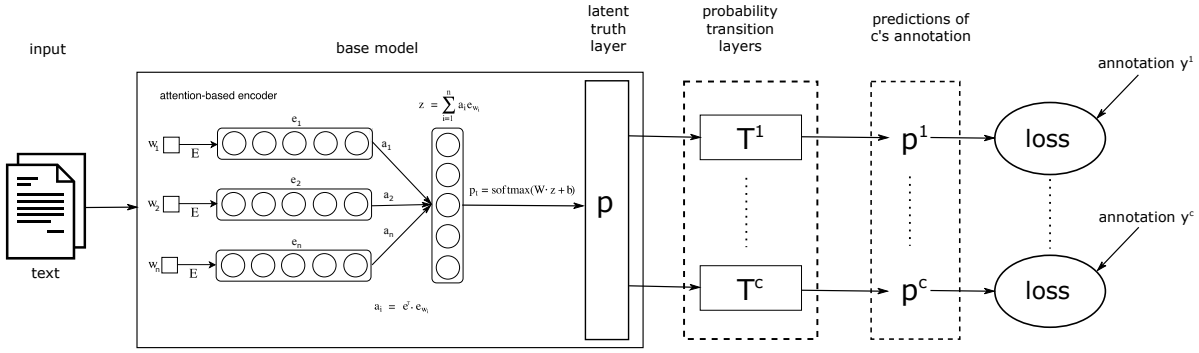


Figure 1: Architecture of the end-to-end trainable LTNet (Zeng et al., 2018). The base model is a simple attention model with a single trainable attention vector e and linear layer with parameters W and b. The transition matrices T^c are the bias matrices from the annotators c. "Each row of the transition matrix T is constrained to be summed to 1" (Zeng et al., 2018). The base model is inspired by ABAE (He et al., 2017).

cross entropy loss function is then defined as $-\sum_{n=1}^{C}\sum_{n=1}^{N}\log(p_n^{c\intercal}\cdot y_n^c)$.

3.3 The Effect of Logarithm Removal on Cross Entropy

The logarithm in the cross entropy formula leads to an exponential increase in the loss for false negative predictions, i.e., when the predicted probability $[p_n^c]_i$ for a ground truth class i is close to 0 and $[y_n^c]_i$ is 1. This increase can be helpful in conditions with numerical underflow, but at the same time this introduces a disproportionate high loss of the other class due to constantly misclassified items. This happens in crowdsourcing, for example, when one annotator is a spammer assigning a high degree of random annotations, which in turn leads to a disproportionally higher loss caused by that annotator's many indistinguishable false negative annotations. Consequentially, the bias matrix of that annotator would be biased towards the false classes. Moreover, this annotator would cause overall more loss than other annotators, which can harm the model training for layers which are shared among all annotators, e.g., the latent truth layer when it is actually trained.

By omitting the log function, these effects are removed and all annotators and datapoints contribute with the same weight to the overall gradient and to the trainable annotator bias matrices, independent of the annotator and his respective annotation behavior. As a consequence, the annotator matrices are capable of modeling the real annotator bias, which is the mismatch between an annotation y_n^c of coder c and the latent truth prediction p_n. If p_n is one-hot encoded, this results to the according classification ratios of samples and is equal to the confusion matrix, without an algorithmically encoded bias towards a certain group of items. This is shown mathematically in the following, where it is assumed that the base network is fixed, i.e., backpropagation is performed through the bias matrices and stops at the latent truth layer.

We define $N=\sum_{k=1}^{L}N_k$ as the number of all samples and N_k of class $k=1,...,L$. L is the number of classes, $T^c=[\tau_{ij}^c]_{L\times L}$ the bias matrix of coder c, p_n the latent truth vector of sample $n=1,...,N$, and p_n^c the annotator prediction. p_{km} is the latent truth of the m-th sample of class k with $m=1,...,N_k$, same for x_{km} and y_{km}^c. The loss without logarithm is

$$
\begin{aligned}
\boldsymbol{O} &= -\sum_{n=1}^{N} p_n^{c\intercal}\cdot y_n^c \\
&= -\sum_{k=1}^{L}\sum_{m=1}^{N_k} p_{km}^{\intercal}\cdot T^c\cdot y_{km}^c \\
&= -\sum_{k=1}^{L}\sum_{m=1}^{N_k} p_{km}^{\intercal}\cdot \begin{pmatrix}\tau_{1k}^c\\ \vdots \\ \tau_{Lk}^c\end{pmatrix} \\
&= \sum_{k=1}^{L}\sum_{m=1}^{N_k}\sum_{h=1}^{L} -[p_{km}]_h\cdot\tau_{hk}^c
\end{aligned}
$$

Apparently, the derivation step between the second and third line would not work if there would be the logarithm from the standard cross entropy. Now, let the learning rate be α, the number of epochs E and the starting values of the initialized bias matrix $(\tau_{lh}^c)_0$. The bias parameters τ_{lh}^c of the bias matrix T^c are updated according to

Figure 2: Male and female bias (top) and confusion (bottom) matrices which are trained using cross entropy loss with and without logarithm in two different settings. The left side has only the original annotations, whereas the right side has 80% random male labels.

$$
\begin{aligned}
(\tau^c_{hk})_E &= (\tau^c_{hk})_0 + \sum_{i=1}^{E} \alpha \left(\frac{\partial \boldsymbol{O}}{\partial \tau^c_{hk}} \right)_i \\
&= (\tau^c_{hk})_0 + \sum_{i=1}^{E} \alpha \left[\sum_{m=1}^{N_k} - [p_{km}]_h \right]_i \\
&= (\tau^c_{hk})_0 - \alpha E \underbrace{\sum_{m=1}^{N_k} [p_{km}]_h}_{=:Z_{hk}}
\end{aligned}
$$

For sufficiently large E the starting values $(\tau^c_{hk})_0$ become infinitesimally small in comparison to the second additive term and thus negligible. As we are normalizing the rows of $(T^c)_E$ after training so that the bias fulfills our probability constraint defined in section 3.2, the linear factor $-\alpha E$ is canceled out, too. Thus, the bias matrix T^c results in the row normalized version of $[Z_{hk}]_{L\times L}$. Z_{hk} is the sum of the latent truth probabilities for class h on all samples of a ground truth class k. If we assume that the latent truth is one hot encoded, $[Z_{hk}]_{L\times L}$ equals to the confusion matrix, of which the k-th column sums up to the number of samples in class k: $\sum_{h=1}^{L} Z_{hk} = \sum_{h=1}^{L} \sum_{m=1}^{N_k} [p_{km}]_h = \sum_{m=1}^{N_k} 1 = N_k$.

4 Experiments

4.1 Bias Convergence

The following experiment compares how training with and without the logarithm in the cross entropy loss affects the LTNet bias matrices empirically. The mathematical explanations in section 3.3 suggest that the logarithm removal from cross entropy leads to an annotator bias matrix identical to the confusion matrix, which would not be the case for the normal cross entropy.

Experiment Description. For the data, we use the TripAdvisor dataset from Thelwall et al. consisting of 11,900 English consumer reviews about hotels from male and female reviewers plus their self-assigned sentiment ratings (Thelwall, 2018). We use the gender information to split the data into two annotator groups, male and female, from which we model each one with a corresponding bias matrix. We exclude neutral ratings and binarize the rest to be either positive or negative. As the dataset is by default completely balanced regarding gender and sentiment at each rating level, it is a natural candidate for correct bias approximation. Throughout our experiments, we use 70% of the obtained data as training, 20% as validation and the 10% remaining as test sets.

Similar to the explanation in 3.3, the base model with its latent truth predictions is pre-trained on all samples and then frozen when the bias matrices are trained. The stochastic gradient descent method is used to optimize the parameters, as other widespread optimizers, such as Adam and AdaGrad (the latter introduced that feature first), introduce an – in our case undesired – bias towards certain directions in the gradient space, namely by using the previous learning steps to increase or decrease the weights along dimensions with larger or smaller gradients (Kingma and Ba, 2014). For all four sub-experiments, we train the base models with varying hyperparameters and pick the best based on accuracy. We train the transition matrices 50 times with different learning rates from the interval $[1e{-}6, 1e{-}3]$. The batch size is 64. In addition to a normal training setting, we add random annotations to 80% of the instances annotated by male

subjects, such that 40% from them are wrongly annotated. This results in four models: with and without logarithm in the cross entropy, with and without random male annotations, each time respectively with two annotator group matrices, male and female – see figure 2.

Results. The bias matrices of the models with the best accuracy are picked and presented in figure 2 in the top row. The corresponding confusion matrices depict the mismatch between latent truth predictions and annotator-group labels in the bottom row. The bias matrices trained without logarithm in the cross entropy are almost identical to the confusion matrices in all cases, which never holds for the normal cross entropy. This confirms our mathematically justified hypothesis given in section 3.3 that the logarithm removal from cross entropy leads to a correctly end-to-end-trained bias. In this context, it is relevant that the related work shows the same mismatch between bias and confusion matrix when applying cross entropy loss without explaining nor tackling this difference, see Zeng et al. (2018, figure 5) and Rodrigues and Pereira (2018, figure 3).

It is worth mentioning for the 80% random male annotations that these are correctly modeled without cross entropy, too, as opposed to normal cross entropy. If the goal is to model the annotator bias correctly in an end-to-end manner, this might be considered as particularly useful to analyze annotator behavior, e.g., spammer detection, later on.

Finally, we report how much variation the bias matrices show during training for cross entropy with and without logarithm. As mentioned in the experiment description, we trained each model 50 times. The elements of the resulting bias matrices with standard cross entropy have on average 7.7% standard deviation compared to 2.8% without logarithm. It can be concluded that the bias produced by standard cross entropy is less stable during training, which raises questions about the overall reliability of its outcome.

In summary, the observations confirm our assumptions that cross entropy without logarithm captures annotator bias correctly in contrast to standard cross entropy. This carries the potential to detect spammer annotators and leads to an overall more stable training.

4.2 Ground Truth Estimation

In the following paragraphs, we demonstrate how to estimate the ground truth based on the latent truth from LTNet. This is then compared to two other kinds of ground truth estimates. All of them can be applied in a single label crowdsourcing setting.

The Dawid-Skene algorithm (Sinha et al., 2018) is a common approach to calculate a ground truth in crowdsourcing settings where there are multiple annotations given on each sample. This method is, for instance, comparable to majority voting, which tends to give similar results for ground truth estimation. However, in single label crowdsourcing settings, these approaches are not feasible. Under single label conditions, the Dawid-Skene ground truth estimates equal to the single label annotations.

This is given by Sinha et al. (2018, formula 1) in the expectation step, where the probability for a class $k \in 1, 2, ..., L$ given the annotations is defined as

$$P(Y_n = k|k_{n_1}, k_{n_2}, ..., k_{n_L}) = \frac{\left(\prod_{c=1}^{C} P(k_{n_c}|Y_n = k)\right) \cdot P(Y_n = k)}{\sum_{k=1}^{L}\left(\prod_{c=1}^{C} P(k_{n_c}|Y_n = k)\right) \cdot P(Y_n = k)}.$$

Here, n is the sample to be estimated, C the number of annotators for that sample, $n_1, n_2, ..., n_C$ the set of annotators who labeled this sample, $k_{n_1}, k_{n_2}, ..., k_{n_C}$ the set of annotation choices chosen by these C participants for sample n, and Y_n the correct (or aggregated) label to be estimated for the sample n (Sinha et al., 2018).

In the single label case C equals to 1, which reduces the formula to $P(Y_n = k|k_{n_1}, k_{n_2}, ..., k_{n_C}) = P(Y_n = k|k_{n_1})$. This in turn equals to 1 if k is the assigned class label to sample n by annotator n_1, or 0 otherwise. In other words, if there is only one annotation per sample, this annotation defines the ground truth. Since different annotators do not assign labels on the same samples, there is also no way to model mutual dependencies of each other.

LTNet, however, provides estimates for all variables from this formula. $P(Y_n = k)$ is the prior and is approximated by the latent truth probability for class k of sample n. $P(k_{n_c}|Y_n = k)$ is the probability that, assuming k would be the given class, sample n is labeled as k_{n_c} by annotator n_c. This equals to $\tau^c_{k_{n_c},k}$, i.e., the entries of the LTNet bias matrix T^c of annotator c.

Eventually, the LTNet ground truth can be derived by choosing k such that the probability $P(Y_n = k|k_{n_1}, ...)$ is maximized:

$$k_{\text{ground truth}} = \arg\max_k P(Y_n = k | k_{n_1}, ...).$$

We will leverage this formula to derive and evaluate the ground truth generated by LTNet.

Experiment We calculate the LTNet ground truth according to the previous formula on the organic dataset, a singly labeled crowdsourcing dataset, which is described in Section 4.3. To demonstrate the feasibility and the soundness of the approach, we compare it with two other ways of deriving a ground truth. Firstly, we apply the fast Dawid-Skene algorithm on the annotator-wise class predictions from the LTNet model. Secondly, we train a base network on all annotations while ignoring which annotator annotated which samples. Eventually, we compare the ground truth estimates of all three methods by calculating Cohen's kappa coefficient (Cohen, 1960), which is a commonly used standard to analyze correspondence of annotations between two annotators or pseudo annotators. The training procedures and the dataset are identical to the ones from the classification experiments in Section 4.3.

Results As can be seen on Table 1, the three ground truth estimators are all highly correlated to each other, since the minimal Cohen's kappa score is 0.98. Apparently, there are only minor differences in the ground truth estimates, if any at all. Thus, it appears that the ground truths generated by the utilized methods are mostly identical. Especially, the LTNet and Dawid-Skene ground truths are highly correlated with a kappa of 99%. The base model, which is completely unaware of which annotator labeled which sample, is slightly more distant with kappas between 98% – 99%. So with respect to the ground truth itself, we do not see a specific benefit of any method, since they are almost identical.

However, it must be noted that LTNet additionally produces correct bias matrices of every annotator during model training, which is not the case for the base model. Correct biases have the potential to help improving model performance by analyzing which annotators tend to be more problematic and weighting them accordingly.

4.3 Classification

We conduct classification comparing LTNet in different configurations on three datasets with crowdsourced sentiment annotations to discuss the potential related benefits and drawbacks of our proposed loss modification.

Ground truths	Dawid Skene	LTNet	Basic Model
Dawid Skene	1.0000	0.9905	0.9832
LTNet	0.9905	1.0000	0.9918
Base Model	0.9832	0.9918	1.0000

Table 1: Cohen's kappa scores between three different ground truth estimation methods applied on the singly labeled crowdsourced organic dataset.

Emotion Dataset. The emotion dataset consists of 100 headlines and their ratings for valence by multiple paid Amazon Mechanical Turk annotators (Snow et al., 2008). Each headline is annotated by 10 annotators, and each annotated several but not all headlines. We split the interval-based valence annotations to positive, neutral, or negative. Throughout our experiments, we used 70% of the obtained data as training, 20% as validation and 10% as test sets.

Organic Food Dataset. With this paper, we publish our dataset containing social media texts discussing organic food related topics.

Source. The dataset was crawled in late 2017 from Quora, a social question-and-answer website. To retrieve relevant articles from the platform, the search terms "organic", "organic food", "organic agriculture", and "organic farming" are used. The texts are deemed relevant by a domain expert if articles and comments deal with organic food or agriculture and discuss the characteristics, advantages, and disadvantages of organic food production and consumption. From the filtered data, 1,373 comments are chosen and 10,439 sentences annotated.

Annotation Scheme. Each sentence has sentiment (positive, negative, neutral) and entity, the sentiment target, annotated. We isolate sentiments expressed about organic against non-organic entities, whereas for classification only singly labeled samples annotated as organic entity are considered. Consumers discuss organic or non-organic products, farming practices, and companies.

Annotation Procedure. The data is annotated by each of the 10 coders separately; it is divided into 10 batches of $1,000$ sentences for each annotator and none of these batches shared any sentences between each other. 4616 sentences contain organic entities with 39% neutral, 32% positive, and 29% negative sentiments. After annotation, the

Dataset	Model	F1 %	Acc %
TripAdvisor	Base Model	88.92	88.91
	LTNet w/o log	**89.71**	**89.71**
	LTNet	89.39	89.39
Organic	Base Model	32.08	45.75
	LTNet w/o log	**44.71**	**50.54**
	LTNet	40.51	47.77
Emotion	Base Model	51.74	56.00
	LTNet w/o log	58.15	63.00
	LTNet	**61.23**	**66.00**
	Base Model DS	44.17	54.00

Table 2: Macro F1 scores and accuracy measured in the classification experiment.

data splits are 80% training, 10% validation, and 10% test set. The data distribution over sentiments, entities, and attributes remains similar on all splits.

Experiment Description. The experiment is conducted on the TripAdvisor, organic, and emotion datasets introduced in section 4.3. We compare the classification of the base network with three different LTNet configurations. Two of them are trained using cross entropy with and without logarithm. For the emotion dataset, we compute the bias matrices and the ground truth for the base model using the fast Dawid-Skene algorithm (Sinha et al., 2018). This is possible for the emotion dataset, since each sample is annotated by several annotators.

We apply pre-training for each dataset by training several base models with different hyperparameters and pick the best based on accuracy. Eventually, we train the LTNet model on the crowdsourcing annotation targets by fine-tuning the best base model together with the bias matrices for the respective annotators. The bias matrices are initialized as row normalized identity matrices plus uniform noise around 0.1. The models are trained 50 times with varying learning rates sampled from between $[1e{-}6, 1e{-}3]$. A batch size of 64 is used.

Results. The classification results of the models are presented in table 2 with their macro F1 score and accuracy as derived via predictions on the test sets. LTNet generally shows a significant classification advantage over the base model. On all three databases, LTNet approaches performed better on the test datasets. The LTNet improvement has a big delta of $11\% + / - 1\%$ when there is a low annotation reliability (organic and emotion datasets) and a small delta $< 1\%$ with high reliability (TripAdvisor) [3]. Apparently, model each annotator separately gives significant advantages.

Regarding the comparison between cross entropy (CE) loss with and without logarithm on LTNet, the removed logarithm shows better classification results on organic ($+3\%$) and TripAdvisor data ($+0.3\%$) and worse on the emotion dataset (-3%). This means that on both of the singly labeled crowdsourcing datasets, the removal of the logarithm from the loss function leads to better predictions than the standard CE loss. On the multi-labeled emotion dataset, however, this does not appear to be beneficial. As this data has only a very small test set of 100 samples, it is not clear if this result is an artifact or not. Concluding, the log removal appears to be beneficial on large datasets, where the bias is correctly represented in the training and test data splits, such that it can be modeled correctly by the denoted approach. It shall be noted, that it is not clear if that observation would hold generally. We advice to run the same experiments multiple times on many more datasets to substantiate this finding.

5 Conclusion

We showed the efficacy of LTNet for modeling crowdsourced data and the inherent bias accurately and robustly. The bias matrices produced by our modified LTNet improve such that they are more similar to the actual bias between the latent truth and ground truth. Moreover, the produced bias shows high robustness under very noisy conditions making the approach potentially usable outside of lab conditions. The latent truth, which is a hidden layer below all annotator biases, can be used for ground truth estimation in our single label crowdsourcing scenario, providing almost identical ground truth estimates as pseudo labeling. Classification on three crowdsourced datasets show that LTNet approaches outperfom naive approaches not considering each annotator separately. The proposed log removal from the loss function showed better results on singly labeled crowdsourced datasets, but this observation needs further experiments to be substantiated. Furthermore, there might be many use cases to explore the approach on other tasks than sentiment analysis.

[3] Unreliable means that the provided annotations have a low Cohen's kappa inter-rater reliability on the organic 51.09% and emotion (27.47%) dataset. On the organic dataset we prepared a separate data partition of 300 sentences annotated by all annotators for that purpose. For the TripAdvisor dataset, it is apparent that the correspondence of annotations between the two annotator groups (male and female) is high as can be seen in figure 2 for cross entropy without logarithm.

References

Michael P. J. Camilleri and Christopher K. I. Williams. 2020. The extended dawid-skene model. In *Machine Learning and Knowledge Discovery in Databases*, pages 121–136, Cham. Springer International Publishing.

Bob Carpenter. 2008. Multilevel bayesian models of categorical data annotation.

Zhendong Chu, Jing Ma, and Hongning Wang. 2020. Learning from crowds by modeling common confusions.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 28(1):20–28.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397, Vancouver, Canada. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia. Association for Computational Linguistics.

Ashish Khetan, Zachary C. Lipton, and Anima Anandkumar. 2017. Learning from noisy singly-labeled data.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, K. Star, E. Hajiyev, and M. Pantic. 2021. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1022–1040.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA. Association for Computational Linguistics.

M. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen. 2014. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*, 5(2):101–111.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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 188–197, Hong Kong, China. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2013. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 196–204, Sofia, Bulgaria. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Emmanouil Antonios Platanios, Maruan Al-Shedivat, Eric Xing, and Tom Mitchell. 2020. Learning from imperfect annotations.

Vikas C. Raykar and Shipeng Yu. 2012. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *J. Mach. Learn. Res.*, 13:491–518.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Anna K. Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. 2009. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *ICML*, volume 382 of *ACM International Conference Proceeding Series*, pages 889–896. ACM.

Filipe Rodrigues and Francisco C. Pereira. 2018. Deep learning from crowds. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):1611–1618.

Vaibhav B Sinha, Sukrut Rao, and Vineeth N Balasubramanian. 2018. Fast dawid-skene: A fast vote aggregation scheme for sentiment classification.

Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. 1995. Inferring ground truth from subjective labelling of venus images. In *Advances in Neural Information Processing Systems*, volume 7, pages 1085–1092, San Diego, CA. MIT Press.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Morristown, NJ, USA. Association for Computational Linguistics.

DJ Spiegelhalter and PGI Stovin. 1983. An analysis of repeated biopsies following cardiac transplantation. *Statistics in medicine*, 2(1):33–40.

Mike Thelwall. 2018. Gender bias in sentiment analysis. *Online Information Review*, 42(3):343–354.

TIAN TIAN and Jun Zhu. 2015. Max-margin majority voting for learning from crowds. In *Advances in Neural Information Processing Systems*, volume 28, pages 1621–1629, San Diego, CA. Curran Associates, Inc.

Hao Wang, Bing Liu, Chaozhuo Li, Yan Yang, and Tianrui Li. 2019. Learning with noisy labels for sentence-level sentiment classification. *CoRR*, abs/1909.00124.

Fabian L. Wauthier and Michael I. Jordan. 2011. Bayesian bias mitigation for crowdsourcing. In *NIPS*, volume 24, pages 1800–1808, San Diego, CA. Curran Associates, Inc.

Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. 2010. The multidimensional wisdom of crowds. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, Red Hook, NY, USA. Curran Associates Inc.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier R. Movellan. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *NIPS*, pages 2035–2043, San Diego, CA. Curran Associates, Inc.

Yan Yan, Rmer Rosales, Glenn Fung, Mark W. Schmidt, Gerardo Hermosillo Valadez, Luca Bogoni, Linda Moy, and Jennifer G. Dy. 2010. Modeling annotator expertise: Learning when everybody knows a bit of something. In *AISTATS*, volume 9 of *JMLR Proceedings*, pages 932–939, Chia Laguna Resort, Sardinia, Italy. JMLR.org.

Jiabei Zeng, Shiguang Shan, and Xilin Chen. 2018. Facial expression recognition with inconsistently annotated datasets. In *ECCV (13)*, volume 11217 of *Lecture Notes in Computer Science*, pages 227–243, Red Hook, NY, USA. Springer.

6 Dataset: Comments From Teaching Course Evaluations

6.1 An Analysis of Programming Course Evaluations Before and After the Introduction of an Autograder

The publication on the consecutive pages is relevant to the examination. It was accepted after peer-review as full paper at the 2021 19th International Conference on Information Technology Based Higher Education and Training. Gerhard Johann Hagerer, the author of the present thesis, is the first author of that paper. His author role, the reprinting permission, and his following contributions are acknowledged by all authors Hagerer, Lahesoo, Anschütz, and Krusche [2021b]:

> *"Gerhard Johann Hagerer headed the research project. He developed the research idea, the concept, and the basic methodology of the paper, and he collected the dataset. Furthermore, he lead the implementation process and reviewed the source code deeply. Regarding the writing of the paper, he created the outline, directed the drafting, and contributed significant content to the paper, by writing large textual parts, incorporating extensive reviewer feedback, and paraphrasing, correcting, combining, and otherwise improving drafted material."*

Publication Summary

The following publication investigates feedback given in evaluation questionnaires by students who participated in several large scale programming courses at the university. The goal here is to depict the possible impact of an introduced autograding system on the satisfaction of the students. We find a strong correlation between topic modeling on open-ended comments with numerical Likert scale answers, which goes hand-in-hand with the findings of the related work. This is relevant for the dissertation at hand, since we discuss how natural language processing (NLP) and opinion mining could support or complement representative opinion surveys. Furthermore, the study shows that the autograding system changed students' satisfaction with the teaching courses, and topic modeling is helpful to investigate the possible reasons for this change. The research is supposed to highlight the importance of unsupervised methods and transfer learning for a representative form of opinion mining.

2021 19th International Conference on Information Technology Based Higher Education and Training (ITHET) | 978-1-7281-8883-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ITHET50392.2021.9759809

An Analysis of Programming Course Evaluations Before and After the Introduction of an Autograder

Gerhard Hagerer, Laura Lahesoo, Miriam Anschütz, Stephan Krusche and Georg Groh
Department of Informatics, Technical University of Munich
Boltzmannstraße 3, 85748 Garching nearby Munich, Germany
Email: {gerhard.hagerer,laura.lahesoo,m.anschuetz,krusche}@tum.de, grohg@in.tum.de

***Abstract*—Commonly, introductory programming courses in higher education institutions have hundreds of participating students eager to learn to program. The manual effort for reviewing the submitted source code and for providing feedback can no longer be managed. Manually reviewing the submitted homework can be subjective and unfair, particularly if many tutors are responsible for grading. Different autograders can help in this situation; however, there is a lack of knowledge about how autograders can impact students' overall perception of programming classes and teaching. This is relevant for course organizers and institutions to keep their programming courses attractive while coping with increasing students.**

This paper studies the answers to the standardized university evaluation questionnaires of multiple large-scale foundational computer science courses which recently introduced autograding. The differences before and after this intervention are analyzed. By incorporating additional observations, we hypothesize how the autograder might have contributed to the significant changes in the data, such as, improved interactions between tutors and students, improved overall course quality, improved learning success, increased time spent, and reduced difficulty. This qualitative study aims to provide hypotheses for future research to define and conduct quantitative surveys and data analysis. The autograder technology can be validated as a teaching method to improve student satisfaction with programming courses.

***Index Terms*—educational software, educational technology, automated grading, assessment tools, higher education, computer science, course assessment, feedback, teaching evaluations**

I. Introduction

At the Technical University of Munich, the number of freshmen computer science students doubled in recent years and reached more than 2500. Programming is a crucial skill for their academic and professional careers in engineering and natural and social sciences. Therefore, many instructors apply autograding to programming exercises to provide immediate feedback to students and lower the manual grading effort.

There are different existing autograders such as WebCat [1], JACK [2], Praktomat [3], GraJa [4], and Artemis [5]. These systems have been thoroughly evaluated with user studies incorporating quantitative analyses of user behavior and questionnaires. Also, the impact on learning success in terms of students' grades has been analyzed [6], [5].

However, little is known regarding how autograding relates to student satisfaction with the course and its varied teaching aspects. Therefore, course evaluations are a standard method for course organizers and lecturers to understand which parts of the course the students liked and which did not. Especially open-ended comments can be insightful since they contain opinions about emerging course aspects that are not being asked in Likert scale questions. Here, a two-pronged analysis approach using text mining in addition to basic statistics ensures no information is lost. In that regard, topic modeling is an established method in qualitative research to derive theories and hypotheses about how certain factors shape the human experience in specific environments, e.g., how an autograding tool can influence the interaction between students and tutors in teaching sessions[1].

We investigate if there are consistent patterns of how the course evaluations changed when the autograder Artemis was introduced in three different courses over at least two years, i.e., before and after the intervention. In particular, we focus on the expressed satisfaction with different aspects of the courses according to the students' opinions. We are interested in the following research questions:

- RQ1 How did students report on their learning experience in course evaluations, and how did it change?
- RQ2 How did the reported interaction between students and tutors change?
- RQ3 How did the perceived difficulty of the practical programming parts of the courses change?
- RQ4 How did the perceived overall course quality change?

We collected university course evaluations consisting of Likert scale questions and free text comments for three courses. In each evaluation, 100-500 students, see Table I, provided ratings alongside positive and critical comments about the course implementation, software, teachers, structure, and exercises. We analyze the evaluation differences before and after the Artemis introduction using statistics and topic modeling.

Section II provides related work regarding text mining on course evaluation comments and how autograders have been evaluated so far. Section III describes the autograder Artemis and its features in more detail. In Section IV, we present the study design with the course profiles and the evaluation analysis. Section V presents the results and Section VI contains the main findings of this paper. Finally, section VII concludes the paper.

[1]Tutors are students who have passed the course and take on the role of a student lecturer. Tutors hold tutor groups, in which they support students in working on assignments, answer questions, discuss solutions, help with problems, and grade the assignments based on pre-defined grading criteria.

978-1-7281-8883-6/21/$31.00 ©2021 IEEE

II. Related Work

This section motivates why we additionally leverage text mining on course evaluations and why they are a meaningful source for evaluating autograders.

A. Text Mining Applications on Course Evaluations

Evaluation questionnaires at universities often provide numerical Likert scale questions along with open-ended questions to their participants, such that they can express additional thoughts about the course as free-text comments. Manually processing them can be labor-intensive, leading to a trend to process comments automatically with text mining for reproducible results [7], [8]. This can be applied to multiple courses to harmonize the evaluations and to make the results comparable [9], [10]. The approach carries potential for new applications in educational marketing, e.g., comparing courses on a faculty or university level to improve the overall teaching quality and giving insights about the perception towards an institution [11].

Relevant text mining techniques for course evaluation analysis are topic modeling and sentiment analysis. These are used in automated tools such as Palaute [12], [7] and SMF [13], visualizing the underlying topic distribution with non-negative matrix factorization (NMF) or latent Dirichlet allocation (LDA) and students' sentiment [14], [10]. Even though there is a plethora of more advanced state-of-the-art text mining methods, we do not consider any which have not been previously applied on course evaluations. Instead, there are promising method evaluations of the previously mentioned techniques [15], [9], [11], for instance, to depict topic trends of a course over several years [16]. This motivates our choice for NMF-based topic modeling applied on course evaluations of multiple courses and years to measure the impact of a new autograding tool. We relate this type of topic distributions in free-text comments with Likert scale answers, as this technique showed statistically significant correlations between both [9]. This has the potential to compensate for missing data points and biased open- or close-ended questions. However, the relation between both answer types has not been demonstrated qualitatively, and there is a lack of related applied research regarding the impact of new teaching concepts or tools.

B. Autograder Evaluations

In this section, we introduce autograding tools and compare how these have been evaluated so far. Evaluation is important as autograders have the potential to improve scaling and fairness, whereas concerns exist regarding the potential lack of personalized feedback and thus decreased motivation.

1) Autograders: JACK [2] is a web-based autograding tool introduced at the university of Duisburg-Essen. The Praktomat [3] autograding tool, used by the KIT and the University of Passau, can grade submissions to multiple languages such as Java, C++ or Haskell. The Web-CAT [1] autograder is a very customizable and extensible autograder that is used as a basis for the GraJa [4] autograder, for example. Vocareum classroom [17] is a commercial software for software development labs that also includes autograding functionality. Keuning [18] provides an overview of automated feedback systems, of which a minority are evaluated with questionnaires and a sufficiently large n. Autograding is not mentioned as functionality.

2) Perceived Usability: Given that web-based autograders store the usage data, the according usage patterns can be analyzed. Previous work evaluated questions, such as, if students liked the appearance and how easy the tool was to use [19], how many submissions were graded and which technical issues the students encountered [4], and what the students liked best and what they liked worst in immediate in-class feedback [20]. While these studies show that autograders can be used successfully by many students in practice to learn programming, it is not clear if and how they contributed to the overall course quality from the students' perspective.

3) Learning Experience: A relevant autograder evaluation aims at testing a gamification concept to improve student engagement in learning programming with Web-CAT [21]. The study uses feedback questionnaires about the learning experience, but it does not explain the role or impact of the autograder.

To evaluate the JACK autograder, the overall course evaluation questionnaire [22] was analyzed, where students referred to the autograder as a positive feature. As only one year of one course is evaluated, it is not clear if the autograder would be seen as beneficial for other courses as well, and it is not compared with when not using an autograder. In a follow-up study, the exam grades were compared to the grades of the previous year that did not use JACK [6]. They report an improvement in the average grade of students using JACK. Other course aspects, e.g., teaching quality, are not considered.

III. Autograder

Artemis [5] is a learning management system with individual feedback [23] that supports interactive learning [24], [25] and is scalable to large courses [26]. It is open source[2] and used by multiple universities and courses. It includes autograding functionality to provide feedback to students regarding programming exercises in interactive instructions which change their status and color based on the progress of students. Completed tasks and correctly implemented model elements are marked in green, incomplete and not yet implemented ones are marked in red. This helps students to identify which parts of the exercise they have already solved correctly and improves the understanding of the source code on the model level. When they submit their current solution, the interactive instructions dynamically update.

The programming exercise workflow is as follows: An instructor sets up a version control repository containing the exercise code handed out to students and test cases to verify students' submissions (*template repository*). It includes a small sample project with predefined classes and dependencies to libraries. The instructor stores the tests for autograding in a separate *test repository*, inaccessible to students. A combination of behavioral (black-box), structural (white-box) tests and static

[2]https://github.com/ls1intum/Artemis

TABLE I
COURSES (ARTEMIS IN BOLD) AND NUMBER OF ANSWERS IN THE DATA .

Course	#Answers	Homework submission	Communication
A.2019	299	Moodle, TUMjudge	
A.2020	**269**	**Artemis**	**Zulip, Tweedback**
O.2016	138	Git	
O.2017	116	Git	
O.2018	182	Moodle	Moodle
O.2019	**169**	**Artemis**	**Moodle**
P.2018	519	Moodle	Piazza
P.2019	**553**	**Artemis**	**Rocket.Chat**

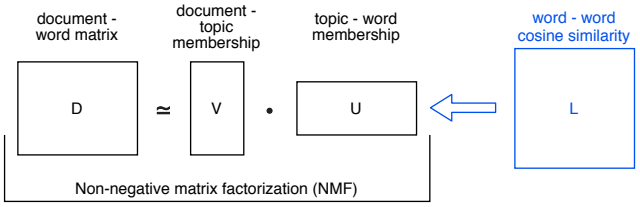


Fig. 1. Black: Topic modeling based on non-negative matrix factorization (NMF). Blue: Transfer learning using word-word cosine similarities of fastText embeddings. The latter ensures that the NMF and fastText word embeddings correspond to each other. This is referred to as knowledge-guided NMF.

code analysis allows to check for functionality, implementation details, and code quality of the submission.

After setting up the template, test, and solution repositories, the instructor configures the build plan on the continuous integration server which compiles and tests the exercise code using the previously defined test cases and the static code analysis configuration (*template build plan*). The build plan includes tasks to pull the source code from the template and test repository whenever changes occur and to combine them so that the tests can be executed in the second step. A final task, which is executed when compilation or test execution fails, notifies Artemis about the new result.

A student starts an exercise with a single click, triggering the setup process: Artemis creates a personal copy of the template and the *student repository* and grants access only to this student. It also creates a personal copy of the template build plan and the *student build plan* and configures it to be triggered when the particular student uploads changes to this repository. The student can not access the build plan to hide its complexity. Personalized means that each student gets one repository and one build plan. When 2,000 students participate in an exercise, Artemis creates 2,000 student repositories and 2,000 student build plans. Students only have access to their personal repository without access other student repositories.

After the setup, Artemis allows the student to work in a local IDE or in the online editor. When the student submits a new solution, the build plan compiles the code and executes the tests defined by the instructor in a docker container. It uploads the results to Artemis, so that the students can immediately review the feedback and iteratively improve the solution. In case of an incorrect solution, the feedback shows a message for each failed test. The student can reattempt to solve the exercise and submit new solutions. The instructor can review results, gain insights, and react to errors and problems.

Artemis includes a web editor allowing inexperienced students to participate in exercises without dealing with complex version control and integrated development environments. It supports the manual review of submissions after the due date. Tutors can see the automatic feedback trough tests and static code analysis and enhance it with manual feedback. This makes it possible to review aspects difficult to assess automatically, e.g., the internal structure and specific code quality.

Artemis also features autograding functionality for text exercises [27] and modeling exercises [28] using a semi-automatic approach based on supervised machine learning. During the manual assessment, Artemis learns which model elements or text segments are correct or incorrect and applies this information including the qualititative feedback of the manual assessment to similiar model elements or text segments in other students' submissions. This approach allows multiple correct solutions by students and therfore does not limit their creativity. The knowledge increases over time and can be reused in subsequent years when exercises are reused.

IV. STUDY DESIGN

At the Technical University of Munich, students can evaluate the courses they have taken through anonymous feedback. Questionnaires are pre-defined and distributed every semester for every course by the student council. They contain open-ended text fields where students describe in own words aspects of the course they appreciated and which could be improved. In other questions students rate different aspects of the course on a Likert scale, e.g., lecture materials or homework exercises.

A. Course Profiles

In the scope of this study, we collected evaluation results from 8 courses held between 2016 and 2020. These belonged to 3 distinct modules of the study plan of informatics, i.e., some courses were repeating instances of the same module held in different years. In the following, we abbreviate each course denoting the module (A, O, P), followed by the year in which the course was held.

Modules A and O are lecture courses accompanied by tutor exercises and homework assignments. Programming tasks are a significant part of the homework in each course. However, good performance in homework assignments is not a central goal for the students, as homework solutions can only yield a grade bonus to improve the final exam grade. In contrast, the module P is a practical programming course. There, the student's grade is determined from individually submitted solutions to the programming assignments. Although P is commonly taken together with a introductory informatics lecture, the modules are graded separately.

B. Evaluation Analysis

The analysis of the questionnaire responses deals with two types of data: On free text replies, the topic distribution is

TABLE II
TOP WORDS FOR EACH TOPIC TRANSLATED TO ENGLISH. EACH COMMA-SEPARATED TERM REPRESENTS A SINGLE WORD IN THE ORIGINAL RESULTS. THE WORDS ARE SORTED IN DESCENDING ORDER BASED ON THE TOPIC-WORD MEMBERSHIP SCORE. ANONYMIZED TOP WORDS ARE DENOTED BY <>.

Topic label	Top words
Course implementation (IMP)	lecture, tutor exercise, hold, synchronize, content, find, work through, lesson, super, book
Sample solutions (SOL)	homework, sample solution, publish, solution, homework, build upon, helpful, on top of each other, exercise task
Tutoring (TUT)	materials, explain, tutor, understand, understanding, quick, task, detailed, discuss, help
Homework (HW)	task, week, practical, difficult, explanation, task description, learn, actually, <course P>, process
Learning programming (PRO)	learn, programming, practical, application, actually, <Artemis>, c, apply, practical course, java
Homework platforms (ONL)	<Artemis>, test, run, interactive, hand-in, interaction, work, platform, quiz, <Moodle>

shown for each course, and on Likert scale answers, the mean response rating is computed. We manually map each topic to a related question and plot both together in Section Results.

1) Topic Modeling on Comments: In all programming course evaluation questionnaires, participants write free-text answers to two questions: a) Which aspects of the course did you appreciate? (positive); b) What should be improved? (negative). With topic modeling [29], these comments are clustered into semantically coherent topics. This shows the themes of interest for the students.

We use non-negative matrix factorization (NMF), an established technique for topic modeling [30]. NMF takes a document-word matrix D as input, in which each row represents a document, each column a word, and each cell the term frequency-inverse document frequency (TF-IDF). It is a weighted count of word occurrences within a document, such that a word occurring in all documents gets a low score [31]. The matrix D is factorized into the document-topic matrix V and the topic-word matrix U – see Figure 1. The scores in V and U show how much a document or word is related to a topic. As we are working with few short texts, D becomes sparse, leading to semantically incoherent word lists for the extracted topics. Therefore, we include external knowledge about the words, i.e., their similarities according to external corpora, by using knowledge-guided NMF (KG-NMF) [32].

2) Likert Scale Questions: In the standardized evaluation, Likert scale questions can have one of three formats: a) Students are asked to state a grade between 1 "best" and 5 "worst"; b) students evaluate their agreement with a given statement ranging from 1 "fully agree" to 5 "completely disagree"; c) students can pick a position on a five-point scale between two extremes, e.g., the speed of a lecture could be 1 "too fast" and 5 "too slow", with 3 implying "just right".

The Likert scale answers to the following survey questions and statements are discussed in the Results:

- *"Which overall grade would you give this course?"*
- *"Which grade would you give your tutor?"*
- *"All in all, the homework sample solutions are helpful."*
- *"The difficulty level of the homework exercises is adequate."*
- *"I have learned to solve problems typical for the course's domain."*
- *"The online offering for the course is good."*

As exact phrasing of questions varied across years, synonymous questions are grouped together manually. Some questions are removed from the questionnaires by some lecturers, since the standardized university evaluation can be modified every year. The questionnaire for A.2020 especially focused on the challenges of digital study and omitted standard questions.

C. Addressing the Research Questions

The research questions in the Introduction are concerned with the impact of the Artemis autograder on the students' experience of programming courses. We analyze the impact by qualitatively comparing course evaluation statistics before and after introducing the tool. The statistics are a) the averaged Likert scale answers to the questions in subsection IV-B2, and b) corresponding topic distributions of comments about positive and negative aspects in the courses – see Table II.

Each research question is related to an evaluation question and topic, such that the answers and topic distributions together provide a meaningful answer. For a topic and a question about the same teaching aspect, we also depict if the topic appearance is reflected in the mean question response. If, e.g., the positive comments and the Likert answers increase over the years, we would conclude the regarding course aspect improved.

V. RESULTS

The results of the course evaluation questionnaires are shown as diagrams. Each consists a bar and a scatter plot. The scatter plot depicts the average of all numerical answers to the close-ended question of a course year. Values of the same course module over years are connected with lines. The last year of each module is highlighted, as this was the first time when Artemis was used in that course. To this year the p-value of a significance test is annotated[3]. The smaller the number, the less likely it is that both distributions are equal. As in similar studies [33], [34], [35], we accept p-values < 0.05 as statistically significant, which holds for all measured p-values.

The bar plots are obtained by summing up the comment-topic-membership scores over the positive or negative comments of the given topic. The highest-scoring top words of each topic are in Table II. We labeled each relevant topic manually based on top words and comments. The number of topics for NMF is 13, after evaluating the top word coherence manually for different topic counts. From these, 6 relevant topics are chosen. These are mapped to the most related questions and both are shown together in the diagrams and subsections.

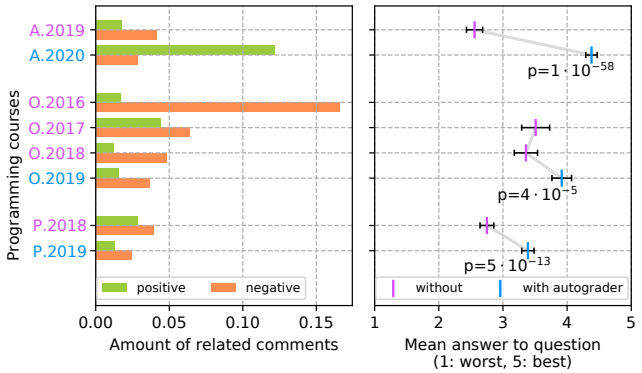


Fig. 2. Left: Number of positive/negative comments on topic IMP. Right: mean of numerical answers with black confidence intervals[5]; annotated p-values[3].

A. Course Quality

Question: *"Which overall grade would you give this course?"*
Topic: IMP collects general feedback about the lectures and tutor exercises held during the course. In the topic's comments, students frequently commented about the quality of the course and the themes as these are discussed in lectures and the related tutor exercises. The top words of the cluster highlight the connection between *lectures* and *tutor exercises*.
Results: Figure 2 shows that all courses received a higher rating after the introduction of Artemis. Simultaneously, either the negative comments about the course implementation receded or positive comments increased.

B. Tutoring

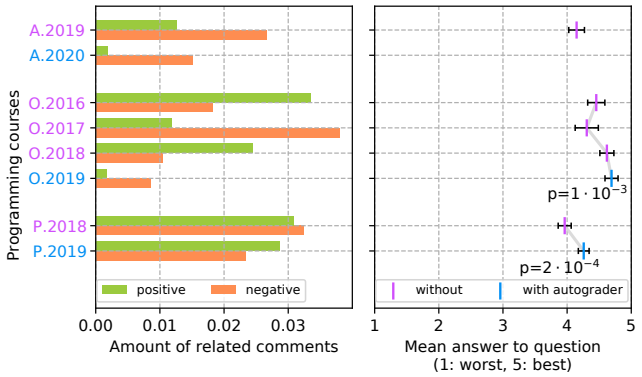


Fig. 3. Left: Number of positive/negative comments on topic TUT. Right: mean of numerical answers with black confidence intervals[5]; annotated p-values[3].

Question: *"Which grade would you give your tutor?"*
Topic: TUT's top words reflect the interactions between the students and tutors, where we find *explain, understand, detailed, discuss*, and *help*.
Results: The rating for the tutor was higher after the introduction of Artemis, see the courses O and P in Figure 3. The number of negative comments corresponds inversely with the grade students gave for their tutors. This shows that the tutors were perceived increasingly positively after Artemis was used. The course A.2020 questionnaire omitted this and other questions, such that no comparisons can be made for it.

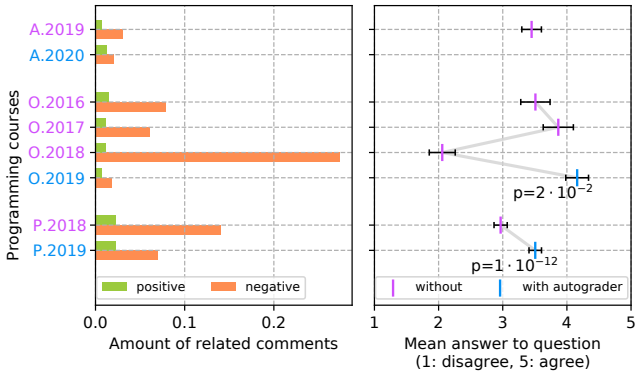


Fig. 4. Left: Number of positive/negative comments on topic SOL. Right: mean of numerical answers with black confidence intervals[5]; annotated p-values[3].

C. Sample Solutions

Question: *"All in all, the homework sample solutions are helpful."*
Topic: SOL collects feedback on *homework* exercises with an emphasis on the *publishing* of *sample solutions*, which *build upon* and *on top of each other*.
Results: Course O.2018 received many negative SOL comments and a low rating. According to the interviews with tutors in section VI-B2, students had difficulties with homework assignments built upon solutions of previous weeks, which were not published in time and thus hindered them from solving those assignments. Apart from that, sample solutions overall were rated more positively and commented less negatively after the Artemis intervention.

D. Artemis

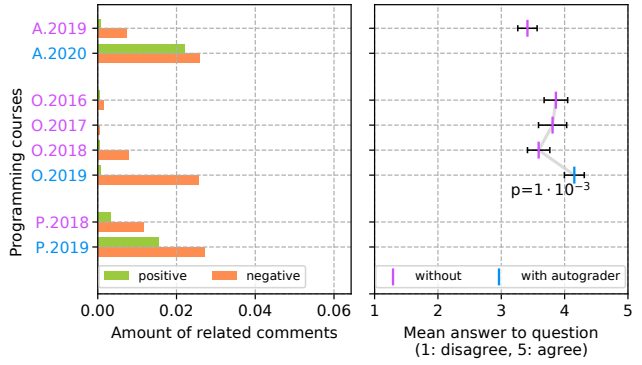


Fig. 5. Left: Number of positive/negative comments on topic ONL. Right: mean of numerical answers with black confidence intervals[5]; annotated p-values[3].

Question: *"The online offering for the course is good."*
Topic: ONL's top words regard to student interactions with online platforms (*Artemis, Moodle*) and their functionalities (*work, run, test, hand-in, quiz*).
Results: With the introduction of Artemis, the rating improved in O.2019 and the overall number of ONL comments increased. In the negative comments for Topic ONL, the students frequently provided bug reports or change requests for the newly introduced Artemis. As a consequence, the negative comments from the students do not depict strong negative sentiment, but constructive criticism, which can be seen on the Likert scale answers.

E. Homework Exercises

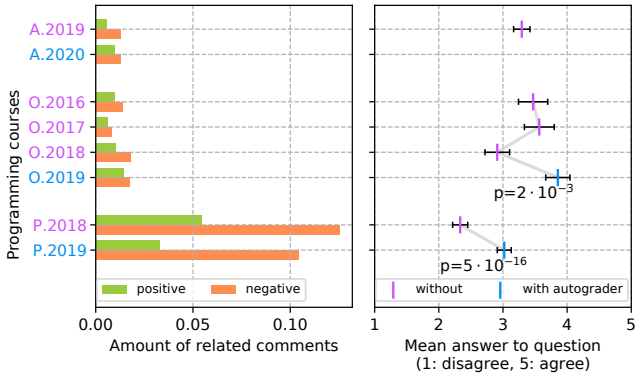


Fig. 6. Left: Number of positive/negative comments on topic HW. Right: mean of numerical answers with black confidence intervals[5]; annotated p-values[3].

Question: *"The difficulty level of the homework exercises is adequate."*
Topic: HW is about *weekly tasks* meant to help students *learn* course contents *practically*. In Figure 6, the topic appeared primarily in the P-courses, which are focused on extensive homework based on which the complete grade is calculated. In other courses, good homework performance gives a small bonus on the exam grade.
Results: Except for the denoted outlier O.2018, homework assignments are perceived as more adequate by each consecutive year in Figure 6. The negative comments of the homework topic are distributed similarly for the course series with sufficient data-points, i.e., O and P. However, the correspondence with positive comments is not clear for P.

F. Time Consumption of Homework

Question: *"On average, how many hours per week do you spend on homework assignments?"*
Results: As no topics and only incomplete data are available, we only provide the average numbers for the course O as follows. The average time consumption of homework exercises remained constant at 3 hours untill 2017, increased to 6 hours in 2018 due to subsection V-C, and decreased to 4 hours in 2019. This means that in 2019, when Artemis was introduced, the reported time spent on homework was higher than in 2016 and 2017. Concurrently, there was an increase in the course satisfaction (Figure 4), the adequateness of the homework tasks (Figure 6), and the learning outcome (Figure 7). Thus, it cannot be excluded that Artemis led to higher motivation and increased time spent on homework. However, the data is only available for one course series, and further research should investigate if other courses also show increased time for solving homework assignments, especially in the long run.

[5]These are confidence intervals with a 95% chance of containing the true population mean answers. We have $n > 100$ samples (see Table I) and assume a normal distribution for the means due to the central limit theorem.

[3]Null hypothesis: Both distributions are equal. Alternative hypothesis: The distribution from the year with autograder is higher than from the years before. A one-sided Wilcoxon rank-sum test [36] is applied, since the distributions are not normally distributed, and the years with autograder have higher ratings. For the sample sizes see Table I.

G. Learning Programming

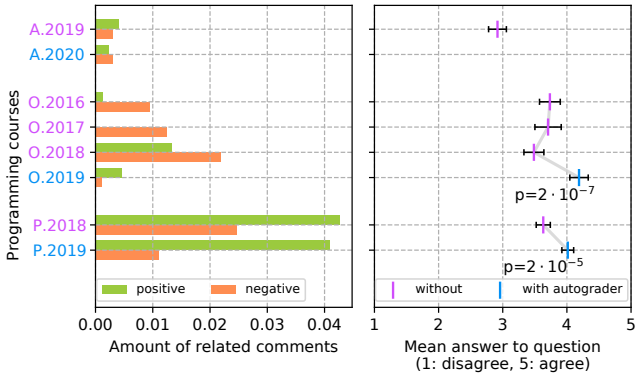


Fig. 7. Left: Number of positive/negative comments on topic PRO. Right: mean of numerical answers with black confidence intervals[5]; annotated p-values[3].

Question: *"I have learned to solve problems typical for the course's domain."*
Topic: PRO collects feedback on the development of programming skills. Its top words relate to *programming*, programming languages (*C, Java*), *applying practical* skills, and *application* development. *Artemis* is also among them, indicating a connection with learning programming in students' feedback.
Results: The students' perceived programming learning success was rated highest for both courses P and O in the year when Artemis was introduced. Similarly and at the same time, the number of negative comments in Figure 7 decreased much more than the positive ones. The practical course P, where students are mainly focused on solving programming tasks, shows the highest proportion of comments in Figure 7.

H. Further Observations

In addition to the results shown above, further data shows positive trends in years when Artemis was introduced. Where comparison data is available, we observe the students' average numerical response improving for any other question related to tutor exercises, homework, or student learning progress. Such questions and prompts include:

- *The difficulty of tutor exercise content is ...*
- *I can now provide an overview about the topic of the course.*
- *I can now explain important concepts from the course.*
- *How well were you able to acquire the required knowledge using the provided materials?*

In all these cases, we observe that modules have a higher numeric score for a course with Artemis compared to all other scores within the module for the same question.

VI. Findings

A. Answers to Research Questions

The research questions from the Introduction regarding how the course evaluation surveys changed after the introduction of Artemis can be answered as follows.

1) RQ1 How did students report on their learning experience in course evaluations, and how did it change?: The students' perception of their ability to solve domain-typical programming problems increased in all courses with according data (V-G), while they spent more time on the exercises (Section V-F), potentially because of improved reported learning experiences (Section V-H).

2) RQ2 How did the reported interaction between students and tutors change?: Section V-B shows tutors and their competences are perceived more positively by students. Supporting arguments are given by interviews we conducted with tutors and by the related work – see the section VI-B2.

3) RQ3 How did the perceived difficulty of the practical programming parts of the courses change?: From Section V-E, it can be concluded that the difficulty of the programming exercises overall was seen as more adequate by the course participants than it was before after Artemis was introduced.

4) RQ4 How did the perceived overall course quality change?: The overall rating of all courses increased and the number of critical comments decreased after the introduction of Artemis as outlined Section V-A. Potential reasons include the previously reported finding, i.e., improved interactions between tutors and students, improved overall course quality, and improved learning success.

B. Hypotheses About the Impact of Artemis

In the following paragraphs, we connect the above findings with several other factors, including the Artemis autograder. This is supposed to inform hypotheses regarding how autograding can contribute to improving student satisfaction in programming courses.

1) Higher Motivation for Homework: After the introduction of Artemis, students reported that they spent more time on homework exercises, even though they perceived the exercises to be less difficult compared to students who did not have an autograder. It appears that the learning experience is more rewarding when homework is solved the homework. Accordingly, the course satisfaction and learning outcome is increased, too. This can be observed across all courses after the introduction of Artemis.

In this context, it is worth mentioning that unit tests are provided on Artemis, which are triggered when uploading the homework exercises. The output of the tests become immediately visible to the student. This motivates the student to upload an improved version of the code before the deadline. So, although it seems the student spends more time solving the homework, the feedback loop keeps motivation high. This is confirmed by the literature [18] and some of the evaluation comments.

2) Better Tutoring Sessions: In Section Tutoring we state *"the rating for the tutors was consistently higher after the introduction of Artemis"*. To explain this observation, we interviewed tutors of the course P from both years about the advantages of Artemis during their teaching sessions. They stated or confirmed the following statement:

> *"What is very helpful, however, are the automatic tests for the in-class exercises. This means that students do not have to test their code so extensively themselves (especially if it works). In the past, they often needed help or support with these tests. As a result, I have more time for the people who really need help."*

Similar feedback was reported in another paper:

> *"Finally, this methodology allows the professor to make better use of her time. Thanks to the use of the automated tool, her time is mostly spent with the students who have not been able to obtain correct answers; the rest of the students already know their solutions are correct, and they have moved on to more challenging problems." [20]*

With the integration of Artemis, more advanced students can work more independently and thus progress faster. They need less feedback – especially at the beginning to get going – as this can be given by automatic tests. Consequentially, tutors can devote their limited time to help the less advanced students. We conclude that the usage of an autograding system can have a positive impact on the teaching quality in tutor sessions. This is a possible explanation why tutors who used an autograder received a more positive evaluation.

3) Fairer Homework Corrections: Artemis assigns homework submissions from students randomly to tutors for correction. The process is anonymized and the students do not know which tutor corrected their submissions. This feature is called double-blind homework correction and might have impacted the evaluations. It was a new practice for all given courses, as previously one tutor was always responsible for the homework correction of one fixed group of students throughout the whole semester. This tutor was also known to them as their responsible supervisor. With the new approach, the overall homework points should depend less on the bias of one single tutor. This can reduce unfair treatment in the homework corrections and impact how students evaluate tutors, practical programming parts, or the overall course implementation.

C. Limitations

The data of this study is not gathered under controlled conditions. This means that in addition to Artemis, there are other factors that impact the student satisfaction. These factors have different degrees of impact and can change from year to year. In the following, we list factors that can impact the satisfaction of users to a high degree. However, it should be kept in mind that all teaching courses we take into account in this study are foundational computer science courses. This means that their content and organization do not significantly change.

The year 2018 of the course O was evaluated worse than all other years in several categories, for example, clarity, amount, and speed of content (the last two are not discussed in the results). As mentioned in section V-C, this was caused by an organizational problem with homework assignments. One open question is how this damaged the overall course impression

of the students, or if not, which other factors played a role in the ratings. We address the problem by omitting evaluations from the course O, which significantly deviate negatively for each Likert question when significance tests are calculated. Consequently, the mean answers for O with and without Artemis are more similar but always statistically significantly different with all p-values < 0.05.

Even though a large sample size of more than 100 evaluation participants (see Table I) should give robustness against single-student bias, sympathetic or otherwise highly regarded professors or tutors might have increased positive valuation from students even for unrelated questions [37], especially as the professors changed in every course from year to year.

The introduction of Artemis frequently coincides with better numerical ratings from students in the presented lectures. A causal link, however, cannot be established without taking a look at lectures or tutor sessions from the same years without Artemis. This would rule out the possibility that the courses are improving each year due to unrelated reasons.

Finally, courses with Artemis should be investigated for selection bias. Since changes to the technical solution of a course require additional effort, it is possible that only course organizers putting more effort towards the course would consider transitioning to a new LMS. Positive factors that arise from motivated course organizers (e.g., better study materials, quicker answering to student questions) might cause higher scores in student evaluations unrelated to Artemis.

VII. Conclusion

In this paper, significant changes in the evaluations of several foundational computer science course series with numerous participants (> 1000) are depicted after introducing an autograding software. Regarding most research questions, consistent improvements from the perspective of the course participants are observed. These include improved interactions between tutors and students, improved course quality, improved learning success, increased time spent, and reduced difficulty. As possible reasons, we identify helpful automated feedback from unit tests, fairer and more objective grading, reduced correction bias, enhanced course implementation, and more available time for tutors to focus on students' needs during teaching sessions.

The analysis is based on a qualitative interpretation of statistics and topic modeling. While the measures provide mostly coherent results, the conclusions need to be confirmed by applying thorough quantitative correlation measures, e.g., as conducted in other studies [9]. Furthermore, we aim to include sentiment analysis, as the sentiment dimension differs from the positive vs. improvable categories given in the questionnaires used in this study. Regarding the data, we want to include more data to see if the changes of Artemis are sustainable after more than one year.

References

[1] S. H. Edwards and M. A. Perez-Quinones, "Web-cat: Automatically grading programming assignments," in *Proceedings of the 13th Annual Conference on Innovation and Technology in Computer Science Education*, ser. ITiCSE '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 328. [Online]. Available: https://doi.org/10.1145/1384271.1384371

[2] M. Goedicke, "10 jahre automatische bewertung von programmieraufgaben mit JACK - rückblick und ausblick," *INFORMATIK 2017*, vol. P-275, pp. 279–283, 2017.

[3] J. Krinke, M. Störzer, and A. Zeller, "Web-basierte programmierpraktika mit praktomat," *Softwaretechnik-Trends*, vol. 22, no. 3, pp. 51–53, 2002.

[4] A. Stöcker, S. Becker, R. Garmann, F. Heine, C. Kleiner, and O. J. Bott, "Evaluation automatisierter programmbewertung bei der vermittlung der sprachen java und sql mit den gradern "asqlg" und "graja" aus studentischer perspektive," in *DeLFI 2013: Die 11 e-Learning Fachtagung Informatik*, A. Breiter and C. Rensing, Eds. Bonn: Gesellschaft für Informatik e.V., 2013, pp. 233–238.

[5] S. Krusche and A. Seitz, "Artemis: An automatic assessment management system for interactive learning," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 284–289. [Online]. Available: https://doi.org/10.1145/3159450.3159602

[6] B. Otto, T. Massing, N. Schwinning, N. Reckmann, A. Blasberg, S. Schumann, C. Hanck, and M. Goedicke, "Evaluation einer statistik-lehrveranstaltung mit dem jack r-modul," in *Bildungsräume 2017*, C. Igel, C. Ullrich, and W. Martin, Eds. Bonn: Gesellschaft für Informatik, 2017, pp. 75–86.

[7] N. Grönberg, "Palaute : an online tool for text mining course feedback using topic modeling and emotion analysis," Master's thesis, LUT University, School of Engineering Science, Computer Science, 2020.

[8] Z. Nasim, Q. Rajput, and S. Haider, "Sentiment analysis of student feedback using machine learning and lexicon based approaches," in *2017 International Conference on Research and Innovation in Information Systems (ICRIIS)*. IEEE, July 2017, pp. 1–6.

[9] M. Hujala, A. Knutas, T. Hynninen, and H. Arminen, "Improving the quality of teaching by utilising written student feedback: A streamlined process," *Computers & Education*, vol. 157, p. 103965, 2020.

[10] S. Gottipati, V. Shankararaman, and S. Gan, "A conceptual framework for analyzing students' feedback," in *2017 IEEE Frontiers in Education Conference (FIE)*. IEEE Computer Society, Oct 2017, pp. 1–8.

[11] S. Srinivas and S. Rajendran, "Topic-based knowledge mining of online student reviews for strategic planning in universities," *Computers & Industrial Engineering*, vol. 128, pp. 974–984, 2019.

[12] N. Grönberg, A. Knutas, T. Hynninen, and M. Hujala, "An online tool for analyzing written student feedback," in *Koli Calling '20: Proceedings of the 20th Koli Calling International Conference on Computing Education Research*, ser. Koli Calling '20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3428029.3428565

[13] G. I. Nitin, S. Gottipati, and V. Shankararaman, "Analyzing educational comments for topics and sentiments: A text analytics approach." in *2015 IEEE Frontiers in Education Conference (FIE)*. IEEE Computer Society, 2015, pp. 1–9. [Online]. Available: https://ieeexplore.ieee.org/xpl/conhome/7344010/proceeding

[14] S. Cunningham-Nelson, M. Baktashmotlagh, and W. Boles, "Visualizing student opinion through text analysis," *IEEE Transactions on Education*, vol. 62, no. 4, pp. 305–311, 2019.

[15] S. Cunningham-Nelson, M. Laundon, and A. Cathcart, "Beyond satisfaction scores: visualising student comments for whole-of-course evaluation," *Assessment & Evaluation in Higher Education*, vol. 0, no. 0, pp. 1–16, 2020.

[16] T. Hynninen, A. Knutas, M. Hujala, and H. Arminen, "Distinguishing the themes emerging from masses of open student feedback," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, May 2019, pp. 557–561.

[17] I. Vocareum, "Vocareum - cloud learning labs," Aug 2021. [Online]. Available: https://www.vocareum.com/#classroom

[18] H. Keuning, J. Jeuring, and B. Heeren, "A systematic literature review of automated feedback generation for programming exercises," *ACM Transactions on Computing Education (TOCE)*, vol. 19, no. 1, pp. 1–43, 2018.

[19] Y. Paredes, P. Huang, H. Murphy, and I. Hsiao, "A subjective evaluation of web-based programming grading assistant: Harnessing digital footprints from paper-based assessments," *CEUR Workshop Proceedings*, vol. 1828, pp. 23–30, 2017.

[20] R. Landa and Y. Martinez-Treviño, "Relevance of immediate feedback in an introduction to programming course," in *2019 ASEE Annual Conference & Exposition*, no. 10.18260/1-2–33235. Tampa, Florida: ASEE Conferences, June 2019, https://peer.asee.org/33235.

[21] A. B. Goldman, "Using daily missions to promote incremental progress on programming assignments," Master's thesis, Virginia Polytechnic Institute and State University, 2019.

[22] M. Goedicke, M. Striewe, and M. Balz, "Computer aided assessments and programming exercises with jack," ICB-Research Report, Essen, Tech. Rep. 28, 2008.

[23] S. Krusche, N. von Frankenberg, and S. Afifi, "Experiences of a Software Engineering Course based on Interactive Learning," in *Tagungsband des 15. Workshops Software Engineering im Unterricht der Hochschulen (SEUH)*. CEUR, 2017, pp. 32–40.

[24] S. Krusche, A. Seitz, J. Börstler, and B. Bruegge, "Interactive Learning: Increasing Student Participation Through Shorter Exercise Cycles," in *Proceedings of the 19th Australasian Computing Education Conference*. ACM, 2017, pp. 17–26.

[25] S. Krusche, N. von Frankenberg, L. M. Reimer, and B. Bruegge, "An Interactive Learning Method to Engage Students in Modeling," in *Proceedings of the 42nd International Conference on Software Engineering: Software Engineering Education and Training*. ACM, 2020, pp. 12–22.

[26] S. Krusche and A. Seitz, "Increasing the Interactivity in Software Engineering MOOCs - A Case Study," in *52nd Hawaii International Conference on System Sciences*, 2019, pp. 1–10.

[27] J. P. Bernius, S. Krusche, and B. Bruegge, "A Machine Learning Approach for Suggesting Feedback in Textual Exercises in Large Courses," in *Proceedings of the 8th Conference on Learning at Scale*. ACM, 2021.

[28] S. Krusche, "Semi-Automatic Assessment of Modeling Exercises using Supervised Machine Learning," in *55nd Hawaii International Conference on System Sciences*, 2022.

[29] W. H. Finch, M. E. Hernández Finch, C. E. McIntosh, and C. Braun, "The use of topic modeling with latent dirichlet analysis with open-ended survey items." *Translational Issues in Psychological Science*, vol. 4, no. 4, p. 403, 2018.

[30] A. Cichocki and A.-H. Phan, "Fast local algorithms for large scale nonnegative matrix and tensor factorizations," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 92, no. 3, pp. 708–721, 2009.

[31] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval," Cornell University, Ithaca, NY, USA, Publication, 1987.

[32] Y. Chen, H. Zhang, R. Liu, Z. Ye, and J. Lin, "Experimental explorations on short text topic mining between lda and nmf based schemes," *Knowledge-Based Systems*, vol. 163, pp. 1–13, 2019.

[33] R. H. Maki, W. S. Maki, M. Patterson, and P. D. Whittaker, "Evaluation of a web-based introductory psychology course: I. learning and satisfaction in on-line versus lecture courses," *Behavior research methods, instruments, & computers*, vol. 32, no. 2, pp. 230–239, 2000.

[34] D. A. Lake, "Student performance and perceptions of a lecture-based course compared with the same course utilizing group discussion," *Physical therapy*, vol. 81, no. 3, pp. 896–902, 2001.

[35] K. A. Hoag, J. K. Lillie, and R. B. Hoppe, "Piloting case-based instruction in a didactic clinical immunology course," *American Society for Clinical Laboratory Science*, vol. 18, no. 4, pp. 213–220, 2005.

[36] F. Wilcoxon, *Individual Comparisons by Ranking Methods*. New York, NY: Springer New York, 1992, pp. 196–202. [Online]. Available: https://doi.org/10.1007/978-1-4612-4380-9_16

[37] D. E. Clayson and M. J. Sheffet, "Personality and the student evaluation of teaching," *Journal of Marketing Education*, vol. 28, no. 2, pp. 149–160, 2006.

Part III

Appendix: Publications Not Relevant for Examination

A Dataset: Social Media Discourse About Organic Food

A.1 SocialVisTUM: An Interactive Visualization Toolkit for Correlated Neural Topic Models on Social Media Opinion Mining

The publication on the consecutive pages was accepted after peer-review as demonstration paper at the 2021 International Conference on Recent Advances in Natural Language Processing. Gerhard Johann Hagerer, the author of the present thesis, is the first author of that paper. His author role, the reprinting permission, and his following contributions are acknowledged by all authors Hagerer, Kirchhoff, Danner, Ghosh, Roy, Zhao, and Groh [2021a]:

> *"Gerhard Johann Hagerer headed the research project. He developed the original research idea, the concept, and the primary methodology of the paper. Furthermore, he directed the implementation process over multiple stages and reviewed the source code deeply. Regarding the writing of the paper, he created the outline, directed the drafting, wrote large textual parts, incorporated extensive reviewer feedback, and paraphrased, corrected, combined, and otherwise improved drafted material."*

Publication Summary

The following publication is a demonstration for how state-of-the-art (SOTA) unsupervised opinion mining methods relate to qualitative content studies. It is found to be possible to automatically reproduce and visualize findings from previous qualitative market research study on social media texts. The method is based on recent advancements of neural aspect extraction, and we designed it such that it is applicable for domain experts without technical expertise. For instance, it is simple to use and implement the technology, which comes with automatic hyperparameter estimation and sentiment analysis included. Further features are graph-based visualizations, which enable to explore correlated neural topic models, such that similar and related topics are clustered together. The topic distributions and topic-related sentiments in the social media corpus are visualized and give an explanation about the content, opinions, and attitudes in unstructured quantities of social texts.

SocialVisTUM: An Interactive Visualization Toolkit for Correlated Neural Topic Models on Social Media Opinion Mining

Gerhard Hagerer[2], **Martin Kirchhoff**[3], **Hannah Danner**[2], **Robert Pesch**[3]
Mainak Ghosh[1], **Archishman Roy**[2], **Jiaxi Zhao**[2], **Georg Groh**[2]
[1]Max Planck Institute for Innovation and Competition
[2]Technical University of Munich
[3]inovex GmbH
{gerhard.hagerer,hannah.danner,archishman.roy,jiaxi.zhao}@tum.de,grohg@mytum.de
{martin.kirchhoff,robert.pesch}@inovex.de,mainak.ghosh@ip.mpg.de

Abstract

Recent research in opinion mining proposed word embedding-based topic modeling methods that provide superior coherence compared to traditional topic modeling. In this paper, we demonstrate how these methods can be used to display correlated topic models on social media texts using SocialVisTUM, our proposed interactive visualization toolkit. It displays a graph with topics as nodes and their correlations as edges. Further details are displayed interactively to support the exploration of large text collections, e.g., representative words and sentences of topics, topic and sentiment distributions, hierarchical topic clustering, and customizable, predefined topic labels. The toolkit optimizes automatically on custom data for optimal coherence. We show a working instance of the toolkit on data crawled from English social media discussions about organic food consumption. The visualization confirms findings of a qualitative consumer research study. SocialVisTUM and its training procedures are accessible online[1].

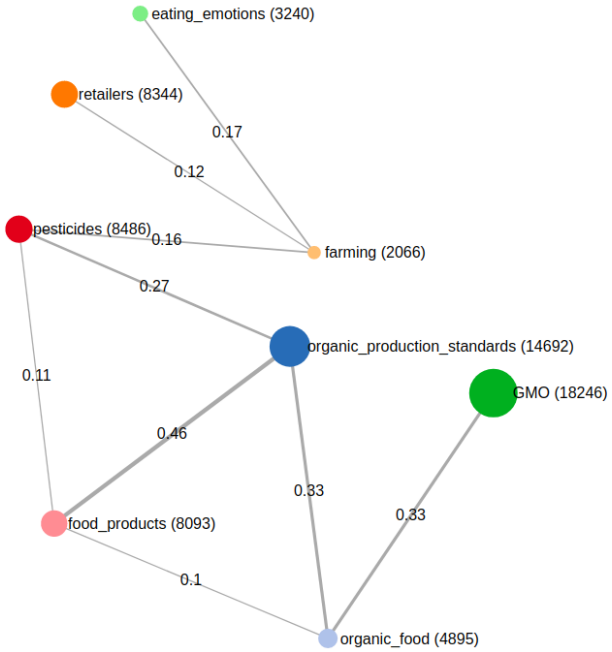


Figure 1: SocialVisTUM applied to our use case *organic food* - The topics, their occurrence (in brackets) and respective correlations.

1 Introduction

Web sources, such as social networks, internet forums, and customer reviews from online shops, provide large amounts of unstructured text data. Along with the steady development of new platforms and the increasing number of internet users, the interest in methods that automatically extract the expressed opinions along with the corresponding topics and sentiments in text data has increased in recent years. Scholars and organizations from different fields can utilize such methods to identify patterns and generate new insights. Examples are opinion researchers investigating current opinions on political and societal issues, consumer researchers interested in consumers' beliefs about the consumption and production of goods (Danner et al., 2020), and marketing managers curious about the public perception of their products and services (Berger et al., 2020; Murphy et al., 2014). (Kirchhoff, 2019)

These domain-specific use cases are of interest for research disciplines which taken by itself are not directly related to natural language processing (NLP). Consequentially, there is a constant need to provide state-of-the-art NLP methods such that domain researchers from other fields can take advantage of them. The requirements therefore are simple usage, automatic hyperparameter optimization, minimal effort for manual labeling of text data, and built-in visualizations to give an abstract overview of the discussed topics and their relation

[1] https://github.com/ghagerer/SocialVisTum

with each other.

While these practical requirements are important for domain experts, modern opinion mining approaches target specific machine learning objectives. Recently, there is a trend towards unsupervised neural methods for opinion target detection. Attention-based aspect extraction (ABAE) enables clustering of short review texts with significantly higher coherence as traditional LDA-based topic modeling, and it gives 70% F1 score for classification (He et al., 2017). This is improved recently (Karamanolakis et al., 2019; Angelidis and Lapata, 2018; Luo et al., 2019), which underlines the recent impact and potential of related techniques.

However, these have not been utilized for visualizations based on correlated topic modeling (Blei and Lafferty, 2006), where all pairs of topics "are" analyzed to determine if two topics generally tend to occur in the same texts of a given dataset. Thus, the similarity between topics can be defined. This is successfully used to connect topics (nodes) among each other based on their correlations (edges) leading to more abstract and more meaningful meta topics (graph-clusters) which additionally improves topic coherence. Consequentially, these meta topics, e.g., company-related events or research subdisciplines (Liu et al., 2014; Maiya and Rolfe, 2014), can be successfully identified by graph-based visualization techniques. However, there is a lack of related prototypes on texts discussing consumption related issues in product reviews or social media. To the best of our knowledge, there is also no related integration of sentiment analysis into a system available for potential end users, i.e., domain experts. As according text data from customers is available on a large scale in social media, this can be considered as a shortcoming in the field.

To address all denoted issues, we propose the *SocialVisTUM* toolkit, a new visualization and labeling tool to give users a comprehensible overview of the topics discussed in social media texts. It integrates a neural method for unsupervised sentence and comment clustering based on word vectors and attention. We denote the respective clusters as topics hereafter. In addition, we provide a graph-based visualization showing the topics as labeled nodes and the correlation between them as edges. A force-directed graph layout maintains readability even while many relevant topics and topic relations are displayed. (Kirchhoff, 2019)

In our interactive graphical user interface, the number of topics displayed and the correlation threshold required to display a connection between two topics can be dynamically adjusted. Further, contextual topic information is provided, such as the number of respective topic occurrences in the social media texts as node diameter, the correlation between the topic occurrences as edge width, example sentences from the data for each topic, a list of representative words for each topic, and the regarding sentiment distribution of a topic. It is a common practice to represent topics merely by word lists (Blei et al., 2003; Chen et al., 2014), which tend to be insufficient to comprehensively express a topic on our given dataset. (Kirchhoff, 2019)

To avoid manual labeling and to give users an immediate impression of each topic, topic labels are generated automatically based on a custom algorithm utilizing the most common WordNet hypernym in a topic's top words. Furthermore, we find that topic hypernym statistics can serve as a metric for automatic hyperparameter optimization, which in our case gives practical advantages over widely used coherence scoring metrics.

In addition to a more detailed description of our SocialVisTUM toolkit, we show the results of a case study based on social media texts from online commenters debating about organic food consumption. We demonstrate that the correlated topics give a meaningful graph representation of the social media discussions supporting the understanding of the concerns of consumers. In this regard, we also show how the combined illustration of different types of relevant topic and sentiment information and automatic labeling of clusters are a contribution.

2 Related Work

Correlated topic models were introduced 2006 (Blei and Lafferty, 2006; Li and McCallum, 2006) to improve topic coherence and to provide graph visualizations based on topics as nodes and their correlations as edges. This shows potential to improve text mining for the end user as "powerful means of exploring, characterizing, and summarizing large collections of unstructured text documents" (Maiya and Rolfe, 2014). Meta topics, such as research domains and their inter-disciplinary overlaps, can thus be described clearly, automatically, and empirically (Blei and Lafferty, 2007).

These correlated topic models are applied for

more sophisticated visualization approaches. *TopicPanorama* models technology-related topics from various text corpora, including newspaper articles, blogs, and tweets (Liu et al., 2014). Here, the domain expert is given the option to interactively modify the matching result of the labeled topic graph. Another topic visualization called *topic similarity networks* is particularly addressing the visualization of large document sets (Maiya and Rolfe, 2014). While claiming good scalability regarding the number of documents, beneficial methods to achieve automatic topic labeling are successfully quantified. *TopicAtlas* provides a graphical user interface to explore text networks, such as hyperlinked webpages and academic citation networks. For manual mining purposes, topic models are generated and related to one another to facilitate manual navigation and finding of relevant documents (He et al., 2016). These examples show a steady, meaningful, and promising development regarding the visualization of correlated topic modeling, partially also applied to social media texts such as micro-blogs. However, these examples do not include sentiment analysis as means to conduct market research and quantify customer satisfaction in specific and not yet explored market domains. Furthermore, the widely used latent Dirichlet allocation (LDA) technique tends to be incoherent on short texts, such as, product reviews or social media comments, and thus insufficient to detect opinion targets in an unsupervised manner (He et al., 2017).

Automatic topic coherence optimization can be seen as desireable for a topic modeling visualization toolkit such as SocialVisTUM, which tries to minimize manual optimization efforts for non-technical users. Therefore, we refer to two widely used coherence definitions (Ghosh, 2020). Firstly, word co-occurrence-based methods measure how often pairs of representative topic words co-occur in the training data set or in an external reference data set. In that regard, it has been shown that the evaluation methods UMass, UCI and NPMI correlate with human judges (Stevens et al., 2012; Newman et al., 2010; Mimno et al., 2011; Bouma, 2009; Ding et al., 2018) and are considered to be a default metric for topic coherence. Secondly, word embedding similarity based coherence scores are recently utilized as these are also based on word co-occurrence statistics (Pennington et al., 2014) and behave similar to NPMI coherence scoring (Ding et al., 2018), resulting in high correlation with human perception. These methodologies show the undesirable effect of no distinct local optimum when the hyperparameters of the models are changed, e.g., number of clusters or vocabulary size. On our data, these parameters increase together with the coherence scores, while the subjective performance, i.e., the human perception, actually decreases. We describe this effect and our solution in the case study section.

3 Clustering Architecture

The unsupervised neural network model called attention-based aspect extraction (ABAE) (He et al., 2017) clusters sentences based on GloVe word embeddings (Pennington et al., 2014) and attention (Bahdanau et al., 2014) to focus on the most important words in a sentence. Every sentence s is represented by a vector z_s that is defined as the weighted average of all the word vectors of that sentence. The weights are attentions calculated based on the contribution of the respective words to the meaning of the sentence and the relevance to the topics. These topics are defined by the actual centroids. In their publication, the topics are mapped to aspect classes for unsupervised aspect extraction, which we do not do for our case. (Kirchhoff, 2019)

The topics are initialized as the resulting centroids of k-means clustering on the word embeddings of the corpus dictionary. These are then stacked as topic embedding matrix $\mathbf{T}$. During training, ABAE calculates sentence reconstructions $\mathbf{r}_s$ for each sentence. These are linear combinations of the topic embeddings from $\mathbf{T}$ and defined as

$$\mathbf{r}_s = \mathbf{T}^\top \cdot \mathbf{p}_t, \tag{1}$$

where $\mathbf{p}_t$ is the weight vector over K topic embeddings. Each weight corresponds to the probability that the input sentence belongs to the associated topic. $\mathbf{p}_t$ is obtained by reducing the dimension of $\mathbf{z}_s$ to the number of topics K and applying softmax such that

$$\mathbf{p}_t = softmax\left(\mathbf{W} \cdot \mathbf{z}_s + \mathbf{b}\right), \tag{2}$$

where $\mathbf{W}$ and $\mathbf{b}$ are trainable and matrix weights and a bias vector respectively. The topic embeddings T are updated during training to minimize the reconstruction error $J(\theta)$ between $\mathbf{r}_s$ and $\mathbf{z}_s$ based on the contrastive max-margin objective function. Since words and topics share the same dimensionality, cosine similarity between both can be used to

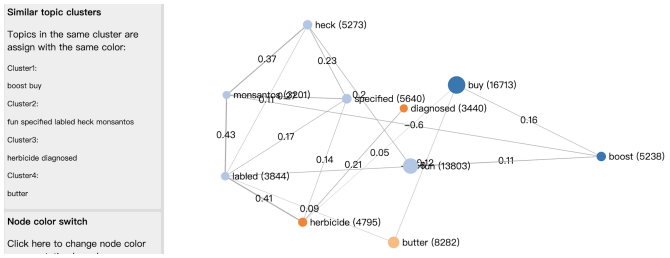


Figure 2: Highly correlated topics are colored by the same color respectively.

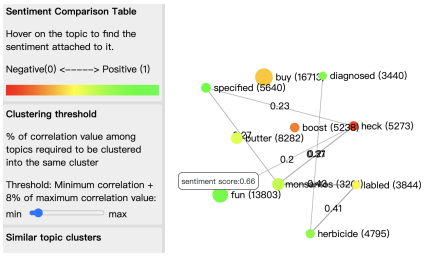


Figure 3: The sentiment for each topic is shown as green (positive) and red (negative).

look up the most similar words representing each topic, similar to the way LDA (Blei et al., 2003) represents topics as word distributions. (Kirchhoff, 2019)

4 The SocialVisTUM Toolkit

Visualization Figure 2 shows an example of the visualization and labeling tool. Topics are represented as nodes with according labels, and the number of texts assigned to the topics is given in parentheses next to the label. The node size increases based on the number of *topic occurrences*. The edges of the topic connections are labeled by the *topic correlations*. The link thickness increases with a higher positive correlation. A graph layout based on repelling forces between nodes helps to avoid overlaps, which is especially useful when many nodes and links are displayed. A second force keeps the graph centered. Users can also move nodes around to get a more comprehensible overview. (Kirchhoff, 2019)

Topic Nodes and Correlations After training the ABAE model, the sentences are assigned to topics based on the maximum topic probability from $\mathbf{p}_t$, see formula 2. The correlation between two topics i and j is calculated based on the probabilities $(\mathbf{p}_t)_i$ and $(\mathbf{p}_t)_j$ of each given sentence t. This yields a value in the range of $[-1, 1]$ for every pair of topics specifying the strength of the corresponding relatedness. (Kirchhoff, 2019)

Hiding Insignificant Topics An occurrence threshold slider defines the percentage of sentences that must be about a topic to display the associated node. Another slider can be used to set the correlation threshold to define the required positive or/and negative correlation to display the associated connections. These sliders are especially helpful to maintain a clear visualization by limiting the number of shown topics and connections when there are many of them available.

Topic Inspection Users can double-click a node to receive additional information about a topic, i.e., representative words and sentences, as shown in Figure 5 on the left and right respectively. As representative words, the top 10 words are shown sorted by the distance of their embeddings to the selected topic centroid in ascending order. The representative sentences are the ones with the highest probability from p_t for the given topic. During topic inspection mode, only nodes that are connected to the clicked node and the associated links are displayed. A double click on the same node brings back the whole graph again.

Colorization of Topic Nodes In an updated version of SocialVisTUM, we introduce two meaningful colorings of the topic nodes for correlated topic clustering and sentiment analysis.

Firstly, we perform a hierarchical clustering algorithm such that those topics which are strongly correlated with each other are colorized in one and the same color respectively. A dynamic slider GUI element helps to adjust the correlation threshold accordingly. One example outcome is shown in 2.

Secondly, we perform sentiment analysis using the Valance Aware Dictionary and sEntiment Reasoner or VADER method (Gilbert, 2014). It is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media settings. It gives positive, negative, and neutral scores to represent the proportion of text in that sentence that falls in these categories. For each sentence, we use the compound score, i.e., the sum of all lexicon ratings normalized between -1 (most negative) and +1 (most positive). We then calculate the average sentiment score for each topic based on all respective topic sentences. In Figure 3, positive sentiment is shown as green topic nodes, and negative as red. (Roy and Zhao, 2020)

Topic Label	Representative words	# Hypernyms
animal (102)	insect, ant, habitat, rodent, herbivore	218
compound (91)	amino, enzyme, metabolism, potassium, molecule	158
chemical (74)	fungicide, insecticide, weedkiller, preservative, bpa	131
systematically (0)	systematically, adequately, cleaned, properly, milked	0

Table 1: Example topics, automatically assigned topic labels, and representative words. The value next to the topic label denotes how often the label occurs as a shared hypernym. The number of hypernyms on the right tells in how many word comparisons a shared hypernym is identified. Taken from (Kirchhoff, 2019).

Automatic Topic Labels We introduce an approach to label topic nodes automatically. It is based on *shared hypernyms*, i.e., the lowest common denominator for words, which we identify using the representative topic words denoted in the previous paragraph and the lexical database WordNet (Miller, 1995). First, we retrieve the hypernym hierarchy for every representative topic word, as shown in Figure 4, and compare every word with every other word in the word list. Next, at each comparison, we save the hypernym with the lowest distance to the compared words in the hypernym hierarchy. We denote these as *shared hypernyms*. We only consider hypernyms if their distance to the word is smaller than half of the distance to the root hypernym to avoid unspecific labels like *entity* and *abstraction*. Eventually, we employ the hypernym that occurs most often as topic labels. If no hypernym can be identified, we use the most representative word instead. In the example shown in figure 4, we identify *dairy_product* as the lowest shared hypernym of *yoghurt* and *butter*, and *food* as lowest shared hypernym of *yoghurt* and *bread*. (Kirchhoff, 2019)

The quality of a shared hypernym chosen as topic label can be approximated by inspecting the number of its hypernym occurrences – see table 1. Topic labels occurring frequently as shared hypernym are usually suitable (e.g., animal (102) and compound (91)) in contrast to topic labels that rarely occur (e.g, group_action (9) or smuckers (0)). Thus, we conclude that the number of hypernym occurrences of each topic is suitable to estimate the topic coherence for hyperparameter optimization – see section 5.2 later on. (Kirchhoff, 2019)

Changing Labels To change the label of a topic, the user can click on the associated label of a node. This opens a prompt allowing the user to change the topic label. The user can download a JSON file with the updated labels by clicking on the *Create file* button on the sidebar. (Kirchhoff, 2019)

5 Case Study

We demonstrate SocialVisTUM's potential for social media data exploration on a new data set.

5.1 Data Set

We crawled online user comments on organic food from multiple forums (e.g., Reddit, Quora, Disqus) and the comment sections of news websites (The Washington Post, The New York Times, Chicago Tribune, HuffPost, and many more). The goal is to discover the discussed topics and opinions in social media regarding the organic food consumption.

Relevant articles from the platforms are found by the search terms "organic food", "organic agriculture", and "organic farming". We further filtered for domain relevance by applying naive Bayes classification on bag of words trained on 1000 random and accordingly labeled texts (84.70% accuracy with 10-fold cross validation). From the left texts, we retain comments containing either of the words *food* and *organic*. The left data set consists of 515.347 sentences totaling 83.938 posts, which are used to train the ABAE topic model. We use the 300-dimensional pre-trained GloVe embeddings and fine-tuned them on the data. (Kirchhoff, 2019)

5.2 Hyperparameter Estimation

Some hyperparameters of the utilized ABAE topic model are the number of topic clusters and the vocabulary size. To optimize these automatically, we define a new metric, the *average number of shared hypernyms* (ANH). We first derive the frequency of all shared hypernyms for each topic, which is already done for automatic topic labeling in section 4. The ANH is the sum of hypernym frequencies over all topics divided by the number of topics. (Kirchhoff, 2019)

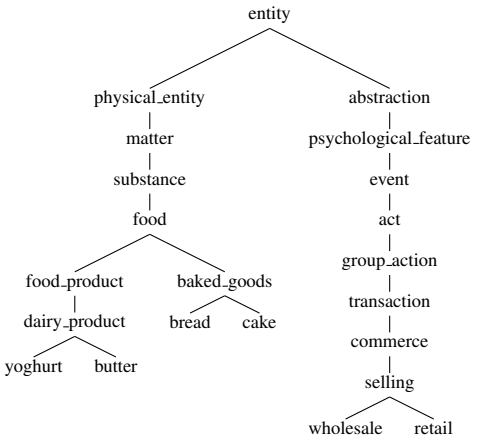


Figure 4: Hypernyms for *yoghurt*, *butter*, *bread*, *cake*, *wholesale*, and *retail*. Taken from. (Kirchhoff, 2019)

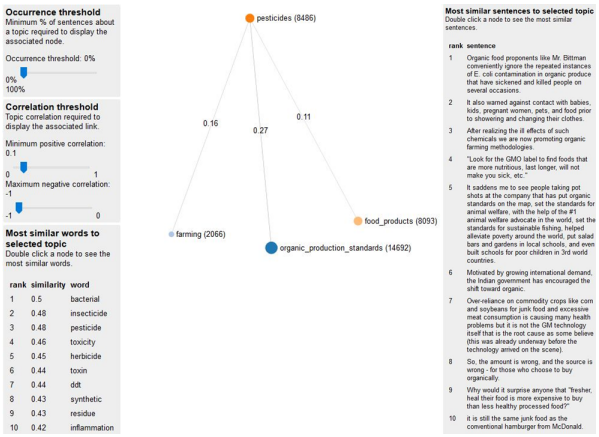


Figure 5: SocialVisTUM applied to our *organic food* use case. Topic inspection of the *pesticides* topic.

In our case study, we identified the following advantages of ANH over the widely used coherence score (CS). First, an increasing number of topics does not always increase the ANH, as a high number of topics leads to many incoherent topics with fewer shared hypernyms, i.e., a lower ANH. Second, a medium-sized vocabulary (~10,000 words) produces the most coherent topics according to ANH and manual inspection. Table 2 shows an excerpt of the results for varying parameters.

5.3 Interpretation

We applied SocialVisTUM to our case of organic food yielding the topics displayed in figure 1. A consumer researcher in the domain of organic food manually refined the automatic labeling based on the most similar words of each topic. The topics reflect previous findings of a qualitative content analysis on a small sub-sample of our data set (Danner and Menapace, 2020). The correlated topics allow market researchers to investigate the context in which topics are discussed.

Figure 5 takes a closer look at the example topic *pesticides*, which is concerned with different pesticides and their toxicity. The topic *pesticides* is correlated with the topic *organic_production_standards*, which references different organic or related production methods, such as bio-dynamic, hydroponic, or bio-intensive agriculture. This correlation suggests that, for the commenting users in our data set, the non-use of chemical-synthetic pesticides is an important characteristic of organic compared to non-organic production. Further topics correlated with *pesticides* propose that the commenting users are concerned about the use of *pesticides* in *farming* and that they discuss the issue of *pesticides*, possibly the residues thereof, in the context of different *food_products*.

6 Conclusion

In this paper, a case of the proposed SocialVisTUM demonstrates the visualization of coherent topics on a given corpus of social media texts about organic food. The graph-based visualization with topics as nodes and topic correlations as edges reflects the topics and patterns found in a related qualitative content analysis (Danner and Menapace, 2020). The presentation of additional topic information, such as word lists, representative sentences, topic importance, and meaningful predefined labels, provide a basis for the understanding and interpretation of a topic for domain experts. The integrated hyperparameter optimization automatically yields interpretable topics and helps tailoring the model to the given data set. For future work, we plan to evaluate the correlated topics on other corpora and in other use cases. In addition to Pearson correlation, other correlations could improve the approach. We plan to integrate multi-lingual word features, such as BERT (Devlin et al., 2018), for cross-cultural comparisons.

Acknowledgments

We thank inovex GmbH for supporting the research of this paper by providing computational resources, Robert Pesch and Martin Kirchhoff for their important contributions. Paragraphs adopted from student works are followed by the corresponding citation after the punctuation.

References

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. *CoRR*, abs/1808.08858.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Jonah Berger, Ashlee Humphreys, Stephan Ludwig, Wendy W. Moe, Oded Netzer, and David A. Schweidel. 2020. Uniting the tribes: Using text for marketing insight. *Journal of Marketing*, 84(1):1–25.

D. Blei and J. Lafferty. 2006. Correlated topic models. In *Advances in Neural Information Processing Systems 18*, volume 18, page 147, Cambridge, MA. MIT; 1998.

D. Blei and J. Lafferty. 2007. A correlated topic model of science. *Annals of Applied Statistics*, 1:17–35.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Zhiyuan Chen, Arjun Mukherjee, and Bing Liu. 2014. Aspect extraction with automated prior knowledge learning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 347–358.

Hannah Danner, Gerhard Hagerer, Florian Kasischke, and Georg Groh. 2020. Combining content analysis and neural networks to analyze discussion topics in online comments about organic food. In *Conference Proceedings of "3rd International Conference on Advanced Research Methods and Analytics"*, Valencia, Spain. Editorial Universitat Politècnica de València.

Hannah Danner and Luisa Menapace. 2020. Using online comments to explore consumer beliefs regarding organic food in german-speaking countries and the united states. *Food Quality and Preference*, 83:103912.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ran Ding, Ramesh Nallapati, and Bing Xiang. 2018. Coherence-aware neural topic modeling. *arXiv preprint arXiv:1809.02687*.

Mainak Ghosh. 2020. Multilingual opinion mining on social media comments using unsupervised neural clustering methods. Master's thesis, Technical University Munich, Department of Informatics, Boltzmannstr. 3, 85748 Garching, Germany. Unpublished; supervised by Gerhard Hagerer and Georg Groh.

CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*.

Junxian He, Ying Huang, Changfeng Liu, Jiaming Shen, Yuting Jia, and Xinbing Wang. 2016. Text network exploration via heterogeneous web of topics. In *ICDM Workshops*, volume abs/1610.00219, pages 99–106. IEEE Computer Society.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397.

Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. 2019. Leveraging just a few keywords for fine-grained aspect detection through weakly supervised co-training. *CoRR*, abs/1909.00415.

Martin Kirchhoff. 2019. Summarizing opinions using neural topic modeling and graph-based visualizations. Master's thesis, Technical University of Munich, Arcisstraße 21, 11. Unpublished; supervised by Gerhard Hagerer, M.Sc, and Prof. Dr. Georg Groh.

Wei Li and Andrew McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 577–584, New York, NY, USA. Association for Computing Machinery.

Shixia Liu, Xiting Wang, Jianfei Chen, Jim Zhu, and Baining Guo. 2014. Topicpanorama: A full picture of relevant topics. In *IEEE VAST*, pages 183–192. IEEE Computer Society.

Ling Luo, Xiang Ao, Yan Song, Jinyao Li, Xiaopeng Yang, Qing He, and Dong Yu. 2019. Unsupervised neural aspect extraction with sememes. In *IJCAI*, pages 5123–5129. ijcai.org.

Arun S. Maiya and Robert M. Rolfe. 2014. Topic similarity networks: Visual analytics for large document sets. *CoRR*, abs/1409.7591:364–372.

George Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272. Association for Computational Linguistics.

Joe Murphy, Michael W Link, Jennifer Hunter Childs, Casey Langer Tesfaye, Elizabeth Dean, Michael Stern, Josh Pasek, Jon Cohen, Mario Callegaro, and Paul Harwood. 2014. Social media in public opinion research: Executive summary of the aapor task force on emerging technologies in public opinion research. *Public Opinion Quarterly*, 78(4):788–794.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, page 100–108, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Archishman Roy and Jiaxi Zhao. 2020. Enhancing socialvistum. Unpublished lab course report at Technical University of Munich; supervised by Gerhard Hagerer, M.Sc., and Prof. Dr. Georg Groh.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961. Association for Computational Linguistics.

Appendix A. Comparison of Coherence Metrics

# Topics	Voc. Size	CS	ANH
5	1,000	-1104	28.6
	10,000	-765	**68.0**
	18,000	-403	5.2
15	1,000	-366	33.3
	10,000	-270	**40.0**
	18,000	-197	33.8
50	1,000	-110	30.4
	10,000	-70	**51.8**
	18,000	-54	49.7

Table 2: Comparing two coherence metrics: coherence score (CS) and average number of shared hypernyms (ANH). The advantage of ANH is that it has its global optimum always in the middle as opposed to CS. This property is beneficial for hyperparameter optimization.

A.2 The News Media and its Audience: Agenda Setting on Organic Food in the United States and Germany

The publication on the consecutive pages is currently under peer-review (2nd revision) to be published as full paper in the Journal of Cleaner Production. Gerhard Johann Hagerer, the author of the present thesis, is the second author of that paper. His author role, the reprinting permission, and his following contributions are acknowledged by all authors Danner, Hagerer, Pan, and Groh [2022]:

> *"Gerhard Johann Hagerer developed several core components of the paper methodology. Furthermore, he co-directed the implementation process and reviewed the source code deeply. Regarding the writing of the paper, he wrote the methodology part and paraphrased, corrected, combined, and otherwise improved the experiment and result parts."*

Publication Summary

The following paper is a qualitative content study investigating how online newspaper articles about organic food influence the readers' opinions formulated in the corresponding comment sections of the respective articles in the time period of 2007 – 2020. The multi-lingual topic model presented in Hagerer et al. [2021d] is used to create explanatory multi-lingual vector representations of articles and comments and to measure their correlation in the time domain based on cross-lagged correlation. The method gives meaningful insights into the multi-lingual corpus, its topic distribution, and how it develops over time in articles and comments. While the news media is found to drive opinion formation of commenters, the topic coverage and the overall discourse in news articles and comments is clearly different between German speaking and US countries, which is relevant for the organic food industry and their global marketing strategies. Methods which enable cross-lingual and cross-cultural analysis of opinions are shown to be highly relevant for globalized economies, especially for products with ethical standards, such as, organic food.

The News Media and its Audience: Agenda Setting on Organic Food in the United States and Germany

Hannah Danner[a,*], Gerhard Hagerer[b], Yan Pan[b], Georg Groh[b]

[a] *Technical University of Munich, TUM School of Management, Chair of Marketing and Consumer Research, Alte Akademie 16, 85354 Freising-Weihenstephan, Germany*
[b] *Technical University of Munich, Department of Informatics, Research Group Social Computing, Boltzmannstr. 3, 85748 Garching, Germany*

Abstract

What are the agenda-setting effects between the news media and its audience regarding organic food? This longitudinal text-mining study investigates the relationship between topics mentioned in news articles and reader comments published the online news outlets *nytimes.com* (USA) and *spiegel.de* (Germany) from 2007 to 2020. Topics are modeled using a neural network approach based on clustered multilingual sentence embeddings. Results show that the salience of topics in news articles significantly influences their salience in reader comments but not vice versa. Metrics for agenda distance and agenda diversity confirm the media's agenda-setting role and additionally point out periods of time when events caused the media and public attention to diverge. The news media drives public opinion on organic food in the US and Germany by determining the discussion topics and is thus an important player in the promotion of organic food consumption to be considered by marketers and policy makers.

Keywords: Agenda Setting, Media Coverage, Public Opinion, Organic Food, Text Mining, Topic Modeling

*Corresponding author
Email address: hannah.danner@tum.de (Hannah Danner)

1. Introduction

People's opinions are influenced by what they read in the media. The agenda-setting theory postulated by McCombs and Shaw (1972) laid the groundwork for a large body of research on how the mass media as an important source of information influences what is salient on people's minds (Abdi-Herrle, 2018, Conway-Silva et al., 2018, Gerber et al., 2009, Hester and Gibson, 2003). The original agenda-setting hypothesis focused on the media determining *which issues* are salient. Later research postulated that—"on a second level"—the media also successfully determines *how* an issue is framed (framing) and *which attributes or topics* of an issue are salient (attribute agenda-setting) (Ghanem, 1997, McCombs, 2014, Weaver, 2007, Wanta et al., 1995). Thus, attribute agenda-setting focuses on thematic variety of an issue and the relative salience of the respective topics and is the type of agenda setting investigated in this research.

The agenda-setting theory has its origins in public opinion and communication research (McCombs and Shaw, 1972). Therefore, most agenda-setting research focuses how mass media influences political opinion (Conway-Silva et al., 2018, Gerber et al., 2009, Weaver, 2007), but the influence on consumer opinions remains little investigated. This study examines the presence of agenda-setting effects in a consumption context, more specifically organic food. The mass media determines what is salient in people's minds (McCombs and Shaw, 1972). The media has been shown to not only influence political opinion, but also environmental concern and attitudes, as antecedents of pro-environmental behavior (Trivedi et al., 2018, Junsheng et al., 2019). It could thus play a decisive role in guiding desirable consumption behaviors such as buying environmentally friendly products like organic food.

Sustainability (Holt and Barkemeyer, 2012) and organic food issues (Lockie, 2006, Meyers and Abrams, 2010, Danner and Thøgersen, 2021) increasingly make it on the agenda of the news media. In addition, Danner

and Thøgersen (2021) have shown that priming organic food topics that are very frequently covered in German online news articles can nudge consumers toward buying organic food products and voting in favor of policies supporting organic agriculture. Thus, to understand the drivers of organic food consumption, it is crucial to investigate the media's influence on public opinion regarding organic food.

However, agenda setting in the context of organic food has been little researched. Thus far, news coverage on organic food (Lockie, 2006) has been analyzed separately from public opinion on organic food (i.e., attitude) (Lee and Hwang, 2016, Rodríguez-Bermúdez et al., 2020). An exception is Thøgersen (2006), who found first evidence for a connection between the positive and negative framing of Danish media articles and consumers' self-reported attitudes toward organic food.

Moreover, digitalization of the media landscape has left its imprint on agenda-setting research (Takeshita, 2006). Together with media producers and readership, research has taken the shift from print to online news media. Nowadays, many news outlets offer comment sections granting their readers a platform to voice their opinions (Santana, 2011). Comment boards of online newspapers and other online platforms such as social media provide new opportunities for quantitative analyses and insights into public opinion (Neuman et al., 2014). User-generated content such as reader comments or social media data is an indicator for what is salient to people (Ksiazek et al., 2016, Takeshita, 2006). Together with its increasing availability, such data is exploited by marketing and consumer research to gain insights into consumer thinking (Balducci and Marinova, 2018). For instance, Danner and Menapace (2020) and Olson (2017) have detected organic food beliefs and topics in user comments of US and German media.

The classic methodology of agenda-setting research consists of comparing content analyses of (print) media with public opinion surveys (Luo et al., 2019). Disadvantages of this approach are costly surveys, the required matching different units of measurement and scales, and possibly biasing time

gaps between article publication and survey (Thøgersen, 2006). Exploiting the benefits of digitalization, agenda-setting research can now directly can now directly compare topics and sentiment in online articles and comments using the same (automatic) text analysis methods (Neuman et al., 2014). Big data methods can to some extent replace but in any case complement traditional methods and test agenda-setting hypotheses (Neuman et al., 2014). However, online user-generated content on organic food has not yet been exploited in the light of agenda setting.

Against this background, the present text-mining research investigates the relationship between the salience of organic food topics in the news articles and comment sections of *nytimes.com* and *spiegel.de* as two major online US and German news outlets. This study thus focuses on attribute agenda-setting, where the attributes (hereafter referred to as topics) are the commenters' associations with *organic food*. The dynamics of the media and public agendas between January 2007 and February 2020 are analyzed and compared. They are represented by the time-evolving relative distribution of topics in the news articles and comment sections.

To detect the discussion topics, the text data from news articles and reader comments are analyzed with a topic modeling approach based on clustered multilingual sentence embeddings. This technique from the field of natural language processing provides several novelties and advantages over classical approaches such as content analysis (Meyers and Abrams, 2010, Danner and Menapace, 2020): Large amounts of data can be analyzed (Neuman et al., 2014). Further, the topic modeling approach allows topics to emerge from the data without requiring prior knowledge. As of now, there are few examples of studies using such an exploratory and data-driven text analysis approach (Guo et al., 2016). An exception is the study by Pinto et al. (2019), who analyzed how topics in the Argentinian news media are reflected in *Twitter* activity and *Google* searches. Finally, a multilingual topic modeling approach enables the analysis and comparison of agendas in different countries and languages, here English and German. Such a comparison across languages is unprecedented in

agenda-setting literature to the best of the authors' knowledge. In addition, most research has focused solely on the US context (Luo et al., 2019).

In the following, the theoretical background of this study is outlined (section 2). Subsequently, the study's methodology and the employed agenda-setting metrics are described (section 3), before presenting the results on the different agenda-setting metrics (section 4). The paper concludes with a discussion of results (section 5) and conclusions (section 6).

2. Theoretical Background

2.1. Organic Food

Food is a consumption domain that is critical to the environmental impact of households (Tukker et al., 2010). Both consumers and research generally understand as a more sustainable alternative (Thøgersen, 2010, Siegrist et al., 2015). Organic food market shares have been rising in the past decades, reaching 5.7% of total food sales in the US and 5.3% in Germany in 2018 (BÖLW, 2020). According to long-standing consumer behavior theories, such as the Theory of Reasoned Action (Fishbein and Ajzen, 1975) and the subsequent Theory of Planned Behavior (Ajzen, 1991), consumers' attitude drives purchase intention and eventual buying behavior. Attitude is formed by the sum of beliefs regarding the attributes of a product, such as organic food. An extensive body of literature has elicited the attributes consumers associate with organic food via surveys, qualitative interviews, and focus groups. For reviews see for example Hemmerling et al. (2015), Hughner et al. (2007), Kushwah et al. (2019). The reviews present a large variety of organic food attributes regarding the organic products themselves (e.g., healthiness, food safety, price, taste, quality and appearance, nutritional value, naturalness, genetically-modified organism free, availability, local origin), their production process (e.g., organic labeling and certification, environmental impact, animal welfare), and the societal role of organic farming (e.g., food security).

2.2. Agenda Setting

Research on agenda-setting theory McCombs and Shaw (1972) started out with studying the media's influence on *issue salience* and moved on to investigating *attribute salience* and framing effects (Ghanem, 1997, McCombs, 2014, Weaver, 2007, Wanta et al., 1995). A recent meta-analysis by Luo et al. (2019) analyzed 67 agenda-setting studies from 1972 to 2015 and confirmed the media's power to influence the issues and topics salient on the public agenda. The vast body of agenda-setting research has been carried out in the context of political opinions (Luo et al., 2019). For instance, authors have shown how the media influences public opinions on politics and voting behavior (Conway-Silva et al., 2018, Gerber et al., 2009, Groshek and Groshek, 2014). In contrast, research on whether and how the media influences consumer behavior in general and organic food consumption in particular is still scarce or outdated.

Mahlau (1999) had analyzed the image of the agricultural sector in German print media from 1980 to 1994 and compared the facticity of the image in the media and in the population, observing that both media and public agenda have biased views on agriculture. However, Mahlau (1999) did not find an agenda-setting function of the media as topics and biases differed in the media and public agendas. Whereas Ader (1995) documented a causal relationship between salience in media and public agenda regarding the salience environmental pollution, a problem related to farming topics.

In the context of organic food, first evidence for framing—as one type of agenda setting—was found in Denmark (Thøgersen, 2006): From the late 1990s to the early 2000s negatively framed articles became more newsworthy, while at the same time several psychological indicators concerning organic food declined. This indicates a potential connection between the media frame and consumers' attitudes toward organic food (Thøgersen, 2006). The influential role of the media is further underlined by findings showing that in particular negative media reports seem to catch consumers attention (Yadavalli and Jones, 2014). They have also been demonstrated to influence attitude and

purchase intention regarding organic food (Müller and Gaus, 2015).

Attribute agenda-setting—i.e., the relationship between topics discussed in the media and the public (Weaver, 2007)—has not yet been researched in the context of organic food. According to attribute agenda-setting, the media determines the topics and transfers their salience (i.e., accessibility or top-of-mind awareness of a topic) to the public (Ghanem, 1997, Takeshita, 2006, Weaver, 2007). By raising their voice in online comments, readers make a deliberate judgment about topic importance (Ksiazek, 2018), which can affect consumer behavior. For example, concern about social and environmental issues (e.g., about pesticide use in agriculture) was shown to influence organic food purchases in Denmark (Thøgersen and Ölander, 2006). Bitsch et al. (2014) found that the media coverage of food safety incidents in the US and Germany affected consumer risk perception and impacted actual food purchase behavior. Furthermore, (Danner and Thøgersen, 2021) show that a topic salient in the media agenda such as *animal welfare* was effective in priming pro-organic consumer behavior, in opposition to a topic with little salience in the media such as *biodiversity*. According to Spreading-Activation Theory, consumers store product knowledge in associative networks and activate their knowledge and attitude toward a product when faced with the product or its attributes (Collins and Loftus, 1975, Fazio, 1986). The more often this associative network is activated, e.g. through reading respective news articles or user comments, the more likely it is to affect consumer behavior (Berger and Mitchell, 1989). In sum, the media's impact on consumers bears valuable insights into organic food perception and behavior. However, attribute agenda-setting—the relationship between attributes or topics discussed in the media and the public—has not yet been researched in the context of organic food. Against this background, this study hypothesizes:

Hypothesis 1: Online news articles on organic food determine the topics discussed in their reader comments.

Another aspect of agenda setting is time lag, i.e., persisting causal effects over time (Roberts et al., 2002, Luo et al., 2019). Therefore, this study tests

the following hypothesis for the context of organic food:

Hypothesis 2: The topics covered in online news articles on organic food predict the topics discussed in reader comments of subsequent articles.

Furthermore, studies up until the early 2000s mostly documented unidirectional causality of agenda setting. However, influence can be reciprocal (Denham, 2010, Neuman et al., 2014). The manifold possibilities to voice opinions online turned readers into senders (Santana, 2011). When journalists respond to public interests, the public can influence the media agenda. This is also referred to as agenda-building or reverse agenda-setting (Denham, 2010, McCombs, 2014). Research on how audience feedback is incorporated in news production is still scarce (Lee and Tandoc, 2017). This paper investigates reverse agenda-setting in the organic food context. It is hypothesized:

Hypothesis 3: Reader comments made in response to one news article on the issue of organic food influence the topics of the subsequent organic food news article (i.e., reverse agenda-setting).

In addition to the agenda-setting relationships on single topic level, few researchers have studied overall topic representation, i.e., how presence and domination of topics changes over time and differs between media outlets as well as between the media and the public agenda (Boydstun et al., 2014, Pinto et al., 2019). Building on this research, this study compares agenda diversity and the agenda distance between media and public for the two country contexts.

In sum, this study aims at investigating the media and public agendas, the dynamics, and their relationship in a sustainable consumption context, specifically organic food. Therefore, it analyzes articles and reader comments over time in two leading US and German online news media outlets, namely *nytimes.com* and *spiegel.de*.

3. Methodology

3.1. Data

This study used news articles and reader comments of two major news outlets serving as characteristic examples of the German and the United States (US) context, respectively. *Spiegel.de* and *nytimes.com* were selected. They are high-quality, general-purpose media outlets leading in terms of print coverage and online views in their countries (statista, 2019, 2020), with large amounts of data available. For reasons of readability, they will hereafter be referred to as US and Germany. A broad time frame spanning January 2007 to February 2020 was chosen to thoroughly observe the dynamics of the media and public agendas on organic food.

News articles on the issue of organic food were identified using the search terms *organic food* and *organic farming* and the German equivalents *Bio-Lebensmittel* and *Bio-Landwirtschaft*. All articles on organic food within the time frame were collected together with the reader comments in response to those articles. Subsequently, the remaining data was pre-processed for data analysis via tokenization, the removal of meaningless text (e.g., URLs) and stopwords (e.g., "and"), and lemmatization (i.e., removal of inflectional forms). Irrelevant content was filtered out in the topic modeling process (see section 3.2.4). In total, this study analyzed 534 articles and 41,320 comments from the US, and 568 articles and 63,379 comments from Germany. Data stemmed from the years 2007 to 2020 for the US, and for the years 2007 to 2017 and the year 2020 for Germany. Unfortunately, no German comments were available for 2018 and 2019 due to website restructuring. Despite the discontinuity in the German data, the authors chose to include the data available up until 2020 to provide the richest and most current insights for the two countries.

3.2. ***Topic Modeling***

3.2.1. Overview

Alongside with the growing availability of large amounts of text data from online media and social media platforms, big-data analysis techniques have

provided an efficient alternative or complement to traditional methods such as content analysis (Neuman et al., 2014). For an overview of text-mining methods in communication research see Guo et al. (2016), and in consumer research see Berger et al. (2020). The most frequently employed automated text analysis approaches in the social sciences are dictionary-based techniques (e.g., LIWC, Tausczik and Pennebaker, 2010). In addition, methods from computer science—more specifically natural language processing (NLP)—have been gradually introduced in the social sciences (Berger et al., 2020, Guo et al., 2016, Jacobi et al., 2015). Unsupervised topic modeling is frequently employed and allows for exploratory topic analyses without requiring prior knowledge. The most prominent topic-modeling approach is the Latent Dirichlet Allocation (LDA) established by Blei et al. (2003), which generates a probabilistic distribution of words and topics in text documents. However, NLP also offers more advanced topic modeling approaches that are based on deep neural networks and account for the semantic context of words via word embeddings (Mikolov et al., 2013), or of sentences via sentence embeddings (Cer et al., 2018). This text-mining study employed a topic modeling approach based on multilingual sentence embeddings clustered with k-means clustering. This novel approach has been previously tested for the case of organic food (Hagerer et al., 2021). The subsequent paragraphs explain its foundations and implementation.

3.2.2. Sentence Vectors

Recent advances in NLP are subject to the evolutions of deep learning (Devlin et al., 2018). One of the past years' most influential developments has been the development of pre-trained word embeddings (Mikolov et al., 2013, Le and Mikolov, 2014), where each word is represented as a vector that captures semantic and syntactic features as well as the context of a word. Similar words have close and different words distant vector representations. Knowledge regarding semantic similarity is derived from large-scale text corpora (e.g., *Wikipedia* using unsupervised learning. Pre-trained word

embeddings improve the performance on many NLP tasks compared to traditional NLP methods (Hossain et al., 2019). Further improvements are achieved by contextualized word embeddings (Devlin et al., 2018) and transformer networks (Yang et al.), which constitute the state-of-the-art in NLP. This also holds true for sentence embeddings employed in the present study, which are derived from word embeddings (Devlin et al., 2018).

This study targets a multilingual environment in English and German. Therefore, Google's XLING was the model of choice. It allows to coherently embed both English and German texts, as it is additionally trained on respective translations (Chidambaram et al., 2018). It is accessible through Tensorflow Hub (Yang et al., 2018). An input sentence (German or English) is transformed into a 512-dimensional vector as shown in Figure 1. For two sentences in German or English with a similar meaning, the scalar product of the two sentence vectors approaches 1, whereas it approaches -1 in case of semantic dissimilarity—see Figure 1 on the right.

3.2.3. Topic Distributions

In NLP, the bag-of-words technique for document representation has a long tradition (Harris, 1954, Hossain et al., 2019). The occurrence of words is counted resulting in a histogram with the dimensionality of the whole vocabulary. Analogously, in the bag-of-concepts technique (Kim et al., 2017) counts word or sentence vectors (Schmitt and Schuller, 2017), which can be clustered using different clustering techniques to model topics. The topic distribution of a document is thus a histogram representing the number of vectors in each cluster—see Figure 2. There are two approaches to bag-of-concepts in topic modeling. Sridhar (2015) used *word2vec* word embeddings and Gaussian mixture models to derive the topic distributions based on soft quantization. He et al. (2017) proposed attention-based aspect extraction, i.e., an unsupervised algorithm using word2vec to derive meaningful semantic clusters from sentence vectors. In terms of topic coherence, both approaches have clearly outperformed traditional topic

modeling such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) not only for short texts such as Tweets or product reviews, but also for long texts such as scientific and news articles (Xing et al., 2014, Kim et al., 2017). Therefore, clustering pre-trained sentence embeddings was chosen over traditional methods like LDA for the topic modeling in the present study,.

3.2.4. Implementation

The goal of the topic modeling based on clustered pre-trained sentence embeddings was to develop a universal topic model for all data from US and German articles and comments.

First, a sentence vector was calculated for each available sentence in the data set using the pre-trained deep neural network XLING—see Figure 1.

Second, the resulting sentence vectors were clustered into topics using k-means clustering—see Figure 2. The topic distribution per document (i.e., a news article or the aggregated comments of a news article) were derived by counting the vectors in each cluster, which yielded a histogram for each document as depicted in Figure 2. This means that each document was not assigned to only one topic but to different topics in different proportions. For the selection of the appropriate number of clusters k (i.e., topics), the Akaike Information Criterion (AIC) indicated that a model with $k = 15$ was the optimal, parsimonious model. This was confirmed by two domain experts, who evaluated the semantic coherence of the topics based on the most important words (i.e., top words) representing each cluster. In the $k = 15$ model, five clusters were garbage clusters containing semantically incoherent top words and/or content irrelevant to the search terms *organic food* and *organic farming* and the German equivalents. These clusters were excluded from further analysis, resulting in ten topics to be included in the agenda-setting analyses.

Third, the resulting topic clusters were given labels manually based on the top words (see Tables 1 and 2). Two domain experts initially labeled the topics independently. There were only minor differences in the assigned topic labels, which were subsequently resolved by discussion.

3.3. Agenda-Setting Metrics

The subsequent agenda-setting analyses evaluated and compared to which degree the ten identified topics were represented in the US and German media and public agendas. Hence, the unit of measurement applied in this research is relative topic distribution. More precisely, the proportion of a topic in the topic distribution of an article or its comments represents its salience, i.e., the degree of attention paid to the topic.

To investigate agenda setting, the relative topic distributions in the news articles and the comments were compared using the following metrics also outlined in Table 3. On the one hand, correlational analyses between single topic proportions—as they are typical of agenda-setting research (Abdi-Herrle, 2018, Lim, 2006, 2011)—were applied. On the other hand, two metrics from information and probability theory were employed, which consider the distribution of all topics simultaneously: Normalized Shannon entropy for agenda diversity, and Jensen-Shannon distance for agenda distance (Boydstun et al., 2014, Pinto et al., 2019). Further details on these metrics can be found in the appendix. In spite of the lack of German comment data for the years 2018 and 2019, the time span up until 2020 was included to yield an up-to-date picture for the US articles and comments. For Germany, there are still some insights into the agenda diversity in articles up to 2020. However, agenda distance could consequently not be measured for the two missing years, and there were fewer correlational data pairs for the two correlation-based measures.

4. Results

4.1. The Topics

The multilingual topic clustering revealed that the US and German media and public agendas were represented by the following ten topics: *Food Safety & Chemicals & GMO*, *Food Products & Quality*, *Health & Nutrition*, *Environment & Climate Change & Energy*, *Farming*, *Animal Welfare & Meat*

Consumption, *Retailers & Prices*, *Economy & Affordability*, *Politics*, and *Evidence*. The multilingual topic clustering allowed to identify the ten main organic food discussion topics in both countries and their comparison in the subsequent analyses. At the same time, country-specific characteristics were retained. This can be seen from differing top words (see Tables 1 and 2). In particular, the topic *Politics* revealed the political parties and politicians of the respective countries. Further, the topic *Retailers & Prices* indicated which retailers are present in the US and German food markets.

Figure 3 gives an overview of the salience of the ten topics in US and German articles and comments. Differences in topic salience were found between the media and the public agendas within in each country. For example, in the US, the topic *Retailers & Prices* was represented on average with more than 20% in articles and with only approximately 8% in comments. Similarly in Germany, the topic *Farming* was far more salient in the media than among readers. Furthermore, the salience of topics differed between countries. The US media covered in particular the topics of *Farming*, *Retailers & Prices*, *Food Products & Quality*, *Health & Nutrition*, and *Food Safety & Chemicals &GMO*. The public agenda in the US also prioritized these topics; additionally, the topic *Evidence* was salient, indicating that consumers discuss the trustworthiness of organic food and the origin and reliability of information.

The most salient topics in the German media were *Farming*, *Retailers & Prices*, *Economy & Affordability*, *Politics*, and *Food Products & Quality*. The topic overlap between media and public agenda was smaller for Germany indicating slightly different agenda priorities. Commenters were mostly concerned with *Economy & Affordability*, *Politics*, *Evidence*, *Animal Welfare & Meat Consumption*, *Environment & Climate Change & Energy*, and *Food Safety & Chemicals &GMO*.

4.2. Synchronous Topic Correlations

In the following, the general agenda-setting function of the media was investigated. To that end, the linear relationship between the relative proportion of a given topic in the articles and in the comments corresponding to each article was determined by calculating the correlation between the two relative proportions. The coefficients ρ_t for the ten topics are displayed in Table 4.

As hypothesized (H1), there is a positive and significant linear correlation between each topic in an article and the same topic in the comments to that article for both US and Germany—see Table 4. In the US, all correlations were strong with coefficients ranging between $r = 0.540$ for *Evidence* and $r = 0.808$ for *Health & Nutrition*. For Germany, correlation coefficients range from a medium-sized correlation for *evidence* at $r = 0.423$ to strong correlations for all other topics, the strongest correlation being $r = 0.913$ for *Food Products & Quality*. All correlations were significant at $p < 0.001$. In conclusion, the strong associations found indicated that topic proportions in comments and articles were largely similar. This means that the salience of any of the 10 topics in the article largely corresponded to the salience in the respective comments.

4.3. Cross-lagged Topic Correlations

The previous section showed that the topics in the news articles strongly influence the topics of the respective comments. This section investigates whether topics discussed in articles influenced not only their own comments but also comments of future articles $P_{X_1Y_2}$. In addition to classical attribute agenda-setting, it was tested whether the topics discussed in comments at one point influenced the topics covered by the media in future articles $P_{Y_1X_2}$ (reverse agenda-setting). To that end, the cross-lagged correlations between a given topic in articles and comments (Kenny, 1975, Rogosa, 1980) were analyzed. A bi-weekly time lag was used, as both news outlets published articles on organic food on average every two weeks. Following the example of

previous cross-lagged correlational analyses in the context of agenda setting (Abdi-Herrle, 2018, Lim, 2006, 2011, Sweetser et al., 2008), the Rozelle-Campbell baseline (RCB) (Rozelle and Campbell, 1969) tested the validity and direction of the cross-lagged correlation. RCB designates the level of correlation to be expected on the basis of the autocorrelations and synchronous correlations. Valid cross-lagged correlations are present if they exceed the RCB.

First, as hypothesized (H2), total cross-lagged correlations indicated that the media agenda influenced the future public agenda, given $P_{X_1Y_2} > RCB$, in the US and Germany (see Figure 4). There are cross-lagged correlations ($P_{X_1Y_2}$) of $r = 0.375$ in the US and $r = 0.211$ in Germany.

Second, the sizes of cross-lagged correlations differed between topics and countries when looking at the topic-level correlations (see Tables 5 and 6). In Germany, the topics *Environment & Climate Change & Energy* and *Animal Welfare & Meat Consumption* were correlated more strongly than other topics, indicating that the media had made these topics very salient, to the point that their salience in public opinion persisted in the subsequent time window (see Table 6). In the US, correlation sizes were more homogeneous across topics, with *Food Safety & Chemicals & GMO* and *Health & Nutrition* and *Retailers & Prices* being the topics most correlated over time (see Table 5).

Third, disconfirming H3, total cross-lagged correlations indicated that comments did not influence later media articles, given $P_{Y_1X_2} < RCB$ (see Figure 4). For Germany, hardly any correlation was found for $P_{Y_1X_2}$ ($r = 0.008$). For the US, there was a weak correlation ($r = 0.158$), which, however, was not valid as it did not surpass the RCB. At topic-level none of the correlations between comments and articles were significant. For cross-lagged correlations, it is concluded that the influence between articles and comments was unidirectional—articles influenced future comments (confirming H2), but comments did not influence future articles (rejecting H3). Thus, there is no evidence for reverse agenda-setting as article authors did not seem to consider topics discussed in the comments. In addition, topics salient

to commenters in one time period maintained their salience in the subsequent time period. This was shown by significant autocorrelations between the comments in consecutive time windows. For the articles, however, only a very weak indication for such agenda persistence was found for Germany ($r = 0.093$) and the US ($r = 0.158$) (see Figure 4).

In conclusion, articles as well as comments from one time period influenced the topics discussed in later comments. In contrast, the media agenda was not susceptible to influence from previous articles or comments.

4.4. Agenda Diversity

The previous sections provided evidence for agenda setting in the context of organic food. The salience of a topic in the media agenda predicted the salience of the topic in the public agenda. Subsequently, normalized Shannon entropy H was used to investigate how the diversity of topics in the media and public agenda evolved over time. H measures the diffusion of distributions. H increases when the topics are equally distributed, and H decreases when certain topics dominate. Thus, the time periods of low agenda diversity can be interpreted by identifying local minima of H. This answers the questions of whether topics are equally salient from 2007 to 2020, and whether certain topics dominate the agenda in certain time periods. Figure 5 shows the agenda diversity H over time for the US and German media and public agendas. The insights from agenda diversity were two-fold.

First, there were statistically significant differences in the levels of agenda diversity. In the US, the diversity of articles was significantly higher than the diversity of comments ($F(1,154)=55.4$, $p<0.001$). Likewise, in Germany, the articles were more diverse than the comments ($F(1,145) = 11.8$, $p <0.01$). The observation that the media agenda was more diverse than the public agenda is consistent with this article's previous findings from topic correlations, which indicated that commenters stuck closely to the topics discussed in the articles. Across countries, the agenda of the German articles was significantly more diverse compared to those in the US ($F(1,156)=23.7$, $p<0.001$). Also, the

agenda of the German comments was more diverse than in the US ($F(1,143)=13.8$, $p<0.001$). It was noted that German commenters not only cover a wider range of topics, but also discussed more extensively: On average, German comments are more than twice as long than US comments in terms of sentences.

Second, the entropy metric identify time periods in which the agenda diversity is particularly low, i.e., where the agenda was dominated by specific topics. Between 2007 and 2020, there were seven time periods with this type of agenda diversity minima, denoted with *a*, *b*, *c*, *d*, *e*, *f*, and *g* in Figure 5. For these time periods, the radar plots in Figure 6 show the topic distribution and thus, which topic dominated in which agenda. By additionally inspecting the articles and comments from those time periods, the minima in agenda diversity could be linked to real events, leading to the following interpretations: Time period *a* coincides with a peak in global food prices, which were also connected to food riots in the advent of the *Arab Spring*. The German media agenda was dominated by the topics *Economy & Affordability* and *Farming* and discussed financial speculations with agricultural commodities, the strongly subsidized EU farming sector, and rising food prices against the background of global food security. Time period *b* covers the start of Barack Obama's first mandate as US president. The US public agenda was dominated by the topic *Health & Nutrition*, as commenters discuss the president's food policy as well as Michelle Obama promoting healthy nutrition and installing an organic vegetable garden on the White House premises. In time period *c*, the German public agenda showed a strong minimum. The public discussion was dominated by the topic *Environment & Climate Change & Energy* in light of the 2009 *United Nations Climate Change Conference* in Copenhagen. Additionally, commenters debated the environmental policy of the newly elected German conservative-liberal government. In time period *d*, the topics *Animal Welfare & Meat Consumption* and *Economy & Affordability* prevailed the German comments. Users discussed the introduction of the EU organic label in 2010 as well as animal welfare benefits of organic animal husbandry.

The next minimum *e* was in 2014 with *Food Safety & Chemicals & GMO* as the predominant topic among US commenters. At the time, new scientific evidence on the food safety of GMO and organic food had been released and certain US states had introduced mandatory labeling of GMO-ingredients. In this context, consumers pondered the meaning of *naturalness*. In *f*, the topic *Politics* prevailed in the German public agenda given the federal elections of 2016. Finally, in time period *g*, there was a minimum in the US media agenda with *Retailers & Prices* being the dominant topic. The US public agenda also focused on this topic although not classifying as a true local minimum. In 2017, the US company *Amazon* acquired the US retailer *WholeFoods*, which is known for its large organic assortment.

In conclusion, agenda diversity differed significantly across the agendas. There were seven time periods of exceptionally low agenda diversity, which could be traced back to real world events. It is noteworthy that only *a* and *g* were minima in the media agenda, whereas the remaining minima were in the public agendas. This confirms that the media agendas maintained a certain topic diversity in their reporting on organic food. In contrast, commenters temporarily focused their attention to specific topics, which were linked to political and economic events at the national and global level.

4.5. Agenda Distance

In the previous section, the diversity of of the media and public agendas was inspected separately. Agenda diversity was found to differ over time, with specific time periods exhibiting dominant topics. However, the topics did not always dominate media and public agendas to the same extent. Consequently, the question remains how similar those agendas were. Jensen-Shannon distance (JSD) was used to measure the similarity between the relative distribution of topics in the media and public agenda for each country. Similar to the correlational metrics, JSD measured the relationship between articles and comments. However, while the correlational measures compared topic proportions individually, JSD considers the distribution of all topics. If JSD is

close to 0, the topic distributions of articles and comments in this time period are very similar, i.e., the closer media and public agendas were related. At large values of JSD, the media and public agendas diverged. Results are displayed in Figure 7. The insights from the agenda distance metric were manifold.

First, low JSD values indicated that the topic distribution was rather similar in articles and comments (see Figure 7). This confirms the findings from topic correlations above.

Second, the agenda distance differed between countries being significantly higher in Germany compared to the US ($F(1, 143)=15.0$, $p<0.001$). This suggests that German commenters stuck slightly less to the topics discussed in the media articles.

Third, two local maxima were identified: *c* and *d* indicate periods of time where the public paid attention to different topics than the media (see Figure 6). The time periods *c* and *d* coincide with the time periods of low agenda diversity in the public discussion in Germany (see Figure 5 and the topic distributions depicted in Figure 6). Agenda distance was large because—in contrast to the media—commenters had focused their attention on environmental policies in *c*, and animal welfare and economic issues in *d*.

5. Discussion

By successfully applying agenda-setting theory and established and novel respective metrics to the organic food context, this study validated agenda-setting theory with a focus on attribute agenda-setting for the use in future consumer behavior studies. Further, it is the first research investigating attribute agenda-setting on the issue of organic food in an online environment. While Thøgersen (2006) looked into the influence of positive or negative framing in print news articles on organic food, this research is based on an exploratory topic clustering to detect the discussion topics. A methodological contribution is made by applying the novel approach of multilingual topic

clustering, which allowed to compare the online salience of topics in different languages and thus make country comparisons. Several insights on agenda setting regarding organic food were obtained.

First, as hypothesized (H1), there was an attribute agenda-setting relationship between the two news outlets and their audience. Both media outlets strongly influence the topics readers write about: The topic proportions in news articles and their comments were strongly correlated.

Second, cross-lagged correlations revealed that the media agenda predicts the public opinion voiced in comments to subsequent articles, which confirms Hypothesis 2. Surprisingly, the public agenda does not influence the future media agenda and therefore Hypothesis 3 is rejected. Hence, counter to previous political opinion research (Groshek and Groshek, 2014, Conway-Silva et al., 2018), no reverse agenda-setting effects were found in the context of organic food. This is in line with previous literature finding that the traditional agenda-setting direction from the media to the public is still stronger than from the public to the media (Groshek and Groshek, 2014, Conway-Silva et al., 2018). This could be interpreted as article authors of *nytimes.com* and *spiegel.de* not considering reader comments when drafting news articles on organic food. Hence, an article's topic could mostly be determined by the media responding to external events (i.e., political decisions, scandals). However, following up on consumers' concerns about organic food could be important to cater consumers' information needs and eventually help them in their purchase decisions (Trivedi et al., 2018). However, although readers did not influence article authors, readers still influenced each other: Cross-lagged correlations indicated that comments do affect future commenters. One explanation could be that readers potentially take the opinions stated in comments for the public opinion and conform to it (Lee, 2012). In a more recent study, Lee et al. (2017) showed that reader comments make certain aspects of a news story salient, and thus influence how the reported news event is interpreted in public opinion. Prior research has also shown that the tone of the comments influences subsequent comments (Ziegele

et al., 2014) as well as the perception of the news article (Winter et al., 2015).

Third, metrics from information and probability theory provided additional insights into the dynamics within and between agendas (Boydstun et al., 2014, Pinto et al., 2019). The distribution of all topics on the agendas was accounted for by calculating normalized Shannon entropy for agenda diversity and Jensen-Shannon distance for the distance between the media and public agendas, whereas correlations—a classical agenda-setting metric—had considered the proportions of each topic separately. While confirming the close semantic relationship between articles and comments, agenda diversity and agenda distance additionally pointed out a number of time periods where agenda distance was particularly high and where agenda diversity was especially low, because specific events caused the audience to focus on other topics than the media agenda and vice versa. In both countries, the media agenda maintained a higher diversity than the public agenda suggesting that readers tend to discuss a smaller selection of same topics, while the media takes up various topics from contemporary events, for instance. When comparing *nytimes.com* and *spiegel.de*, agenda diversity was significantly higher in Germany compared to the US, meaning that more topics are discussed in Germany. Also, agenda distance was higher in Germany, while US commenters stuck more closely to the topics mentioned in the news articles. This indicates that in Germany public discourse on organic food seems to be more detached from the media agenda than in the US. German consumers possibly use comment boards on news websites to discuss their general opinion and concerns about organic food regardless of current events and respective news articles.

In addition to agenda setting, this study showed how text mining can provide a comprehensive overview of which topics are salient in the media and its audience. Topics ranged from typical product attributes of organic food such as food safety to production characteristics such as animal welfare as well as political and economic aspects of the organic food system. A comparison with extant literature confirms validity. The ten identified topics concur with

previous results on topics salient in online media and user-generated content on organic food (Danner and Menapace, 2020, Lyu and Choi, 2020, Meza and Park, 2016, Olson, 2017), as well as buying motives documented in survey research (Hemmerling et al., 2015, Hughner et al., 2007, Testa et al., 2019). The salience of the ten topics differed in *Nytimes.com* and *spiegel.de*. US media and public agenda concentrated on safety-, health-, quality-, and price-related aspects of organic food, i.e., topics concerning the personal benefits of buying organic food. In contrast, the German media and public agenda were more concerned with external impact of organic food production, i.e., the consequences for the environment and animal welfare. Further, they discussed more political and economic aspects of organic food as well as different retailers. This resounds with qualitative findings of Danner and Menapace (2020), which showed that product-related topics such as health and food safety were more prominent in the US than in Germany. The different emphases of both media and consumers in the two countries could be related to potentially different aspects highlighted in marketing organic products.

In addition, while literature already knows that consumers discuss a wide array of organic food topics (Danner and Menapace, 2020, Hughner et al., 2007), the ten identified topics suggest that the documented media agendas on organic food in both countries have become more diverse over the last decades. According to Lockie (2006), the discussion in the US media from 1996 to 2002 had been limited to mainly food safety issues and the general conflict between organic and conventional foods. An analysis of five US print media, including *New York Times*, from 2005 to 2006 showed that in addition to health, production, industrialization, and in particular ethical topics were covered (Meyers and Abrams, 2010). For Germany, an analysis of news media in the 1980s and 1990s revealed that the media coverage on agriculture in general focused on trade and agricultural policy—subsidies in particular (Mahlau, 1999). Environmental aspects of agriculture were largely and organic farming completely ignored. By contrast, the findings of this study imply that the current media attention is more diverse. This could be grounded in the media

taking into account the consumer perspective by discussing topics such as food safety, product quality, and healthiness. Moreover, the growth of the organic food markets over the past decades (FiBL, 2020) could have caused the media to discuss the issue more comprehensively.

5.1. Research Implications

This study found agenda-setting effects from media to public agenda but no reverse agenda-setting effects from public to media. Future research could further explore the directions of influence between the media and public in the context of consumption. Moreover, this study analyzed one exemplary news outlet from each the US and Germany. However, nowadays, there are reciprocal and dynamic flows of agenda between a large number of news outlets and discussion platforms with reader comments representing only a part of the public opinion (Denham, 2010, Neuman et al., 2014). Following research could analyze more news outlets and discussion platforms and also the influence between different outlets/platforms, i.e., intermedia agenda-setting (Lim, 2011). For political topics, intermedia agenda-setting was documented between the social network *Twitter* and news media (Groshek and Groshek, 2014, Conway-Silva et al., 2018). Meza and Park (2016) discovered that *Twitter* is also a valuable channel for word-of-mouth communication on organic food both among consumers as well as between consumers and companies. Therefore, future research could investigate the interplay between the news media and public opinion platforms like *Twitter* in the context of organic food.

5.2. Practical Implications

Results indicate that the media can foster and direct its readers' discourse on organic food. The media determines issue and attribute importance of organic food, which in turn can potentially affect behavioral outcomes (Danner and Thøgersen, 2021). Buying organic food is seen as one path to more sustainable consumption patterns (Thøgersen, 2010, Siegrist et al., 2015).

News media could support this transition. However, media has also been found to spur criticism of organic food (Thøgersen, 2006, Vittersø and Tangeland, 2015) with negative news receiving more attention (Müller and Gaus, 2015, Yadavalli and Jones, 2014). Thus, the media carries ethical responsibility in selecting the issues and topics on the agenda of consumers, as it is deemed to sustain a certain consensus-building function (McCombs, 2014).

Furthermore, the findings of this study are relevant to different players promoting organic food purchases. Marketers, lobbyists, policy makers, and politicians can consider the media as an important information channel by bringing specific topics to the media's attention, and thus, shape the public perception of organic food. Media campaigns can accompany policy measures such as the introduction, adaptation, and promotion of organic standards, and thereby increase consumer awareness, familiarity, and salience thereof (Aitken et al., 2020, Brach et al., 2018).

6. Conclusions

This study is the first to comprehensively document the attribute agenda-setting function of the online media in the context of consumption using the example of organic food. It investigated the topics in news articles and reader comments of *nytimes.com* *spiegel.de* from 2007 to 2020. A multilingual topic clustering was applied to compare the topics across sources. Correlational analyses show that media determines the topics on the public agenda but not vice versa. Metrics of agenda diversity and distance confirmed a close semantic relationship between media and public agenda and pointed reald world events dominating the agendas. Given the influence of the media on its readership, this research emphasizes the media as an important player in the transition to a more sustainable food system, economy, and society.

CRediT Authorship Contribution Statement

Hannah Danner: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Writing - review & editing, Supervision, Visualization Project administration, Funding acquisition. **Gerhard Hagerer:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Investigation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration. **Pan Yan:** Methodology, Software, Formal analysis, Data curation, Writing original draft, Writing - review & editing, Visualization. **Georg Groh:** Conceptualization, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Funding: This work was supported by the German Academic Scholarship Foundation. The authors thank Jutta Roosen and Luisa Menapace for their helpful comments.

Appendices - Agenda-Setting Metrics

A Synchronous Topic Correlations

Let $t = 1, ..., 10$ be the topic index, T the total number of topics, $i = 1, ..., N$ the article index, N the total number of articles, $x_{i,t}$ the proportion of topic t in article i, and $y_{i,t}$ the proportion of topic t in all comments of article i. Then, for one topic of interest t, $\rho_t = \rho_{x_t y_t}$ is the Pearson correlation coefficient of the topic proportions $(x_{i,t}, y_{i,t})$ with $i = 1, ..., N$.

B Cross-Lagged Topic Correlations

The time dimension was included into the calculation of the Pearson correlation as follows: The years 2007 to 2020 were divided into time windows of two weeks because on average every two weeks a news article on organic food was published in the two media outlets. Within a given two-week time window $w = 1, ..., W$ with W being the total number of time windows, the number of all article sentences is denoted as x'_w and the number of all comment sentences as y'_w. The proportion of topic t in the articles within w is $x_{w,t} = \frac{x'_{w,t}}{x'_w}$, and the proportion of topic t in the comments of w is $y_{w,t} = \frac{y'_{w,t}}{y'_w}$. Now, let w_k be an arbitrary two-week time window and w_{k+1} the subsequent time window. Then, the autocorrelation for articles $P_{X_1X_2} = \rho_{X_1,X_2}$ was calculated based on all value pairs $(x_{w_k,t}, x_{w_{k+1},t})$ with $k = 1, ..., W-1$ and $t = 1, ..., T$. This was done analogously for the autocorrelation of comments $P_{Y_1Y_2}$, as well as for the synchronous correlations $P_{X_1Y_1}$ and $P_{X_2Y_2}$, where the time-window is always the same.

C Agenda Diversity - Normalized Shannon Entropy

The agendas were represented as time-evolving distributions of topics. A time window length of six months with a step size of two months between consecutive windows was selected. For each time window, the agenda diversity was measured with the normalized Shannon entropy H over the topic distribution (Boydstun et al., 2014) (see Eq. (C.1)). H measures the diffusion of the distribution by quantifying the level of information. H increases when the topics are equally distributed (i.e., all topics are equally probable; agenda diversity is high), and H decreases when certain topics dominate (i.e., certain topics have high probability; agenda diversity is low). Thus, identifying local minima of H indicates time periods of low agenda diversity. Building on Tukey's box plot construction, the lower inner fence of H is defined by $\text{LIF} = Q_1 - 1.5 \cdot (Q_3 - Q_1)$, with Q_1 as the lower quartile and Q_3 as the upper quartile. Values of H below the LIF indicate local minima, where agenda diversity is low as certain topics dominate. Sensitivity checks had shown that,

on the one hand, time windows larger than 6 months were not sensitive enough to detect true local minima, and, on the other hand, time windows shorter than 6 months created artificial minima due to lack of data. Therefore, 6 months was considered the appropriate window length for Shannon entropy and Jensen-Shannon distance.

$$H[p] = \frac{-\sum_{i=1}^{T}(p(x_i) * \ln p(x_i))}{\ln T}. \qquad Eq.(C.1)$$

D Agenda Distance - Jensen-Shannon Distance (JSD)

JSD measures the similarity between two distributions. It is a symmetric and smoothed version of the Kullback-Leibler divergence (see Eq. (D.1)). JSD was used to quantified the similarity between topic distributions of all articles (P) and all comments (Q) from a given time period (see Eq. (D.2)). The time period specification is the same as for normalized Shannon entropy. The sum of the normalized distributions of all articles in each time window was computed and re-normalized. This process was repeated for the comments, and the JSD of the two resulting distributions is calculated. If JSD is close to 0, the topic distributions of articles and comments in this time period are very similar. At large values of JSD, the distributions of articles and comments diverge. Such local maxima of JSD were detected for values of JSD above the upper inner fence (UIF) of the box plot construction, which is defined by UIF $= Q_3 + 1.5 \cdot (Q_3 - Q_1)$, analogously to normalized Shannon entropy.

$$KL(P \parallel Q) = \sum P(x) \log \frac{P(x)}{Q(x)}. \qquad Eq.(D.1)$$

$$JSD(P \parallel Q) = \sqrt{\frac{1}{2} KL(P \parallel \frac{P+Q}{2}) + \frac{1}{2} KL(Q \parallel \frac{P+Q}{2})}. \qquad Eq.(D.2)$$

References

Abdi-Herrle, S., 2018. Mediale Themensetzung in Zeiten von Web 2.0. Dissertation. Nomos Verlagsgesellschaft. http://dx.doi.org/10.5771/9783845295909.

Ader, C.R., 1995. A longitudinal study of agenda setting for the issue of environmental pollution. Journal. Mass Commun. Q. 72, 300–311. http://dx.doi.org/10.1177/107769909507200204.

Aitken, R., Watkins, L., Williams, J., Kean, A., 2020. The positive role of labelling on consumers' perceived behavioural control and intention to purchase organic food. J. Clean. Prod. 255, 120334. http://dx.doi.org/10.1016/j.jclepro.2020.120334.

Ajzen, I., 1991. The theory of planned behavior. Organ. Behav. Hum. Decis. Process. 50, 179–211. http://dx.doi.org/10.1016/0749-5978(91)90020-T.

Balducci, B., Marinova, D., 2018. Unstructured data in marketing. J. Acad. Mark. Sci. 46, 557–590. http://dx.doi.org/10.1007/s11747-018-0581-x.

Berger, I.E., Mitchell, A.A., 1989. The effect of advertising on attitude accessibility, attitude confidence, and the attitude-behavior relationship. J. Consum. Res. 16, 269. http://dx.doi.org/10.1086/209213.

Berger, J., Humphreys, A., Ludwig, S., Moe, W.W., Netzer, O., Schweidel, D.A., 2020. Uniting the tribes: Using text for marketing insight. J. Mark. 84, 1–25. http://dx.doi.org/10.1177/0022242919873106.

Bitsch, V., Koković, N., Rombach, M., 2014. Risk communication and market effects during foodborne illnesses: A comparative case study of bacterial outbreaks in the U.S. and in Germany. Int. Food Agribus. Man. 17, 97–114. http://dx.doi.org/10.22004/AG.ECON.183451.

Blei, D., Ng, A., Jordan, M., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

BÖLW, 2020. Branchen Report 2020. Technical Report. Bund Ökologische Lebensmittelwirtschaft e.V.. Berlin.

Boydstun, A.E., Bevan, S., Thomas, H.F., 2014. The importance of attention diversity and how to measure it. Policy Stud. J. 42, 173–196. http://dx.doi.org/10.1111/psj.12055.

Brach, S., Walsh, G., Shaw, D., 2018. Sustainable consumption and third-party certification labels: Consumers' perceptions and reactions. Eur. Manag. J. 36, 254–265. http://dx.doi.org/10.1016/j.emj.2017.03.005.

Cer, D., Yang, Y., yi Kong, S., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.H., Strope, B., Kurzweil, R., 2018. Universal sentence encoder. CoRR abs/1803.11175. https://arxiv.org/pdf/1803.11175.pdf.

Chidambaram, M., Yang, Y., Cer, D., Yuan, S., Sung, Y.H., Strope, B., Kurzweil, R., 2018. Learning cross-lingual sentence representations via a multi-task dual-encoder model. arXiv preprint arXiv:1810.12836 abs/1810.12836. https://arxiv.org/abs/1810.12836.

Collins, A.M., Loftus, E.F., 1975. A spreading-activation theory of semantic processing. Psychol. Rev. 82, 407–428. http://dx.doi.org/10.1037/0033-295X.82.6.407.

Conway-Silva, B.A., Filer, C.R., Kenski, K., Tsetsi, E., 2018. Reassessing Twitter's agenda-building power. Soc. Sci. Comput. Rev. 36, 469–483. http://dx.doi.org/10.1177/0894439317715430.

Danner, H., Menapace, L., 2020. Using online comments to explore consumer beliefs regarding organic food in German-speaking countries and the United States. Food Qual. Prefer. 83, 103912. http://dx.doi.org/10.1016/j.foodqual.2020.103912.

Danner, H., Thøgersen, J., 2021. Does online chatter matter for consumer behaviour? A priming experiment on organic food. Int. J. Consum. Stud. http://dx.doi.org/10.1111/ijcs.12732.

Denham, B.E., 2010. Toward conceptual consistency in studies of agenda-building processes: A scholarly review. Rev. Commun. 10, 306–323. http://dx.doi.org/10.1080/15358593.2010.502593.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 http://arxiv.org/abs/1810.04805.

Fazio, R.H., 1986. How do attitudes guide behavior?, in: Sorrentino, R.M., Higgins, E.T. (Eds.), Handbook of motivation and cognition. Guilford Press, New York, pp. 204–243.

FiBL, 2020. Data on organic agriculture worldwide 2018. the statistics.fibl.org website maintained by the research institute of organic agriculture (fibl). http://statistics.fibl.org/world.html.

Fishbein, M., Ajzen, I., 1975. Belief, attitude, intention and behavior: An introduction to theory and research. Addison-Wesley, Reading.

Gerber, A.S., Karlan, D., Bergan, D., 2009. Does the media matter? A field experiment measuring the effect of newspapers on voting behavior and political opinions. Am. Econ. J. Appl. Econ. 1, 35–52. http://dx.doi.org/10.1257/app.1.2.35.

Ghanem, S., 1997. Filling in the tapestry: The second level of agenda setting, in: McCombs, M.E., Shaw, D.L., Weaver, D.H. (Eds.), Commun. Democr. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 3–14.

Groshek, J., Groshek, M.C., 2014. Agenda trending: Reciprocity and the predictive capacity of social networking sites in intermedia agenda setting across topics over time. Media Commun. 1, 15–27. http://dx.doi.org/10.17645/mac.v1i1.71.

Guo, L., Vargo, C.J., Pan, Z., Ding, W., Ishwar, P., 2016. Big social data analytics in journalism and mass communication. Journal. Mass Commun. Q. 93, 332–359. http://dx.doi.org/10.1177/1077699016639231.

Hagerer, G., Leung, W.S., Danner, H., Groh, G., 2021. A case study and qualitative analysis of simple cross-lingual opinion mining, in: KDIR.

Harris, Z., 1954. Distributional structure. Word 10, 146–162. http://dx.doi.org/https://10.1007/978-94-009-8467-7_1.

He, R., Lee, W.S., Ng, H.T., Dahlmeier, D., 2017. An unsupervised neural attention model for aspect extraction., in: Barzilay, R., Kan, M.Y. (Eds.), ACL (1), Association for Computational Linguistics. pp. 388–397. http://dx.doi.org/10.18653/v1/P17-1036.

Hemmerling, S., Hamm, U., Spiller, A., 2015. Consumption behaviour regarding organic food from a marketing perspective—a literature review. Org. Agric. 5, 277–313. http://dx.doi.org/10.1007/s13165-015-0109-3.

Hester, J.B., Gibson, R., 2003. The economy and second-level agenda setting: A time-series analysis of economic news and public opinion about the economy. Journal. Mass Commun. Q. 80, 73–90. http://dx.doi.org/10.1177/107769900308000106.

Holt, D., Barkemeyer, R., 2012. Media coverage of sustainable development issues - attention cycles or punctuated equilibrium? Sustain. Dev. 20, 1–17. http://dx.doi.org/10.1002/sd.460.

Hossain, N., Krumm, J., Gamon, M., 2019. president vows to cut ¡taxes¿ hair: Dataset and analysis of creative text editing for humorous headlines, 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), pp. 133–142.

Hughner, R.S., McDonagh, P., Prothero, A., Shultz, C.J., Stanton, J., 2007. Who are organic food consumers? A compilation and review of why people

purchase organic food. J. Consum. Behav. 6, 94–110. http://dx.doi.org/10.1002/cb.210.

Jacobi, C., van Atteveldt, W., Welbers, K., 2015. Quantitative analysis of large amounts of journalistic texts using topic modelling. Digit. Journal. 4, 89–106. http://dx.doi.org/10.1080/21670811.2015.1093271.

Junsheng, H., Akhtar, R., Masud, M.M., Rana, M.S., Banna, H., 2019. The role of mass media in communicating climate science: An empirical evidence. J. Clean. Prod. 238, 117934. http://dx.doi.org/10.1016/j.jclepro.2019.117934.

Kenny, D.A., 1975. Cross-lagged panel correlation: A test for spuriousness. Psychol. Bull. 82, 887–903. http://dx.doi.org/10.1037/0033-2909.82.6.887.

Kim, H.K., Kim, H., Cho, S., 2017. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. Neurocomputing 266, 336–352. http://dx.doi.org/10.1016/j.neucom.2017.05.046.

Ksiazek, T.B., 2018. Commenting on the news. Journal. Stud. 19, 650–673. http://dx.doi.org/10.1080/1461670X.2016.1209977.

Ksiazek, T.B., Peer, L., Lessard, K., 2016. User engagement with online news: Conceptualizing interactivity and exploring the relationship between online news videos and user comments. New Media Soc. 18, 502–520. http://dx.doi.org/10.1177/1461444814545073.

Kushwah, S., Dhir, A., Sagar, M., Gupta, B., 2019. Determinants of organic food consumption. A systematic literature review on motives and barriers. Appetite 143, 104402. http://dx.doi.org/10.1016/j.appet.2019.104402.

Le, Q.V., Mikolov, T., 2014. Distributed representations of sentences and documents., in: ICML, pp. 1188–1196.

Lee, E.J., 2012. That's not the way it is: How user-generated comments on the news affect perceived media bias. J. Comput. Mediat. Commun. 18, 32–45. http://dx.doi.org/10.1111/j.1083-6101.2012.01597.x.

Lee, E.J., Kim, H.S., Cho, J., 2017. How user comments affect news processing and reality perception: Activation and refutation of regional prejudice. Commun. Monogr. 84, 75–93. http://dx.doi.org/10.1080/03637751.2016.1231334.

Lee, E.J., Tandoc, E.C., 2017. When news meets the audience: How audience feedback online affects news production and consumption. Hum. Commun. Res. 43, 436–449. http://dx.doi.org/10.1111/hcre.12123.

Lee, H.J., Hwang, J., 2016. The driving role of consumers' perceived credence attributes in organic food purchase decisions: A comparison of two groups of consumers. Food Qual. Prefer. 54, 141–151. http://dx.doi.org/10.1016/j.foodqual.2016.07.011.

Lim, J., 2006. A cross-lagged analysis of agenda setting among online news media. Journal. Mass Commun. Q. 83, 298–312. http://dx.doi.org/10.1177/107769900608300205.

Lim, J., 2011. First-level and second-level intermedia agenda-setting among major news websites. Asian J. Commun. 21, 167–185. http://dx.doi.org/10.1080/01292986.2010.539300.

Lockie, S., 2006. Capturing the sustainability agenda: Organic foods and media discourses on food scares, environment, genetic engineering, and health. Agric. Hum. Values 23, 313–323. http://dx.doi.org/10.1007/s10460-006-9007-3.

Luo, Y., Burley, H., Moe, A., Sui, M., 2019. A meta-analysis of news media's public agenda-setting effects, 1972-2015. Journal. Mass Comm. Q. 96, 150–172. http://dx.doi.org/10.1177/1077699018804500.

Lyu, F., Choi, J., 2020. The forecasting sales volume and satisfaction of organic products through text mining on web customer reviews. Sustainability 12, 4383. http://dx.doi.org/10.3390/su12114383.

Mahlau, G., 1999. Das Image der Landwirtschaft. M. Wehle, Bonn.

McCombs, M.E., 2014. Setting the agenda: The mass media and public opinion. Polity Press, Cambridge.

McCombs, M.E., Shaw, D.L., 1972. The agenda-setting function of mass media. Public Opin. Q. 36, 176–187. http://dx.doi.org/10.1086/267990.

Meyers, C., Abrams, K., 2010. Feeding the debate: a qualitative framing analysis of organic food news media coverage. J. Appl. Commun. 94, 22–37.

Meza, X.V., Park, H.W., 2016. Organic products in Mexico and South Korea on Twitter. J. Bus. Ethics 135, 587–603. http://dx.doi.org/10.1007/s10551-014-2345-y.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality, in: Burges, C.J.C., Bottou, L., Ghahramani, Z., Weinberger, K.Q. (Eds.), Advances in neural information processing systems, pp. 3111–3119.

Müller, C.E., Gaus, H., 2015. Consumer response to negative media information about certified organic food products. J. Consumer Policy 38, 387–409. http://dx.doi.org/10.1007/s10603-015-9299-z.

Neuman, W.R., Guggenheim, L., Mo Jang, S., Bae, S.Y., 2014. The dynamics of public attention: Agenda-setting theory meets big data. J. Commun. 64, 193–214. http://dx.doi.org/10.1111/jcom.12088.

Olson, E.L., 2017. The rationalization and persistence of organic food beliefs in the face of contrary evidence. J. Clean. Prod. 140, 1007–1013. http://dx.doi.org/10.1016/j.jclepro.2016.06.005.

Pinto, S., Albanese, F., Dorso, C., Balenzuela, P., 2019. Quantifying time-dependent media agenda and public opinion by topic modeling. Physica A 524, 614–624.

Roberts, M., Wanta, W., Dzwo, T.H., 2002. Agenda setting and issue salience online. Comm. Res. 29, 452–465. http://dx.doi.org/10.1177/0093650202029004004.

Rodríguez-Bermúdez, R., Miranda, M., Orjales, I., Ginzo-Villamayor, M.J., Al-Soufi, W., López-Alonso, M., 2020. Consumers' perception of and attitudes towards organic food in Galicia (Northern Spain). Int. J. Consum. Stud. 44, 206–219. http://dx.doi.org/10.1111/ijcs.12557.

Rogosa, D., 1980. A critique of cross-lagged correlation. Psychol. Bull. 88, 245–258. http://dx.doi.org/10.1037/0033-2909.88.2.245.

Rozelle, R.M., Campbell, D.T., 1969. More plausible rival hypotheses in the cross-lagged panel correlation technique. Psychol. Bull. 71, 74–80. http://dx.doi.org/10.1037/h0026863.

Santana, A.D., 2011. Online readers' comments represent new opinion pipeline. Newsp. Res. J. 32, 66–81. http://dx.doi.org/10.1177/073953291103200306.

Schmitt, M., Schuller, B.W., 2017. openxbow - introducing the passau open-source crossmodal bag-of-words toolkit. J. Mach. Learn. Res. 18, 96:1–96:5.

Shrimali, V., 2018. Universal sentence encoder. public website. `https://www.learnopencv.com/universal-sentence-encoder/`. (accessed 15 June 2020).

Siegrist, M., Visschers, V.H., Hartmann, C., 2015. Factors influencing changes in sustainability perception of various food behaviors: Results of a longitudinal study. Food Qual. Prefer. 46, 33–39. http://dx.doi.org/10.1016/j.foodqual.2015.07.006.

Sridhar, V.K.R., 2015. Unsupervised topic modeling for short texts using distributed representations of words., in: Blunsom, P., Cohen, S.B., Dhillon, P.S., Liang, P. (Eds.), VS@HLT-NAACL, The Association for Computational Linguistics. pp. 192–200. http://dx.doi.org/10.3115/v1/W15-1526.

Sweetser, K.D., Golan, G.J., Wanta, W., 2008. Intermedia agenda setting in television, advertising, and blogs during the 2004 election. Mass Comm. Soc. 11, 197–216. http://dx.doi.org/10.1080/15205430701590267.

Takeshita, T., 2006. Current critical problems in agenda-setting research. Int. J. Public Opin. Res. 18, 275–296. http://dx.doi.org/10.1093/ijpor/edh104.

Tausczik, Y.R., Pennebaker, J.W., 2010. The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. 29, 24–54. http://dx.doi.org/10.1177/0261927X09351676.

Testa, F., Sarti, S., Frey, M., 2019. Are green consumers really green? Exploring the factors behind the actual consumption of organic food products. Bus. Strategy Environ. 28, 327–338. http://dx.doi.org/10.1002/bse.2234.

Thøgersen, J., 2006. Media attention and the market for 'green' consumer products. Bus. Strategy Environ. 15, 145–156. http://dx.doi.org/10.1002/bse.521.

Thøgersen, J., 2010. Country differences in sustainable consumption: The case of organic food. J. Macromarketing 30, 171–185. http://dx.doi.org/10.1177/0276146710361926.

Thøgersen, J., Ölander, F., 2006. To what degree are environmentally beneficial choices reflective of a general conservation stance? Environ. Behav. 38, 550–569. http://dx.doi.org/10.1177/0013916505283832.

Trivedi, R.H., Patel, J.D., Acharya, N., 2018. Causality analysis of media influence on environmental attitude, intention and behaviors leading to green purchasing. J. Clean. Prod. 196, 11–22. http://dx.doi.org/10.1016/j.jclepro.2018.06.024.

Tukker, A., Cohen, M.J., Hubacek, K., Mont, O., 2010. The impacts of household consumption and options for change. J. Ind. Ecol. 14, 13–30. http://dx.doi.org/10.1111/j.1530-9290.2009.00208.x.

Vittersø, G., Tangeland, T., 2015. The role of consumers in transitions towards sustainable food consumption. the case of organic food in norway. J. Clean. Prod. 92, 91–99. http://dx.doi.org/10.1016/j.jclepro.2014.12.055.

Wanta, W., King, P.t., McCombs, M.E., 1995. A comparison of factors influencing issue diversity in the U.S. and Taiwan. Int. J. Public Opin. Res. 7, 353–365. http://dx.doi.org/10.1093/ijpor/7.4.353.

Weaver, D.H., 2007. Thoughts on agenda setting, framing, and priming. J. Commun. 57, 142–147. http://dx.doi.org/10.1111/j.1460-2466.2006.00333.x.

Winter, S., Brückner, C., Krämer, N.C., 2015. They came, they liked, they commented: Social influence on Facebook news channels. Cyberpsychol. Behav. Soc. Netw. 18, 431–436. http://dx.doi.org/10.1089/cyber.2015.0005.

Xing, C., Wang, D., Zhang, X., Liu, C., 2014. Document classification with distributions of word vectors., in: APSIPA, IEEE. pp. 1–5. http://dx.doi.org/10.1109/APSIPA.2014.7041633.

Yadavalli, A., Jones, K., 2014. Does media influence consumer demand? The case of lean finely textured beef in the United States. Food Policy 49, 219–227. http://dx.doi.org/10.1016/j.foodpol.2014.08.002.

Yang, Y., Abrego, G.H., Yuan, S., Guo, M., Shen, Q., Cer, D., Sung, Y.H., Strope, B., Kurzweil, R., 2018. universal-sentence-encoder-xling/en-de — english and german language-agnostic text encoder. online. `https://tfhub.dev/google/universal-sentence-encoder-xling/en-de/1`. (accessed 15 June 2020).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., Le, Q.V., . Xlnet: Generalized autoregressive pretraining for language understanding. Adv. Neural Inf. Process. Syst. 32, abs/1906.08237.

Ziegele, M., Breiner, T., Quiring, O., 2014. What creates interactivity in online news discussions? An exploratory analysis of discussion factors in user

comments on news items. J. Commun. 64, 1111–1138. http://dx.doi.org/10.1111/jcom.12123.

Topic Name	US Top Words
Food Safety & Chemicals & GMO	gmo, chemical, cancer, pesticide, organic, study, antibiotic, safe, fda, cause
Food Products & Quality	cheese, bread, cook, taste, tomato, recipe, bean, sauce, salad, fresh
Health & Nutrition	food, eat, diet, healthy, fat, sugar, nutrition, health, calorie, lunch
Environment & Climate Change & Energy	water, energy, carbon, climate, gas, heat, fuel, warming, air, emission
Farming	farmer, organic, crop, farming, agriculture, grow, soil, plant, pesticide, land
Animal Welfare & Meat Consumption	meat, animal, eat, vegetarian, vegan, beef, cow, chicken, feed, kill
Retailers & Prices	store, amazon, grocery, price, shop, company, product, market, customer, sell
Economy & Affordability	money, pay, tax, cost, profit, government, rich, school, care, income
Politics	trump, vote, political, government, republicans, president, party, conservative, democratic, obama
Evidence	science, article, study, read, fact, belief, cultural, religion, truth, evidence

Table 1: The most representative words (top words) for each topic in the US media and public agendas.

Topic Name	GER Top Words
Food Safety & Chemicals & GMO	dioxin, bio, bakterien, antibiotika, erreger, ehec, enthalten, gentechnik, gifte, krebs
Food Products & Quality	schmecken, käse, essen, kochen, brot, tomaten, milch, analogkäse, gurken, frisch
Health & Nutrition	lebensmittel, essen, ernährung, fett, gesund, nahrungsmittel, lebensmitteln, kalorien, zucker, ungesund
Environment & Climate Change & Energy	co2, erde, grad, energien, luft, wasser, erneuerbare, windkraft, atmosphäre, e10
Farming	landwirtschaft, bauern, bio, pflanzen, ökologisch, landwirte, anbau, konventionell, saatgut, dnger
Animal Welfare & Meat Consumption	fleisch, tiere, vegetarier, essen, hühner, massentierhaltung, kuh, fleischkonsum, futter, veganer
Retailers & Prices	verbraucher, kaufen, produkte, kunden, aldi, supermarkt, lebensmittel, produkt, qualität, ware
Economy & Affordability	geld, euro, zahlen, mittelschicht, steuern, bezahlen, einkommen, staat, kosten, mehr
Politics	spd, grün, politik, fdp, cdu, link, partei, merkel, deutsch, politiker
Evidence	verseuchung, menschen, schuld, aussagen, verstehen, lesen, diskussion, tragen, glauben, thema

Table 2: The most representative words (top words) for each topic in the German media and public agendas.

Metric	Relationship between Media and Public Agendas
Pearson Correlation ρ_t	Linear relationship between the relative proportions of a topic in an article and its comments
Cross-Lagged Correlations $P_{X_1Y_2}$, $P_{Y_1X_2}$	Linear relationships between the relative proportions of a topic in an article and the comments of future articles, as well as the comments of an article and future articles
Normalized Shannon entropy H	Agenda diversity, i.e., the diffusion of the relative topic distribution in an agenda over time. Local minima indicate time periods of low agenda diversity, i.e., certain topics dominate due to specific events.
Jensen-Shannon distance (JSD)	Agenda distance, i.e., the similarity between the relative distribution of topics in the media and public agenda. Local maxima indicate time periods where the distance between media and public agendas is high

Table 3: The four agenda-setting metrics and how they measure the relationship between the media and public agendas.

Topic	US	GER
Food Safety & Chemicals & GMO	0.712	0.773
Food Products & Quality	0.785	0.913
Health & Nutrition	0.808	0.690
Environment & Climate Change & Energy	0.626	0.764
Farming	0.759	0.754
Animal Welfare & Meat Consumption	0.781	0.757
Retailers & Prices	0.783	0.668
Economy & Affordability	0.679	0.765
Politics	0.711	0.817
Evidence	0.540	0.423
Total correlation	0.751	0.761

Table 4: Topic correlations ρ_t in media and public agenda for the US and Germany. All correlations are statistically significant ($p < 0.001$).

Topic	$P_{X_1Y_2}$		$P_{Y_1X_2}$		RCB
Food Safety & Chemicals & GMO	0.372	***	0.105	ns	0.209
Food Products & Quality	0.206	*	0.023	ns	0.153
Health & Nutrition	0.359	***	0.119	ns	0.106
Environment & Climate Change & Energy	0.180	ns	0.002	ns	0.101
Farming	0.189	*	-0.157	ns	0.148
Animal Welfare & Meat Consumption	0.314	***	-0.021	ns	0.124
Retailers & Prices	0.340	***	0.062	ns	0.216
Economy & Affordability	0.035	ns	-0.001	ns	0.026
Politics	0.284	**	0.006	ns	0.162
Evidence	0.256	**	0.175	ns	0.033
Total correlation	0.375	***	0.158	***	0.210

Table 5: US Cross-lagged correlations between articles in $t1$ and comments in $t2$ ($P_{X_1Y_2}$), and between comments in $t1$ and articles in $t2$ ($P_{Y_1X_2}$). Rozelle-Campbell Baseline (RCB). Statistical significance is denoted with * ($p<0.05$), ** ($p<0.01$), and *** ($p<0.001$).

Topic	$P_{X_1Y_2}$		$P_{Y_1X_2}$		RCB
Food Safety & Chemicals & GMO	0.187	***	0.012	ns	0.181
Food Products & Quality	0.152	***	-0.018	ns	0.230
Health & Nutrition	0.204	***	-0.010	ns	0.137
Environment & Climate Change & Energy	0.566	***	-0.013	ns	0.217
Farming	0.249	***	0.040	ns	0.138
Animal Welfare & Meat Consumption	0.402	***	-0.022	ns	0.182
Retailers & Prices	0.220	***	0.132	ns	0.097
Economy & Affordability	0.055	***	-0.154	ns	0.042
Politics	0.267	***	0.072	ns	0.190
Evidence	0.003	***	-0.016	ns	0.071
Total correlation	0.211	***	0.008	***	0.168

Table 6: German Cross-lagged correlations between articles in $t1$ and comments in $t2$ ($P_{X_1Y_2}$), and between comments in $t1$ and articles in $t2$ ($P_{Y_1X_2}$). Rozelle-Campbell Baseline (RCB). Statistical significance is denoted with * (p<0.05), ** (p<0.01), and *** (p<0.001).

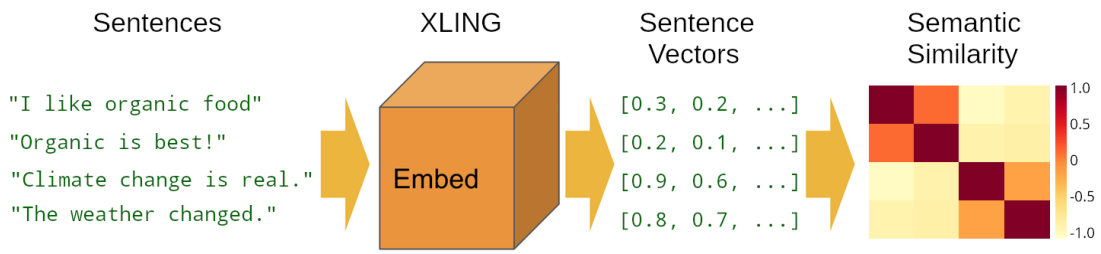


Figure 1: **Sentence embeddings**. Illustrative example of how sentence embeddings are generated using XLING. Inspired by Cer et al. (2018) and Shrimali (2018).

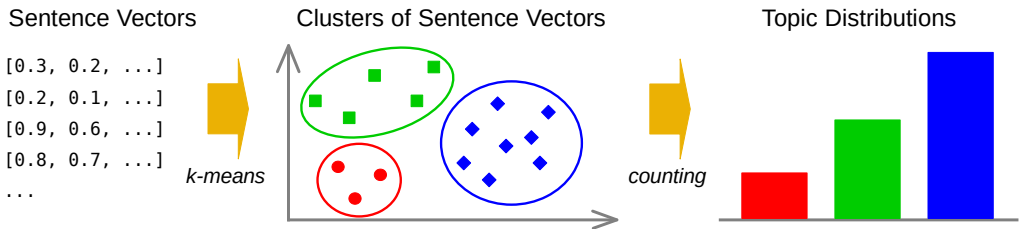


Figure 2: **Topic modeling based on clustering of pre-trained sentence embeddings**. Topic distributions are calculated for all sentences of a document. Here, a document is either a news article or the aggregated comments of a news article.

Figure 3: **Average topic distributions (%) in US and German media and public agenda.**

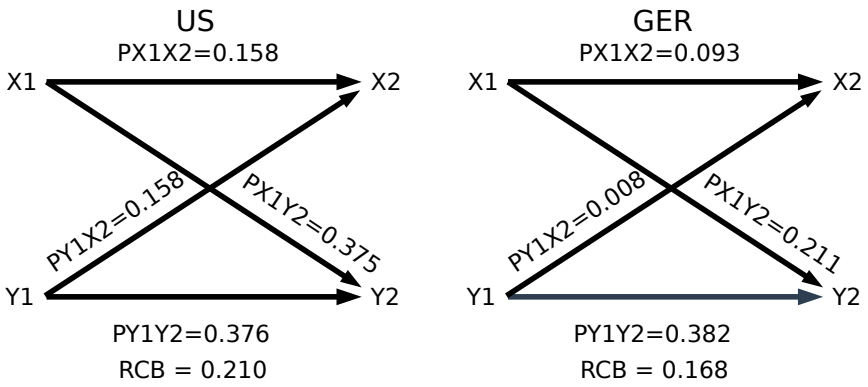


Figure 4: **Total cross-lagged correlations.** Cross-lagged correlations between articles and comments ($P_{X_1Y_2}$), and between comments and articles ($P_{Y_1X_2}$) of consecutive weeks for the US and Germany. Autocorrelations are given among articles ($P_{X_1X_2}$) and among comments ($P_{Y_1Y_2}$). All correlations are statistically significant at $p < 0.001$, except $P_{X_1X_2}$ for Germany, which is significant at $p < 0.05$. Rozelle-Campbell Baseline (RCB). Own illustration according to Kenny (1975).

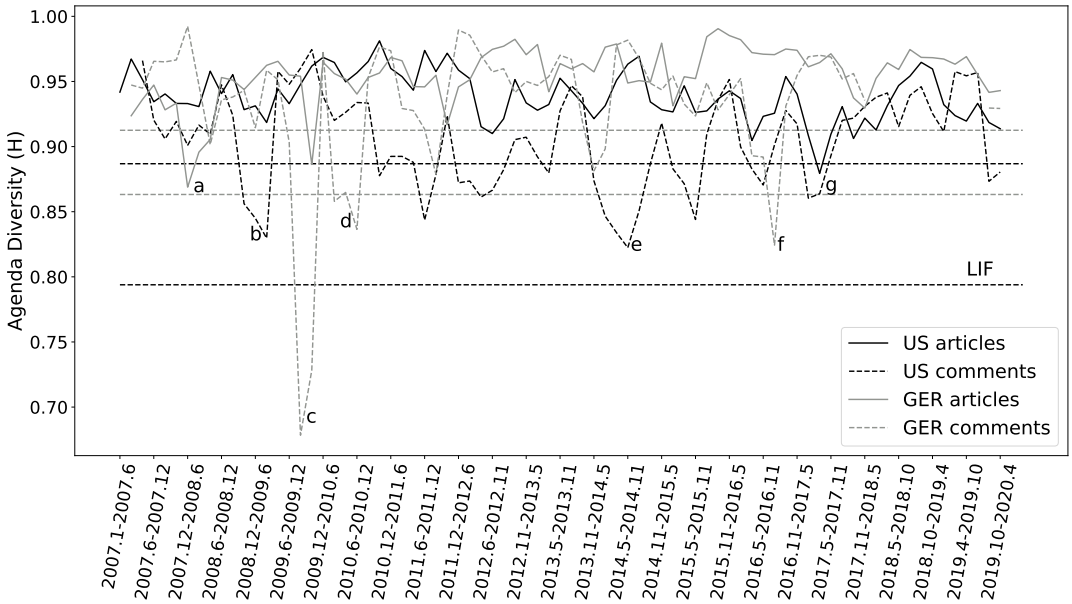


Figure 5: **Agenda diversity over time.** Measured by normalized Shannon entropy (H) for the US and German media and public agendas. The horizontal lines point out the lower inner fences (LIF) to identify local minima, i.e., time periods with low agenda diversity. These time periods are denoted with *a*, *b*, *c*, *d*, *e*, *f*, and *g*.

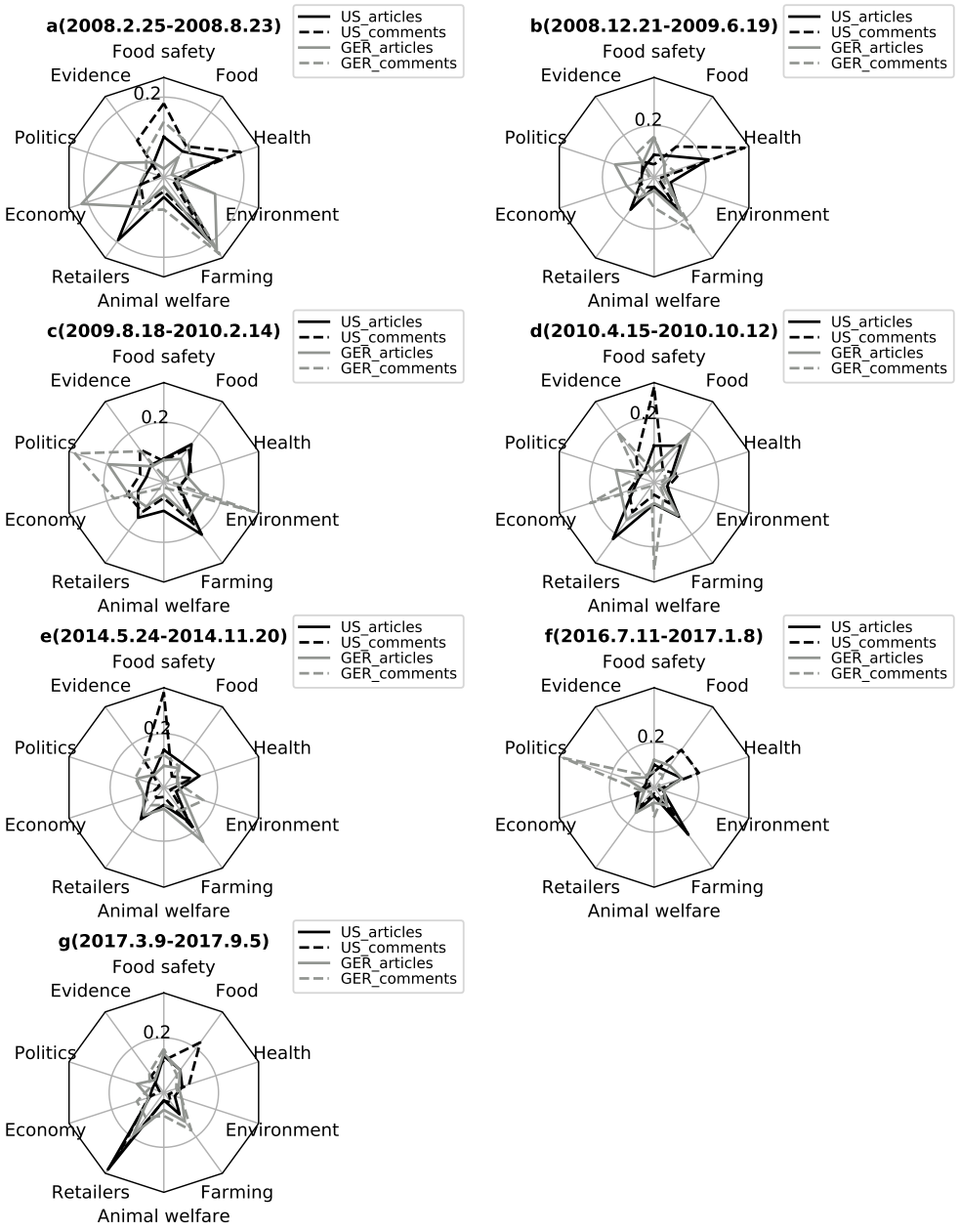


Figure 6: **Topic distributions.** Radar plots for the relative topic distribution in time periods a, b, c, d, e, f, and g identified via normalized Shannon entropy and JSD.

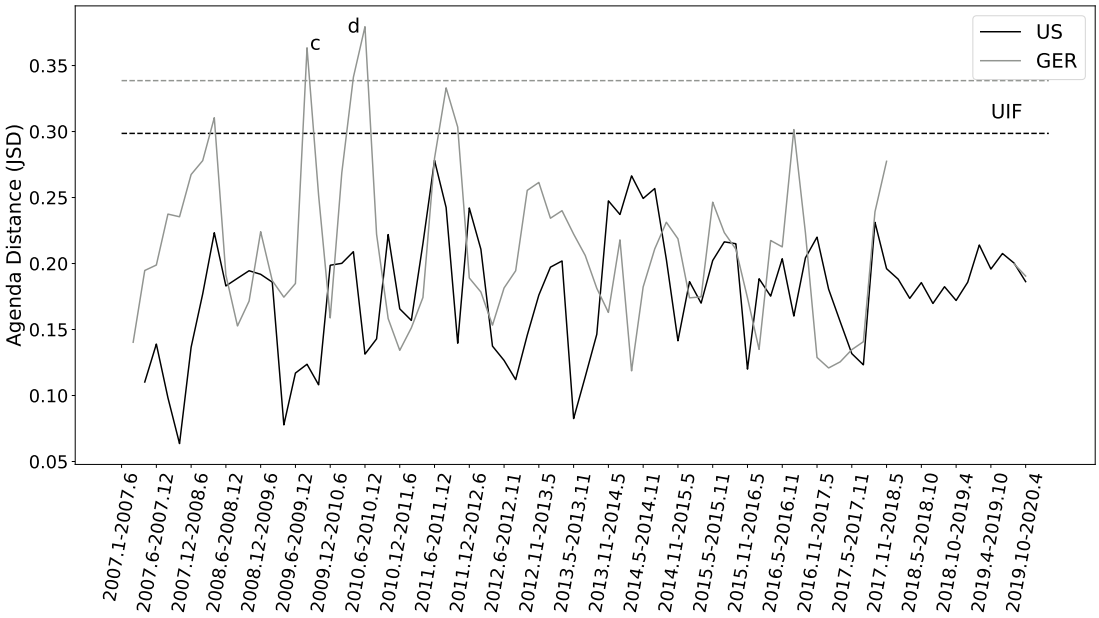


Figure 7: **Agenda distance.** Distance between media and public agenda for the US and Germany over time measured by Jensen-Shannon distance. The horizontal lines point out upper inner fences (UIF) to identify local maxima, i.e., time periods with high agenda distance. These time periods are denoted with *c* and *d*, and coincide with the time periods depicted in the graph and radar plots of Figure 6. The horizontal lines point out the lower inner fences (LIF) to identify local minima, i.e., time periods with low agenda diversity.

A.3 Combining Content Analysis and Neural Networks to Analyze Discussion Topics in Online Comments About Organic Food

The publication on the consecutive pages was accepted after peer-review as full paper at the 2020 3rd International Conference on Advanced Research Methods and Analytics. Gerhard Johann Hagerer, the author of the present thesis, is the second author of that paper. His author role, the reprinting permission, and his following contributions are acknowledged by all authors Danner, Hagerer, Kasischke, and Groh [2020]:

> *"Gerhard Johann Hagerer supervised experimental part of the research project. He developed the experimental methods of the paper. Furthermore, he directed the implementation process and reviewed the source code deeply. Regarding the writing of the paper, he contributed textual parts and paraphrased, corrected, combined, and otherwise improved drafted material."*

Publication Summary

After Hagerer et al. [2020b], this paper was the first application of the Universal Sentence Encoder (USE) onto our organic food social media dataset. This pilot study indicates correlations between the USE features and the annotations given by a domain expert in a previous qualitative content study on our data. The correlation is observed on small annotated data samples as well as on larger, not annotated samples. We conclude that the USE features extrapolate onto larger datasets without annotations and maintain a correlation to expert annotations. For the development of our research, it is important to see features which bear relation with the observations from our domain expert. As such, this was another groundwork for our main papers Hagerer et al. [2021d] and Danner et al. [2022], where the related Universal Sentence Encoder Cross-Lingual (XLING) is used successfully for opinion mining in qualitative content studies.

3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11632

Combining content analysis and neural networks to analyze discussion topics in online comments about organic food

Hannah Danner[1], Gerhard Hagerer[2], Florian Kasischke[2], Georg Groh[2]
[1]TUM School of Management, Technical University of Munich, Germany, [2]TUM Department of Informatics, Technical University of Munich, Germany.

Abstract

Consumers increasingly share their opinions about products in social media. However, the analysis of this user-generated content is limited either to small, in-depth qualitative analyses or to larger but often more superficial analyses based on word frequencies. Using the example of online comments about organic food, we investigate the relationship between qualitative analyses and latest deep neural networks in three steps. First, a qualitative content analysis defines a class system of opinions. Second, a pre-trained neural network, the Universal Sentence Encoder, analyzes semantic features for each class. Third, we show by manual inspection and descriptive statistics that these features match with the given class structure from our qualitative study. We conclude that semantic features from deep pre-trained neural networks have the potential to serve for the analysis of larger data sets, in our case on organic food. We exemplify a way to scale up sample size while maintaining the detail of class systems provided by qualitative content analyses. As the USE is pre-trained on many domains, it can be applied to different domains than organic food and support consumer and public opinion researchers as well as marketing practitioners in further uncovering the potential of insights from user-generated content.

Keywords: *deep neural networks; natural language processing; consumer research; content analysis; social media; organic food.*

1. Introduction

Novel communication technologies sparked the desire of users to publicly share opinions on online platforms (Ziegele et al., 2014). These developments provide an increasing amount of user-generated content, such as online user comments, which can be exploited by marketing and consumer research to gain insights into consumer thinking (Balducci & Marinova, 2018). Beginning with Kozinets' (2002) netnography of online communities, social scientists have increasingly analyzed textual user-generated content with established methods such as content analysis (Krippendorff, 2019). However, due to time and human resources required, such qualitative analyses are limited to small data samples. More recently, advances in automated text analysis and data collection enable consumer researchers to efficiently analyze larger datasets in a short amount of time and facilitate the detection of patterns, and compare measurements over time or between datasets. For an overview of methods see Berger et al., 2020). Frequently employed methods are dictionary-based approaches (e.g., LIWC, Tausczik & Pennebaker, 2010) relying on word frequencies. Researchers using automated text analysis have started to incorporate methods from the field of natural language processing (NLP, such as of data-mining, data-preprocessing, simple classifiers, and topic models (Latent Dirichlet Allocation, Blei, 2012) (for an overview see Vidal et al., 2018). However, to the best our knowledge, there has been little research on how qualitative and NLP methods can be combined fruitfully. Latest advances in NLP are neural networks that account for the semantic context of words, i.e., word embeddings (Mikolov et al., 2013), or sentences, i.e., sentence embeddings (Cer et al., 2018). In this paper, we explore how such embeddings particularly lend themselves to be combined with qualitative text analysis by matching the analysis-depth of the latter with the scope of pre-trained sentence embeddings. In three steps, we present a novel approach for how a qualitative content analysis can be combined and enhanced with deep neural networks for semantic similarity.

We apply the approach to the case of organic food. Not only is a growing share of consumers aware of and buys organic food (Hemmerling et al., 2015)—making it an increasingly important consumer research topic—, consumers also voice their opinions about organic food online (Danner & Menapace, 2020; Meza & Park, 2016; Olson, 2017). The analysis of online user-generated content can thus deliver valuable insights into which product attributes and related topics matter to consumers and what could be potential purchase drivers and barriers.

2. Methodology

In step 1 of our approach, a qualitative text analysis is conducted to develop a class system and manually classify a dataset of interest. In Step 2, we use semantic features from pre-trained neural networks to investigate the semantic characteristics and the respective frequencies for each class. Step 3 presents criteria to combine results of both methods.

2.1. Step 1 – Qualitative Analysis

To exemplify the approach, for step 1, we draw on a recent qualitative content analysis by Danner and Menapace (2020) of online comments about organic food. They manually extracted and classified consumer opinions (referred to as beliefs) about organic food to understand consumers' perception of organic. The authors collected 1069 online comments about organic food from high-coverage US news websites (e.g., nytimes.com, washingtonpost.com) and forums (e.g., reddit.com, quora.com). The 1069 comments consisted of 5510 sentences. Among these 5510 sentences, the two coders identified 1065 containing belief statements about organic food and subsequently classified those belief statements into 64 belief classes and 21 superordinate topics. For example, the sentence stated by a commenter *organic farming is better for nature* was attributed to the belief class *organic farming protects the environment*, which in turn was attributed to the topic class *environment*. By counting the frequencies of belief statements per category, the authors presented a detailed picture of topics salient to the online commenters in the data.

2.2. Step 2 – Universal Sentence Encoder

Using the same data and class system as in step 1, we find similar sentences for each class using the Universal Sentence Encoder (USE). USE is a recent advance in NLP and deep learning (Cer et al., 2018). Its architecture is based on the widely adopted Transformer architecture (Vaswani et al., 2017). USE is a deep neural network model pre-trained on large scale text corpora from many domains. From there, the statistical knowledge in terms of generalizable, intermediate, semantic vector representations, which are also referred to as features or embeddings, can be used to quantify the semantics of specific domains, here organic food. USE works on sentence level providing sentence embeddings. The semantics of a given sentence are expressed by its vector representation. When compared to other sentences, the cosine similarity ranges between 1 (similar) to -1 (dissimilar).

We applied USE to automatically find semantically similar sentences for each of the 64 beliefs identified by Danner and Menapace (2020) (e.g., *organic farming protects the environment*) (Table 1). First, USE transformed each of the 64 beliefs and the 5510 sentences into an embedding. Second, USE measured the cosine similarity, i.e. the angular distance, between the embedding of each of the 64 beliefs (also referred to as seed sentences) and each of the 5510 sentences. When choosing a low threshold level for cosine similarity (i.e., the closer to -1), many sentences are considered as similar, whereas at high levels fewer sentences are considered as similar.

2.3. Step 3 - Evaluation

Eventually, we determine the appropriate level of semantic similarity, i.e., the respective cosine similarity threshold level which yields similar frequencies compared to the qualitative

content analysis as reference. To this end, we inspect the thresholding results for cosine similarity levels from 0.7 to 0.84 based on the following criteria. (1) In the content analysis, 1065 sentences were relevant as in containing beliefs about organic food. A meaningful sentence filtering should yield a similar amount of relevant sentences. (2) The number of sentences assigned into the different classes should be similar for both methods. Therefore, we inspected the relative class frequencies and also calculated the Pearson correlation between the class frequencies for different cosine similarity levels. Figure 1 displays a trade-off between semantic similarity and class frequencies: the lower the cosine similarity (i.e., the less similar the sentences), the higher the correlation between the two methods. (3) Manual inspection should confirm the semantic cohesion between the manually and the automatically assigned sentences. Note that we performed the evaluation at topic level (21 topic classes) as the 64 belief classes are very detailed and in part semantically too similar (e.g., *organic farming is better for the environment* and *conventional farming harms the environment*).

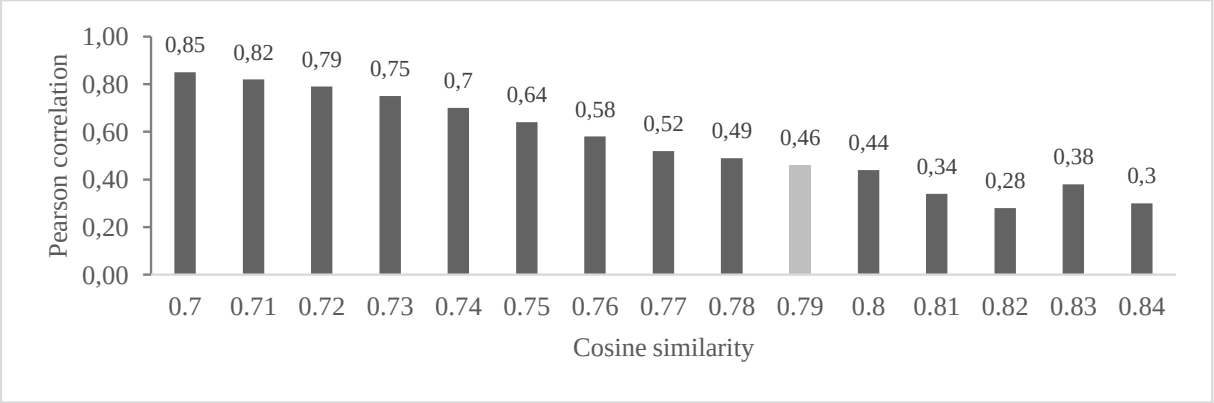


Figure 1. Pearson correlation of class frequencies (21 topic classes) between content analysis and USE. Source: own illustration.

3. Results

Applying the aforementioned evaluation criteria, the thresholding performed best at a cosine similarity of 0.79. (1) At this level of similarity, USE found 1376 relevant sentences, which roughly corresponds to the 1065 relevant sentences identified in the manual analysis. (2) As highlighted in Figure 1, for cosine similarity of 0.79, both methods yielded similar class frequencies, indicated by a correlation of r = 0.46. However, class frequencies do not match perfectly. Looking at the relative class frequencies for each of the 21 topic classes in Figure 2, we find that the class frequencies for both methods are more similar for some topics than for others. For example, the topic *environment* accounts for 11% of sentences in the content analysis and 18% in the similarity thresholding. The most frequent topics in the content analysis were *system integrity*, *food safety*, *environment*; the most frequent topics in USE were *environment*, *system integrity*, *farmer welfare*.

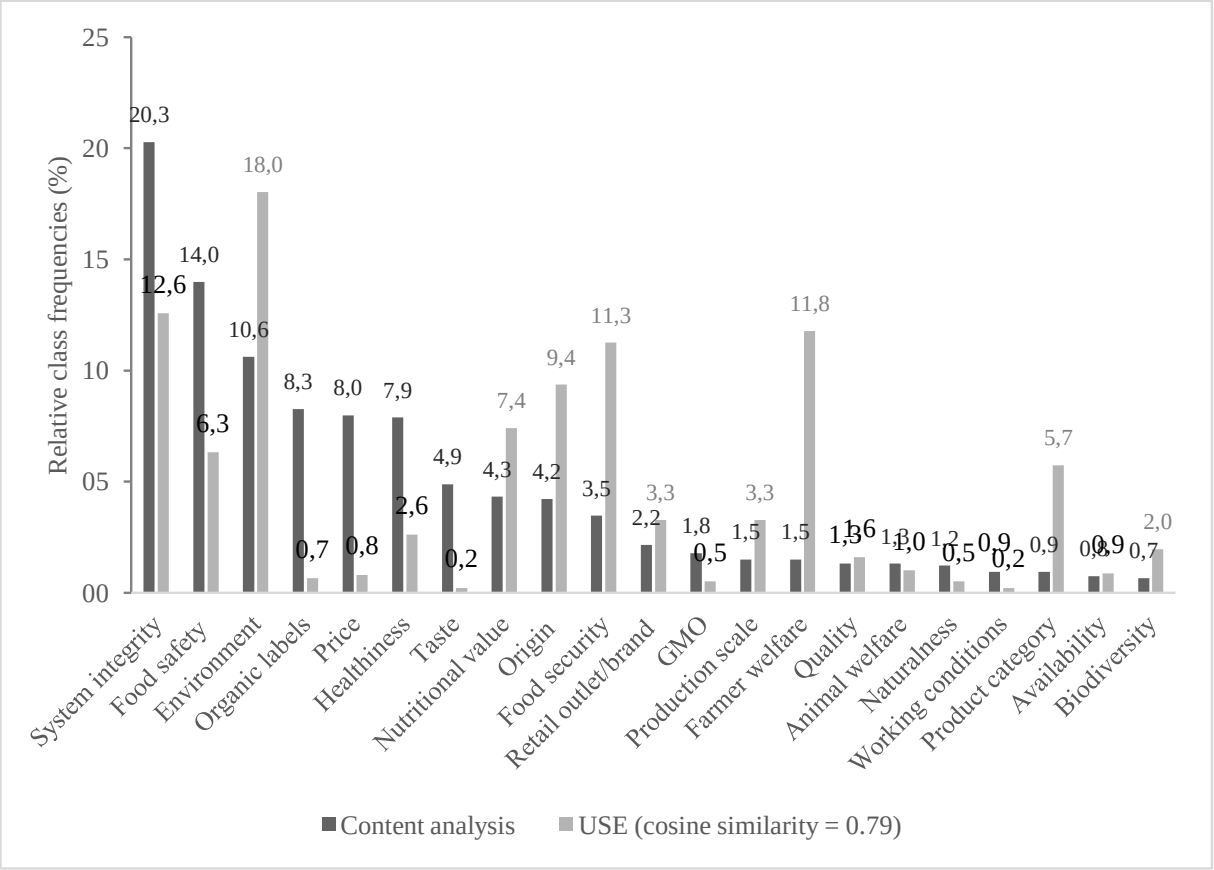


Figure 2. Relative class frequencies (21 topic classes) in content analysis and USE.
Topics are ordered in descending frequency according to the content analysis. Source: own illustration.

Table 1. A seed sentence from content analysis and the 11 sentences identified as similar by USE (cosine similarity = 0.79).

seed	Organic farming protects the environment.
1	Organic farming can help to preserve our environment for future generations.
2	The depletion of the soil and monoculture is what causes factory farming produce to be less nutritious than organic.
3	Mythbusting 101: Organic Farming > Conventional Agriculture
4	A lot of what I've read has said that organic farming is not better for the environment.
5	Organic is for the environment.
6	And from this we hear that organic farming is "devastating" to the environment.
7	Organic farming is much closer to the way Mother Nature farms.
8	GMOs can be super beneficial - to the consumer, the farmer, the environment.
9	Organic farming is greener
10	Besides delivering health benefits, organic farming is better for the environment.
11	Organic is for the environment.

Source: own illustration.

(3) For cosine similarity of 0.79, manual inspection showed very high semantic cohesion between the seed sentences per topic and the sentences identified as similar by USE. Table 1 displays the 11 sentences that USE found to be similar to the belief *organic farming protects the environment* at a cosine similarity of 0.79. All 11 are concerned with the effect of organic farming on the environment. However, sentences 3, 4, and 6 carry negative and thus the sentiment opposite to the seed sentence. Thus, while USE correctly identifies the topic, the sentiment is not always correctly classified, which is one reason why comparisons at topic level were chosen for this study. In addition, the manual inspection of the sentences classified by both methods proved that both methods classified largely the same sentences in the respective classes.

4. Discussion

USE appears to be an effective and easy to use method to analyze large text corpora by searching for sentences that are semantically similar to seed sentences of interest. Seed sentences can originate, for instance, from a small-scale qualitative study—here the belief classes identified by Danner & Menapace (2020). Provided a manually developed class system, it can analyze any unseen dataset, —here 5510 sentences on organic food—,

according to semantic similarity. In the present example, a human researcher selected the required level of similarity by evaluating the features generated by USE based of descriptive statistics and manual inspection. We suggested several criteria to select the appropriate similarity level as an alternative to training a classifier. Training a reliable classifier to classify fine-grained classes as complex as 64 different organic food beliefs requires large amounts of labeled data, which often exceed the resources of common research projects in the field of consumer and opinion research, and as it also applied to the presented example.

The selected similarity threshold was valid as the filtered sentences were widely coherent with the qualitative content analysis. In a subsequent step, USE could be applied to filter a larger unseen data set on organic food. Thus, the potential of the suggested approach lies in its scalability. We can extrapolate the detail of insight characteristic of qualitative research to analyze class frequencies in a larger data set of user-generated content.

Being still in an early phase, our approach bears potential for further refinements. We used a very large class system with 64 belief classes grouped into 21 topics, which also contained classes semantically very similar to each other. Using fewer and more distinct classes could thus improve the coherence between a manual classification and automatic classification based on USE. Furthermore, USE reliably finds the sentences containing similar topics, but does not always correctly distinguish positive and negative sentiment regarding the topic. Therefore, while suitable for topic classification, its use for sentiment analysis is bound to the manual control of a human researcher and domain expert. The imperfect match between manual classification and automatic filtering may also originate from the selection of the unit of analysis, a well-discussed issue in qualitative research (Campbell et al., 2013). The unit of analysis in USE are sentences, whereas in the content analysis, the unit of analysis could also stretch beyond a single sentence, and qualitative researchers can use domain knowledge for understanding and classifying text.

5. Conclusion

In a three-step approach, we suggested how a topic classification of a qualitative content analysis—here of online comments about organic food—can be combined with neural networks like USE to find similar sentences. We proved that embedding techniques largely fit the results of qualitative analysis and point out their methodological potential. USE considers the semantic coherence between words and sentences and delivers in-depth insights by providing the original consumer phrasings (see Table 1) instead of abstract word lists and word frequencies as in more simple approaches of automated text analysis, such as dictionary-based approaches or LDA topic modeling.

Additional potential lies in cross-lingual applications using multilingual USE: Researchers can use the same seed sentences in one language and analyze data sets in different languages

to make cross-country comparisons. Analyzing user-generated content, consumer researchers can learn about which product attributes and topics salient to consumers and potentially serve as purchase drivers or barriers. Based on this, consumer typologies and clusters can be derived. An improved understanding of consumers' opinions can support the design of organic products as well as labeling policies. Another application of USE lies in using items of established scales from survey research as seed sentences and analyze their similarity and prevalence in social media data. In addition, the suggested approach could be promising for market monitoring based on the targeted detection of social media content. For example, social media managers can observe the prevalence and development of certain opinions over time.

References

Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*, *46*(4), 557–590. https://doi.org/10.1007/s11747-018-0581-x

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the Tribes: Using Text for Marketing Insight. *Journal of Marketing*, *84*(1), 1–25. https://doi.org/10.1177/0022242919873106

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding In-depth Semistructured Interviews. *Sociological Methods & Research*, *42*(3), 294–320. https://doi.org/10.1177/0049124113500475

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., . . . Kurzweil, R. (2018). Universal Sentence Encoder. *ArXiv*. Retrieved from http://arxiv.org/pdf/1803.11175v2

Danner, H., & Menapace, L. (2020). Using Online Comments to Explore Consumer Beliefs Regarding Organic Food in German-Speaking Countries and the United States. *Food Quality and Preference*, *83*(103912). https://doi.org/10.1016/j.foodqual.2020.103912

Hemmerling, S., Hamm, U., & Spiller, A. (2015). Consumption behaviour regarding organic food from a marketing perspective—a literature review. *Organic Agriculture*, *5*(4), 277–313. https://doi.org/10.1007/s13165-015-0109-3

Kozinets, R. V. (2002). The Field Behind the Screen: Using Netnography for Marketing Research in Online Communities. *Journal of Marketing Research*, *39*(1), 61–72. https://doi.org/10.1509/jmkr.39.1.61.18935

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (Fourth edition). Los Angeles, London, New Delhi, Singapore: SAGE.

Meza, X. V., & Park, H. W. (2016). Organic products in Mexico and South Korea on Twitter. *Journal of Business Ethics*, *135*(3), 587–603. https://doi.org/10.1007/s10551-014-2345-y

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. Retrieved from http://arxiv.org/pdf/1301.3781v3

Olson, E. L. (2017). The rationalization and persistence of organic food beliefs in the face of contrary evidence. *Journal of Cleaner Production*, *140*, 1007–1013. https://doi.org/10.1016/j.jclepro.2016.06.005

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention Is All You Need. *ArXiv*. Retrieved from http://arxiv.org/pdf/1706.03762v5

Vidal, L., Ares, G., & Jaeger, S. R. (2018). Chapter 6 - Application of Social Media for Consumer Research. In G. Ares & P. Varela (Eds.), *Woodhead Publishing Series in Food Science, Technology and Nutrition. Methods in consumer research* (pp. 125–155). Duxford, United Kingdom: Woodhead Publishing.

Ziegele, M., Breiner, T., & Quiring, O. (2014). What Creates Interactivity in Online News Discussions? An Exploratory Analysis of Discussion Factors in User Comments on News Items. *Journal of Communication*, *64*(6), 1111–1138. https://doi.org/10.1111/jcom.12123

A.4 Classification of Consumer Belief Statements From Social Media

The paper on the consecutive pages is a technical report which has been published on `arxiv.org` without peer-review. Gerhard Johann Hagerer, the author of the present thesis, is the first author of that paper. His author role, the reprinting permission, and his following contributions are acknowledged by all authors Hagerer, Le, Danner, and Groh [2021c]:

> *"Gerhard Johann Hagerer headed the research project. He developed the research idea, the concept, and the methodology of the paper. Furthermore, he directed the implementation process and reviewed the source code deeply. Regarding the writing of the paper, he created the outline, directed the drafting, contributed significant textual parts, incorporated reviewer feedback, and paraphrased, corrected, combined, and otherwise improved drafted material."*

Publication Summary

As shown before, we found pre-trained models producing meaningful semantic features with high correlation to our domain expert annotations. The following study evaluates the feasibility of domain expert annotations for predictive machine learning (ML) on user-generated texts based on these features. Essentially, this aims at uncovering the quality degradation when using our natural language processing (NLP) models to predict the domain expert annotations. This helps to estimate how suitable the predictions would be for the task of Directed Content Analysis. Since there is a lack of regarding methodical research for Directed Content Analysis, it appears necessary to estimate how difficult it is for a ML based NLP model to mimic a domain expert annotator, especially when it comes to predicting fine-grained annotations on small datasets. This is compared with annotations generated by (semi-) automatic clustering routines. It is found that fine-grained expert annotations are problematic for our otherwise helpful NLP prediction models. On the other hand, our clustering approaches generate fine-grained classes which are much more suitable to be classified by the same prediction models. Consequentially, the automated exploration of fine-grained topics and aspects in terms of an Inductive Content Analysis appears to be much more promising than its counterpart of Directed Content Analysis.

Classification of Consumer Belief Statements From Social Media

Gerhard Hagerer[1], **Wenbin Le**[1], **Hannah Danner**[2], **Georg Groh**[1]
[1]Social Computing Research Group, Technical University of Munich
[2]Chair of Marketing and Consumer Research, Technical University of Munich
{ghagerer,grohg}@mytum.de

Abstract

Social media offer plenty of information to perform market research in order to meet the requirements of customers. One way how this research is conducted is that a domain expert gathers and categorizes user-generated content into a complex and fine-grained class structure. In many of such cases, little data meets complex annotations. It is not yet fully understood how this can be leveraged successfully for classification. We examine the classification accuracy of expert labels when used with a) many fine-grained classes and b) few abstract classes. For scenario b) we compare abstract class labels given by the domain expert as baseline and by automatic hierarchical clustering. We compare this to another baseline where the entire class structure is given by a completely unsupervised clustering approach. By doing so, this work can serve as an example of how complex expert annotations are potentially beneficial and can be utilized in the most optimal way for opinion mining in highly specific domains. By exploring across a range of techniques and experiments, we find that automated class abstraction approaches in particular the unsupervised approach performs remarkably well against domain expert baseline on text classification tasks. This has the potential to inspire opinion mining applications in order to support market researchers in practice and to inspire fine-grained automated content analysis on a large scale.

1 Introduction

The rise of social media has enabled and sparked user's desires to share opinions publicly online. The user-generated content often reveals their true *customer beliefs* towards a certain aspect of things and therefore worth being researched. Typical research fields are qualitative social studies, e.g., *content analyses* for market and consumer research as well as political surveys. One of the challenges, however, has been the course of parsing and organizing a large amount of data that is becoming available in the form of natural language into a fine-grained class structure, which provides a more digestible and actionable insight. This can be achieved by injecting the knowledge from domain experts through annotations. In that regard, there is much manual labour and massive associated expenses invested for such content analysis on social media texts.

In the corresponding qualitative research, fine-grained expert annotations are provided, which have a high resolution on few data points compared to what is available within social media. This is contrary to the requirements of automated textual analyses, e.g., *supervised classification* based on natural language processing and machine learning. Consequently, the question if and how that type of *domain-specific, expert-annotated data* can be effectively leveraged for automated large-scale analyses is a challenging research problem. It carries high potential to gain insights into data which goes beyond a few manually selected texts, scaling up the derived opinion mining models to gather relevant statistics over big amounts of related textual social media corpora, i.e., to inform automated opinion mining based analysis.

For their own qualitative content analysis, domain experts provide fine-grained labels. These, in turn, are organized in a meaningful hierarchy, such that coarse-grained classes of texts always contain several fine-grained classes. For example, a class of social media comments about food products might contain statements including fine-grained attributes, such as, *taste*, *safety*, *price*, etc. *Safety* could in turn contain even more fine-grained *belief statements* about *food products containing chemicals*, *safe and regulated foods*, *unsafe nutrition*, and so on. This shows there are labels at different levels inside of the class hierarchy of customer belief statements.

As we are interested in optimizing the conditions towards improved supervised classification performance, we see the potential to compare supervised classification at different levels of such a class hierarchy, as higher, more abstract classes contain more datapoints and thus solidify model training. We investigate if the expert-based class hierarchy is optimal for classification purposes, or if automatic, hierarchical clustering of the classes yields labels, which would be more suitable for supervised classification. This could notably reduce human effort as well as labour cost and increase productivity to thus accelerate the research process. Automated approaches also provide consistent results with less variability than humans. Not to mention, how it could reactivate existing meaningful content analyses to provide new insights at little cost. We perceive a great potential in semi-automated class abstraction, especially for large-scale content analysis. More precisely, we are investigate the following research questions (RQs):

RQ1 How suitable are expert annotations from domain-specific content studies about customer beliefs suitable for automatic classification?

RQ2 How does the combination of fine-grained classes to more coarse classes relate to classification accuracy?

RQ3 Can an automatic combination of fine-grained classes improve classification over manual, expert-based class hierarchies?

RQ4 How does this compare to classes which are derived without any expert knowledge, i.e., by unsupervised text clustering methods?

RQ5 What are the favorable and the unfavorable effects of automatic and manual class combination and how do these relate to each other?

Our research dataset consist of opinions about organic food and related consumer issues on social media in German-speaking countries and the United States, which is described in section 4. We analyze the differences regarding machine learning classification accuracy using labels of varying granularity generated from expert-based and automated class hierarchies as explained in section 3. The latter can be further divided into supervised and unsupervised approaches. In supervised class hierarchies, we form new class hierarchies based on pairwise semantic class similarities and hierarchical clustering of the existing fine-grained expert labels. This differs from unsupervised class abstraction, where the fine-grained classes are not given by the domain expert but by unsupervised semantic text clustering. We describe how to experimentally compare the regarding classification performances in section 4 to see which kind of labeling techniques are beneficial for predictive machine learning models. The results part of section 5 amongst others depicts the effects of class abstraction for classification and according favourable and unfavourable effects. Section 6 answers the research questions and gives an outlook for future work and potential.

2 Related work

Social media are an established source to investigate consumer beliefs of relevant market domains, for instance, [List some examples]. Therefore, content analyses are carried out, which follow the methodology of *grounded theory* (Martin and Turner, 1986). A domain expert labels the given texts with so called codes, which are tags of ideas or concepts describing the related belief statements. These tend to be rather fine-grained and concrete with a high semantic correspondence to the labeled statements from the text. As an outcome, there are many fine-grained labels with few samples per class, which thus need to be combined to categories. This problem is described as follows:

> *"If a label is insufficiently abstract or general, too few observations will fall into that category. One will add few incidents to each concept card as one analyzes the data. In such cases, the label merely restates or rephrases the data. To "work" (Glaser & Strauss, 1967), a conceptual label must occupy a higher level of abstraction than the incidents (facts, observations) it is intended to classify. If the concept label is too abstract, however, too much information will fall into that category."* (Martin and Turner, 1986)

To choose a category system is relevant, because this supports the cognitive and scientific progress of formulating theories and gaining insights over the analyzed behavior and attitudes within that domain. Theoretical works find that this type of content analysis is substantially similar to topic modeling, a technique from text mining and natural language processing (NLP) (Bakharia, 2019; Yu et al., 2011;

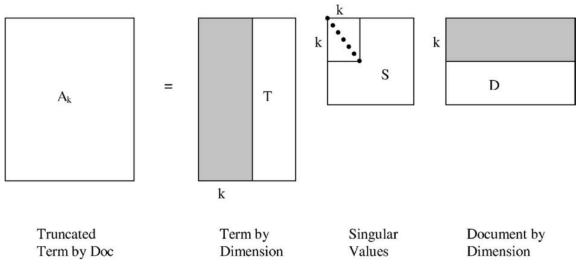


Figure 1: LSI is a SVD decomposing a term-document matrix A into a term-dimension matrix T, a singular-value matrix S, and a document-dimension matrix D, which are all reduced to k latent topics. (Deerwester et al., 1990)

Piepenbrink and Gaur, 2017). As a consequence, these text mining techniques are adopted in practice for content analyses on social media, especially for consumer research (Rocklage and Rucker, 2019). It is a useful means to support decision makers for product development by providing them with information about a market segment, competing companies, and consumer requirements (Xu et al., 2011; He et al., 2013). However, this research does not discuss or leverage NLP methods to detect fine-grained consumer beliefs, which express another dimension of the consumer attitude than the sentiment-related affect (Perner, 2018). In fact, there is little research on if and how consumer beliefs can be classified by NLP classifiers, as these are primarily concerned with the task of sentiment analysis. However, there is evidence that the task of consumer belief classification can be solved by proper NLP methods, including topic modeling as a means to detect new and yet unknown beliefs. The granularity of these, however, is not examined, not to speak of how this relates to expert-based annotations and judgements.

3 Methodology

3.1 Hierarchical Class Clustering

We use agglomerative hierarchical clustering (Maimon and Rokach, 2005) to merge our expert classes into automatically generated class hierarchies based on semantic distance metrics. Since hierarchical clustering is a bottom-up approach, each expert label is first considered as a single-element cluster. We use the weighted-average linkage criteria (WPGMA), which calculates the intra-cluster distance intuitively as the arithmetic mean when forming new clusters.

Language	English	German
Main themes	4	4
Themes	21	21
Beliefs	62	60
Documents	1099	789
Sentences	2275	2334

Table 1: Statistics of the organic dataset

Semantic Similarity Metric For hierarchical clustering of the given expert classes, we use LSI as semantic distance metrics.

LSI (Deerwester et al., 1990) utilizes singular value decomposition (SVD) to reveal latent relationships between documents in a corpus. It yields a low-rank approximation for each document and enables to extract document-document semantic similarity in this low-rank document representation. As illustrated in Fig. 1, SVD decomposes a document-term matrix A into term matrix T, singular value matrix S, and document matrix D, where each matrix will be truncated to k dimensions, i.e., k latent topics. We utilized the document matrix D to construct document representation, in which each document is represented as a linear combination of latent topics. The weights of this combination are taken as the vector representation of the document.

4 Experiments

4.1 Data

The effectiveness of our class combination approaches is evaluated on the organic dataset (Danner and Menapace, 2020). It is gathered to explore organic food beliefs from consumers. The beliefs are annotated by a market researcher on online comments posted on forums and discussion boards of news websites from both English and German-speaking countries. These beliefs are further structured manually into superordinate themes and main themes. The dataset is multi-labeled, since numerous comments have multiple beliefs according to the research. We choose specifically a domain-specific dataset to research how our approach is able to fulfill the requirement even when there is a lack of data. The statistics of the datasets are summarized in Table 1.

Preprocessing Preprocessing consists of 5 steps. Sentence segmentation extracts sentences from documents which will be used to create sentence-level

embeddings for unsupervised approaches. Text cleaning utilizes libraries and regular expressions to filter out URLs, stop words, brackets, quotes, line feeds and blank symbols. Optimal number of clusters are computed using decision criteria such as AIC, BIC scores together with Elbow method.

4.2 Baseline: Expert-Based Class Labels

We take the manually annotated labels and their class hierarchy as given by the domain expert as a baseline and compare them with the class hierarchies obtained from our hierarchical class clustering approach regarding classification performance. The manually annotated label consists of 4 main themes and 21 superordinate themes. Accordingly, we use hierarchical class clustering to generate 4 coarse classes and 21 fine-grained classes.

4.3 Hierarchically Clustered Class Labels

The idea of hierarchically clustered class labels is to analyse the intrinsic semantic class hierarchy of existing expert labels. This is achieved by pairwise combination of fine-grained classes of social media texts according to their overall semantic similarities. The latter is measured by applying LSI as a distance metric on varying document representations. We investigate how the degree of class combination relates to classification accuracy.

There are 2 main steps for supervised class combination approaches, namely semantic similarity extraction and class combination using agglomerative hierarchical clustering. Semantic similarity extraction is realized by calculating distances on document representations. Agglomerative hierarchical clustering combines similar classes into new abstract classes based on distance matrices obtained from semantic similarity extraction. The closest clusters indicated by the Euclidean distance will be successively merged until K (4 or 21) clusters are formed.

4.4 Classes from Unsupervised Clustering

In contrast to the hierarchical clustering approach, which we apply on fine-grained *expert classes* to generate more abstract classes, in the unsupervised approach, we assign labels automatically without incorporating any prior information. This is achieved by K-means clustering on multi-lingually aligned pre-trained textual embeddings. Therefore, we set the number of clusters to 4 or 21 corresponding to the number of main themes and superordinate themes from the expert class hierarchy. This allows us to make a comprehensive comparison of classification performance between labels obtained from the expert-labeled raw dataset, supervised class combination approach, and unsupervised class combination approach.

Each document is first segmented into sentences or words respectively. After clustering all according embeddings from the corpus, the cluster of each document is the most occurring cluster of all its respective sentences or words. A tie results in a multi-labeled document instance.

Since our organic datasets are bilingual, it is preferred to keep the consistency of word and sentence representations irrespective of the language being used. Thus, bilingually aligned representations of textual embeddings between English and German are preferred. Word embeddings are produced by multi-lingually aligned fasttext word vectors computed on Wikipedia (Bojanowski et al., 2017). Sentence embeddings are provided by the multilingual universal sentence encoder XLING (Cer et al., 2018; Chidambaram et al., 2018) trained with a dual-encoder that learns tied representations using translation-based bridge tasks (Chidambaram et al., 2018).

4.5 Classification

We evaluate the performance of our class combination approach on the organic dataset using the gradient boosted decision trees classifier based on tf-idf document features. The dataset can be categorized into two document classification tasks, namely 4-labelled classification, and 21-labelled classification. We use 80% of the data for training and 20% for testing. Besides, we apply 8-fold cross-validation, which divides the data into 8 folds and ensures that each fold is used as testing set in evaluation. The mean value of the scores from each iteration of 8-fold cross validation determines the overall performance of the model.

Since all classification tasks in our experiment are multi-label, exact match, F1 macro, F1 micro, and normalized entropy are the metrics that are used to measure the model performance. Exact match indicates the percentage of samples that have all their labels classified correctly. F1 score is the harmonic mean of precision and recall. It reaches its best value at 1 and worst value at 0. F1 micro and F1 macro differ from each other on the type of averaging performed on the data. F1 micro calculates metrics globally by counting the total true

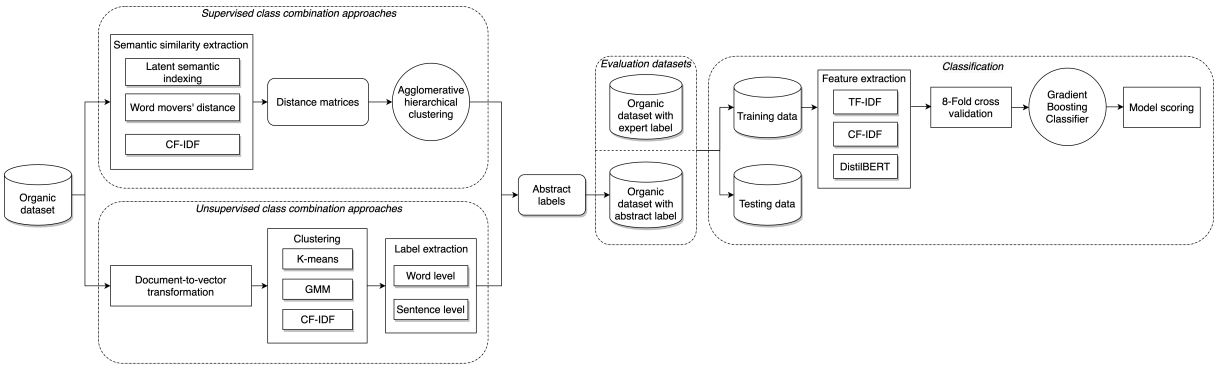


Figure 2: An overview of the experimental design. Abstract labels were extracted using supervised and unsupervised class combination approaches. They will be evaluated in a classification task competing with manually annotated expert label as baseline.

positives, false negatives, and false positives. On the other hand, F1 macro calculates metrics for each class separately and finds their unweighted mean (Pedregosa et al., 2011). Information Entropy, amongst others, is a measure of disorder and uncertainty (Lin, 1991), which we use to depict how balanced the classes are with respect to the number of their containing class samples. Normalization makes the distribution of entropy between 0 and 1, where 1 indicates perfectly balanced classes and 0 maximally unbalanced classes. Using these measures, we aim to research what are the favorable and unfavorable effects of class abstraction and how do these relate to each other.

5 Results

We summarize the main results in Table 2, which shows that the classification results based on hierarchically clustered and unsupervised class labels overall outperform the ones given by the domain expert baseline. The performance gain is especially obvious for hierarchically clustered class labels in exact matches and f1 micro, while for unsupervised classes in F1 macro and normalized entropy. In the remainder of this section, we discuss the results of classification regarding the previous research questions in detail.

5.1 Coarse vs. Fine-Grained Classes

We compare how the number of classes, i.e., few classes with many samples versus many classes with few samples, would relate to classification accuracy against the background of different class hierarchies and methodologies. We consider 4 and 21 as possible number of classes and respective clusters for the unsupervised case.

Regarding macro F1 scores, these appear the best for unsupervised class labels followed by expert class labels. There is the same relation regarding entropy, i.e., there is a clear relation with regards to how well the classes are actually balanced. Micro F1 scores and exact matches tend to better better for hierarchically clustered classes at times where classes are rather imbalanced. From these observations it appears that the issue of balanced classes might be related to how a classifier is going to perform.

5.2 Side Effects

We observe in table 2 that although our hierarchical class clustering approach provides high exact match and F1 micro scores, it also suffers from extremely low F1 macro and normalized entropy scores. From both it can be concluded that this is due to an imbalanced class distribution. It results from those fine-grained expert classes which have a) a low number of documents and b) are semantically highly dissimilar to the other existing classes. This leads to combined classes and regarding class hierarchies containing a comparatively small number of samples. These cannot represent a given class sufficiently for a machine learning classifier. It can be considered as an undesired side effect of hierarchical class clustering, which in its algorithm does not consider class balancing as opposed to k-means for example. As the approach is completely based on existing expert classes which sometimes are unbalanced and dissimilar, this outcome is inevitable. Thus, class imbalance is an unfavourable effect of guided class abstraction.

The expert-based class structure with its 4 and 21 classes, respectively, is more balanced. This means that the way in which the domain expert structures

Corpus	Labels	Expert CA				Supervised CA				Unsupervised CA			
		Em	Fma	Fmi	Ne	Em	Fma	Fmi	Ne	Em	Fma	Fmi	Ne
EN	4	0.39	0.48	0.54	0.92	**0.66**	0.44	**0.76**	0.63	0.45	**0.54**	0.60	**0.98**
	21	0.17	0.22	**0.33**	0.85	0.14	0.18	0.27	0.86	**0.18**	**0.28**	0.31	**0.97**
DE	4	0.25	0.37	0.43	0.92	**0.74**	0.34	**0.80**	0.43	0.32	**0.61**	0.70	**0.94**
	21	0.16	0.13	0.27	0.86	**0.17**	0.13	0.31	0.73	0.08	**0.31**	**0.35**	**0.98**

Table 2: Exact match (Em), F1 macro (Fma), F1 micro (Fmi) , normalized entropy (Ne) for manual domain-expert class abstraction (baseline), best variants of supervised automatic class abstraction (CA) and best variants of unsupervised automatic class abstraction on 4-labelled and 21-labelled English and German text classification tasks based on TF-IDF document representation.

the class hierarchies appears to be more beneficial with respect to class balancing and thus classification with respect to macro F1 scores. However, micro F1 scores are generally small here, which raises questions about how well classification overall might work.

On the other hand, the unsupervised clustering approach does not have these issues and achieves a satisfying result. Its F1 micro, F1 macro, and normalized entropy exceed domain-expert class abstraction by significant margin, while exact match is almost identical. Besides, it can discover new interesting fine-grained classes such as milk that maybe ignored by domain experts. This result shows the beneficial properties of the approach for effective classification, and it demonstrates that it could improve classification over manual, domain expert-based class labels.

6 Discussion

We presented a systematic analysis of 3 class abstraction approaches on the organic datasets for text classification tasks. In accordance with the previously explained results, we answer the research questions from Section Introduction as follows:

RQ1: How suitable are expert annotations from domain-specific content studies about customer beliefs suitable for automatic classification? Table 2 summarizes the classification accuracy of all class abstraction approaches. It can be seen that the domain expert classes offer moderate performances. The macro F1 scores lie between the ones from the hierarchical and unsupervised class labels, which is supported by the fact that the classes tend to be balanced well according to the entropy. However, the micro F1 scores and exact matches are mostly the lowest among all approaches.

RQ2: How does the combination of fine-grained classes to more coarse classes relate to classification accuracy? According to our measurements, a fewer number of classes generally improves classification performance significantly, independent of how the class labels are derived, i.e., by an expert, hierarchical class clustering, or unsupervised clustering. By reducing the number of classes from 21 to 4, macro and micro F1 scores at least doubled for all methods. The improvements are specifically striking for hierarchical class clustering. This means that for classification on small, expert-labeled opinion mining datasets with fine-grained labels, a reduction of the number of classes should be considered as an option.

RQ3: Can an automatic combination of fine-grained classes improve classification over manual, expert-based class hierarchies? The classification results in table 2 show that our hierarchical class clustering approach improves micro F1 classification performance over manual, domain expert-based class labels. However, class imbalancing is a problematic side effect, which can lead to low macro F1 scores. Further research needs to determine if and how this issue can be solved on the given problem domain.

RQ4: How does this compare to classes which are derived without any expert knowledge, i.e., by unsupervised text clustering methods? We observe in Table 2 that our unsupervised clustering approach outperforms expert-based classes with margin for all performance metrics. In the related work, it has been shown that this approach is also able to generate a meaningful intrinsic class hierarchy based on the available data without any human interference (Danner et al., N.D.; Hagerer et al., N.D.). Our finding shows that automated class abstraction can also serve serve as an alternative to domain-expert class labeling for classification. We

perceive a great potential for it and hope the proposed technique would offer valuable guidance for market researchers investigating social media.

RQ5: What are the favorable and the unfavorable effects of automatic and manual class combination and how do these relate to each other? As explained in Subsection Side Effects, hierarchical class clustering suffers from minority classes, which is introduced by small outlier classes of the expert annotations. Those classes which have a low amount of documents are combined and are then not semantically representative enough in classification. Consequentially, they lead to imbalanced abstract class distributions. Thus, achieving not only a satisfying precision but also a high entropy is a demanding task for hierarchical class clustering. Nevertheless, our unsupervised class abstraction approach has overcome this issue and clearly improves over the expert based classes baseline.

7 Conclusion

The described experiments and results show that it is technically feasible to classify belief statements automatically. However, the fine-grained nature of labels given by a domain expert make the task challenging. Combining classes of belief statements is helpful, as it makes the data statistically more representative for machine learning algorithms. Automatized class combinations might lead to imbalanced classes, but overall better classification scores. Classes given by our proposed unsupervised approach yield very balanced classes, which give best classification results. We conclude by recommending our unsupervised clustering approach based on deep, multi-lingual, and pre-trained sentence embeddings, which showed beneficial results in previous opinion mining studies, too (Danner et al., N.D.; Hagerer et al., N.D.).

References

Aneesha Bakharia. 2019. On the equivalence of inductive content analysis and topic modeling. In *Advances in Quantitative Ethnography*, pages 291–298, Cham. Springer International Publishing.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning cross-lingual sentence representations via a multi-task dual-encoder model. *arXiv preprint arXiv:1810.12836*.

Hannah Danner, Gerhard Hagerer, Yan Pan, and Georg Groh. N.D. The news media and its audience: Agenda-setting on organic food in the united states and germany. Currently under review.

Hannah Danner and Luisa Menapace. 2020. Using online comments to explore consumer beliefs regarding organic food in german-speaking countries and the united states. *Food Quality and Preference*, 83:103912.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Gerhard Hagerer, Wing Sheung Leung, Hannah Danner, and Georg Groh. N.D. A case study and qualitative analysis of simple cross-lingual opinion mining. Currently under review.

Wu He, Shenghua Zha, and Ling Li. 2013. Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33(3):464–472.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Oded Maimon and Lior Rokach. 2005. *Data Mining and Knowledge Discovery Handbook*. Springer.

Patricia Yancey Martin and Barry A Turner. 1986. Grounded theory and organizational research. *The journal of applied behavioral science*, 22(2):141–157.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Lars Perner. 2018. Consumer behavior - attitudes. [Online; accessed 29-June-2021].

Anke Piepenbrink and Ajai Singh Gaur. 2017. Topic models as a novel approach to identify themes in content analysis. In *Academy of Management Proceedings*, volume 2017, page 11335. Academy of Management Briarcliff Manor, NY 10510.

Matthew D Rocklage and Derek D Rucker. 2019. Text analysis in consumer research: An overview and tutorial. *Handbook of Research Methods in Consumer Psychology*, pages 385–402.

Kaiquan Xu, Stephen Shaoyi Liao, Jiexun Li, and Yuxia Song. 2011. Mining comparative opinions from customer reviews for competitive intelligence. *Decision Support Systems*, 50(4):743–754. Enterprise Risk and Security Management: Data, Text and Web Mining.

Chong Ho Yu, Angel Jannasch-Pennell, and Samuel DiGangi. 2011. Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. *Qualitative Report*, 16(3):730–744.

B Dataset: Transcripts From Video Reviews

B.1 GraphTMT: Unsupervised Graph-Based Topic Modeling From Video Transcripts

The publication on the consecutive pages was accepted after peer-review as full paper at the 2021 IEEE Seventh International Conference on Multimedia Big Data. Gerhard Johann Hagerer, the author of the present thesis, is the second author of that paper. His author role, the reprinting permission, and his following contributions are acknowledged by all authors Stappen, Thies, Hagerer, Schuller, and Groh [2021]:

> *"Gerhard Johann Hagerer supervised the research project and validated the research methodology of the paper. Regarding the writing of the paper, he contributed parts to the related work, incorporated reviewer feedback, and corrected drafted material."*

Publication Summary

The following paper performs topic modeling on transcripts of YouTube videos containing product reviews about cars. The relevance for this dissertation is to discover new social media data sources for opinion mining and to establish modern methods to extract relevant information from it. Video transcripts are hardly examined in that regard, but these are a rich and feasible source of information as we show in the study. Our proposed topic modeling approach can be considered as novel, since on that kind of data well functioning topic modeling methods are generally not well understood, and especially graph-clustering and word embeddings are not wide-spread. The approach performs well compared to other methods, and it demonstrates a viable and methodically simple solution for this rather difficult and new type of social media data.

GraphTMT: Unsupervised Graph-based Topic Modeling from Video Transcripts

1st Jason Thies
Social Computing
Technical University of Munich
Munich, Germany
jason.thies@tum.de

1st Lukas Stappen
EIHW
University of Augsburg
Augsburg, Germany
stappen@ieee.org

2nd Gerhard Hagerer
Social Computing
Technical University of Munich
Munich, Germany
ghagerer@mytum.de

3rd Björn W. Schuller
GLAM
Imperial College London
London, United Kingdom
bjoern.schuller@imperial.ac.uk

4th Georg Groh
Social Computing
Technical University of Munich
Munich, Germany
grohg@mytum.de

***Abstract*—To unfold the tremendous amount of multimedia data uploaded daily to social media platforms, effective topic modeling techniques are needed. Existing work tends to apply topic models on written text datasets. In this paper, we propose a topic extractor on video transcripts. Exploiting neural word embeddings through graph-based clustering, we aim to improve usability and semantic coherence. Unlike most topic models, this approach works without knowing the true number of topics, which is important when no such assumption can or should be made. Experimental results on the real-life multimodal dataset MuSe-CaR demonstrates that our approach GraphTMT extracts coherent and meaningful topics and outperforms baseline methods. Furthermore, we successfully demonstrate the applicability of our approach on the popular Citysearch corpus.**

***Keywords*-topic modeling, graph connectivity, transcripts, k-components, clustering**

I. Introduction

Hundreds of hours of videos are uploaded to YouTube every minute, enabling studies in various fields of research. For example, educational information on cancer treatment [1] and hearing aids [2] are studied in health-care, the influence on election campaigns in social sciences [3], and large-scale multimodal sentiment in multimodal machine learning [4], [5], [6], [7], [8]. For these approaches, researchers closely examine the videos for collection, labelling, and analysis, whereby visual patterns and metadata, e. g. , authorship, can be exploited. Nowadays, also transcripts – automatically created by YouTube – are available [9]. Since text is the most meaningful modality to understand contextual information, effective computer-assisted text analysis methods are needed.

Topic models that structure information into theme distributions have existed for many years. It has been performed on a range of different texts, including online social network data [10], [11], [12], journals [13], and transcripts [14], [15].

[1]JT and LS contributed equally to this work.

Given a transcript snippet: "*It comes with four turbochargers on [and] has an aught [⇒ naught] to 62 [⇒ 60] time of just 5.2 seconds and [...]*", a typical two-way topic modeling procedure *first*, extracts the aspect terms e. g. , "turbochargers", *second*, clusters the aspects into coherent topic clusters e. g. , "motorisation" = {"turbochargers", "engine", ...}. Automatic transcripts, however, bring unique challenges. Transcripts often have errors like missing words ("*and*"), incorrect ("*62*" ⇒ "*60*"), and similar sounding words ("*aught*" ⇒ "*naught*") due to erroneous speech-2-text processing.

Video transcripts are an emerging data domain, however, the explicit use for topic modeling is understudied [15], [14], [16]. To broaden the perspective on this medium more evaluation and new approaches are needed. Recently, graph connectivity showed promising results on extracting topic from news articles [17]. Compared to other methods [18], [10], [19], the number of expected topics does not have to be explicitly determined a priori. In addition, graph modelling research has gained momentum in several areas, such as text classification [20] and video retrieval [21].

In this work, we propose a *Graph-based Topic Modeling approach for Transcripts* (GraphTMT). For benchmarking, we base our evaluation on *a)* a problem-specific multimedia dataset of car reviews, MuSe-CaR [6], and *b)* the popular written-text dataset Citysearch [22]. MuSe-CaR is one of the largest state-of-the-art video datasets for multimodal sentiment analysis research, containing almost 40 hours of video footage and transcripts of car reviews. The reported word error rate of the automatic transcript is estimated around 28 % [6]. To the best of our knowledge, studies on topic extraction have only been conducted in a supervised fashion [23], [24], [25] on this corpus. Furthermore, Citysearch is utilised to evaluate the applicability of our approach to other datasets. It covers written reviews from restaurant visits and is often featured for the task of aspect and topic modeling in previous works [22], [26], [27].

Our contributions are as follows: We propose a novel graph-based approach for topic modeling for the emerging use case of video transcripts. It is the first time, an unsupervised extraction model is applied to a large-scale, noisy MuSe-CaR dataset packed with typical mistakes of automatic speech-to-text. The performance is extensively benchmarked on this dataset against conventional methods. Here, the semantic consistency of the topics is evaluated by assessing a common coherence measure. Furthermore, for a more human-centred evaluation approach of the results and to determine the semantic validity, we conduct a structured word intrusion user study with 31 subjects. Finally, we evaluate the coherence of our approach on a standard topic modeling dataset of product reviews to assess the potential for other use cases. Our results show that GraphTMT outperforms conventional methods on the MuSe-CaR datasets. For reproducibility, this paper is adjoined with a public Git repository[1].

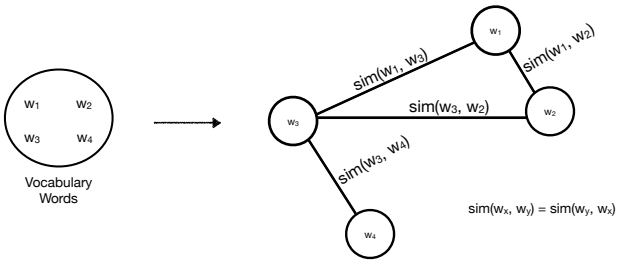


Fig. 1: Illustration of a word embedding graph. Each node represents a word from the vocabulary and each edge is weighted by the similarity between the adjacent nodes. The edges are undirected so that sim(w_i, w_j) = sim(w_j, w_i).

II. Related Work

A. Word Vector Based Topic Models

Topic modeling is often performed by clustering natural language embeddings, grouping semantically similar words together to discover the semantic structure of the underlying corpus [28], [29], [30].

Curiskis [28] compared a traditional topic modeling based on Latent Dirichlet Allocation (LDA) with clustering embedding approaches. All models were applied to Twitter and Reddit textual data. His study indicated that weighted and unweighted embedding clustering has the potential to outperform traditional approaches when using word2vec.

Recently, Sahlgren [29] compared document-based topic modeling to word-based topic modeling. The word-based topic models used utilized embeddings for each prominent word, and the document-based model used document embeddings. The study showed that word-based topic modeling resulted in less or no overlap, more unique topics, and higher average topic coherence. Furthermore, Wang et al. [30] recently evaluated the performance of different topic modeling approaches on Twitter data, applying embedding clustering. The study indicates that more advanced models, such as BERT, do not necessarily outperform approaches on distributed embeddings.

B. Graph-based Topic Models

While these studies used clustering methods to create semantically related word groups, comparatively few have worked with graphs for topic extraction. This paper aims to motivate research in using graph connectivity for topic modeling. While common clustering techniques require strict hyperparameters, e. g. , K-Means requires the true number of topics, K-Components [31] does not. Altuncu et al. [17] used graph connectivity and document embeddings to extract topics. The graph nodes represent documents, and the edges are weighted by the cosine similarity of the respective document pair. The study applied minimum spanning tree and community detection to extract document groups, representing the topics of the corpus. The study concluded that graph connectivity outperforms standard clustering techniques (e. g. , K-Means). Graph-based clustering approaches have been successfully utilized in various applications, e.g. in crime pattern analysis [16] and cohesive subgraphs' discovery for social networks [32].

C. Topic Modeling on Video Transcripts

There are promising applications and use cases of topic modeling related approaches on YouTube video transcripts. Morchid and Linarès [15] used LDA-based topic modeling on self-generated YouTube video transcripts to improve automatic tagging of the uploaded videos. While the overall tagging robustness improved compared to conventional approaches, absolute performance in predicting user-provided tags remained low. The authors argued that this is due to subjectivity and high word error rate of their custom speech recognition system. More recent works are based on the video transcripts provided by YouTube itself. Basu et al. [14] apply preprocessing using automatic spell checking and irrelevant word removal. They utilize LDA for soft assignment of topics to teaching videos and texts.Furthermore, latent semantic indexing, a technique related to topic modeling, has been leveraged for search indexing on YouTube transcripts [33]. Despite existing topic modeling applications, to the authors´ best knowledge, there are no coherence evaluations of topic modeling technology on YouTube transcripts. Such tool would be helpful to extract opinion targets for opinion mining purposes on video product reviews in an unsupervised manner [27], widely established approach on text-based product reviews. Our goal is to foster this research on publicly available video transcripts for market research purposes.

III. Approach: GraphTMT

In this section, we describe our proposed graph-based topic modeling approach. The ultimate goal of GraphTMT is to create and split a word embedding graph, into subgraphs based

[1]Our code can be found at *https://github.com/JaTrev/unsupervised_graph-based*

on edge connectivity. The resulting subgraphs, similar to word embedding clusters, hold semantically related words and are considered the prominent topics of the corpus.

A. Word Embedding Graph

Given a set of vocabulary words W ($|W| = n$), a unique set of the most prominent corpus words, a word embedding graph $G = (N, E)$ is created consisting of $|N| \leq n$ nodes. Each node represents a vocabulary word and each undirected edge $e \in E$ is weighted by the cosine similarity score of the adjacent nodes (cf. Figure 1). Cosine similarity is used to represent the semantic similarity embodied within the trained embeddings [34]. A higher cosine score indicates higher semantic similarity, while an edge weighted with a low cosine score indicates that the adjacent words are not semantically related.

B. Edge Dropping

By weighting the edges, low-weighted edges can be removed from the graph without disconnecting subgraphs of high semantic similarity. To extract insightful topics from the graph, GraphTMT uses a percentile threshold p_t to remove low-weighted edges in E.

C. Graph-based Topic Modeling

Using the resulting (incomplete) graph, the k-component subgraphs [31], [35], [36] are calculated. A k-component is a maximal subgraph of the original graph having (at least) edge connectivity k, a minimum of k edges must be removed from such a k-component subgraph to split it into further subgraphs. These subgraphs are inherently hierarchical; a 1-connected graph can contain several 2-component subgraphs, each of which can contain multiple 3-component subgraphs. In Figure 1, $G_{sub} = (N_{sub}, E_{sub})$, with $N_{sub} = \{w_1, w_2, w_3\}$ and $E_{sub} = \{\text{sim}(w_1, w_3), \text{sim}(w_3, w_2), \text{sim}(w_1, w_2)\}$ is a 2-component subgraph of the given graph. Each k-component subgraph represents a topic discussed in the corpus. The top N representatives of each topic are selected based on node degree and node weights.

IV. Experimental Setup

A. Datasets

We evaluate our method on two real-world datasets. We focus on MuSe-CaR, applying different topic modeling approaches to the unique dataset but include the Citysearch corpus to demonstrate the applicability of GraphTMT outside of video transcripts.

a) MuSe-CaR:: The MuSe-CaR [6] is a multimodal dataset gathered in-the-wild from English YouTube videos centred around car reviews. It was created with different computational tasks in mind, allowing researchers to improve the machine's understanding of how sentiment and topics are connected. The in-the-wild aspect of MuSe-CaR refers to the natural conditions a video is captured in. It varies in recording equipment, recording setting, and soundscapes. The audio captures ambient noises (e. g. , car noises), while the non-acted speech includes colloquialisms and domain-specific terms.

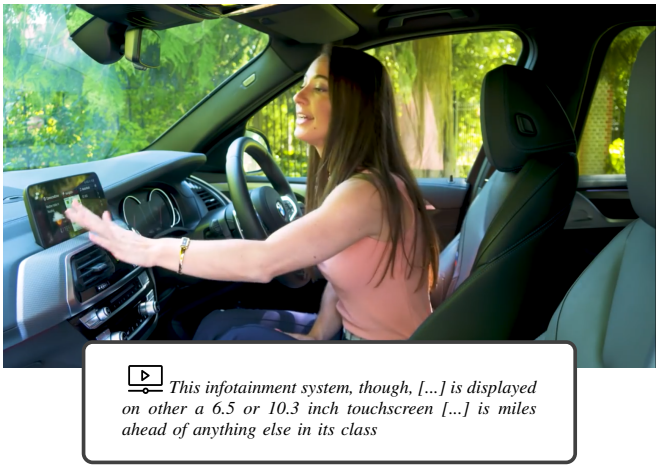


Fig. 2: Frame from MuSe-CaR (video id 2, 4:06) showing a *User Experience* segment and corresponding transcripts.

For our experiments, we use a preprocessed subset of the data featuring labelled topic segments[2], consisting of a total of 35h 39min of YouTube car review videos of approx. 90 speakers [23]. Consisting of real-life opinions about different aspects of modern vehicles, the dataset allows one to apply models to a large volume of user-generated data. The corpus includes 5 467 segments, each consisting of multiple sentences (total: $>$ 20k sentences) with an average of 54 words. Long, encapsuled utterances are typical for transcripts. Video segments are assigned to one of ten topics: *Comfort*, *Costs*, *Exterior Features*, *General Information*, *Handling*, *Interior Features*, *Performance*, *Safety*, *Quality* & *Aesthetic*, and *User Experience*. The transcripts are generated by the authors using automatic Amazon Transcribe speech-to-text pipelines. Due to the in-the-wild factors, the error rate of the automatic transcripts is estimated to be relatively high and specified at around 28 % with outliers of up to 39 % on a subset of 10 hand-transcribed videos [6].

b) Citysearch corpus:: Restaurant reviews from Citysearch[3] have been widely used in previous works [22], [26], [27]. Citysearch was created in 2006. The project aims to provide a better understanding of patterns in user reviews and create tools to better analyse text reviews. The corpus contains over 50 000 restaurants reviews, written by over 30 000 distinct users. Ganu et al. [37] manually labelled a subset of 3 400 sentences using one of six topics: *Ambience*, *Anecdotes*, *Food*, *Miscellaneous*, *Price*, and *Staff*. The topic modeling approaches are evaluated based on this labeled subset.

B. Preprocessing

We begin by extracting the corpus vocabulary $W = \{w_1, w_2, ..., w_n\}$ ($|W| = n$). The Natural Language Toolkit

[2]Download MuSe-Topic: https://zenodo.org/record/4134733

[3]Download Citysearch: http://www.cs.cmu.edu/~mehrbod/RR/, accessed on 29 April 2021

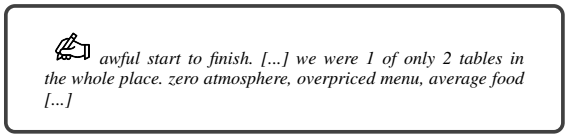


Fig. 3: Snippet from a review from the Citysearch corpus.

(NLTK) [38] part-of-speech (POS) tagger is used to collect POS tags for each word. Word tags have been successfully applied in previous studies [39], [16]. Stop word removal is applied to the Citysearch vocabulary, due to its larger size.

After extracting the corpus vocabulary W, we associate each word to a word embedding. The word2vec model [40] is used to learn these feature vectors, using the following parameters: window size = 15, epoch = 400, hierarchical softmax, and the skip-gram word2vec model [40]. For a fair comparison, this configuration is used in all settings.

Furthermore, we run preliminary experiments on MuSe-CaR and Citysearch utilising the POS tags (cf. Section III). The results indicated that using only nouns performs better on MuSe-CaR, regardless of the method, while the use of all parts-of-speech tags yields slightly better results on Citysearch (cf. Section VII) which we report in the following.

C. Baseline Approaches

Three baseline approaches are compared with GraphTMT: Latent Dirichlet allocation (LDA) [41], K-Means [42], and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)[43].

LDA is a common topic modeling technique, using word co-occurrences to learn semantic clusters. It uses a Dirichlet prior on the topic distribution and the topic representatives distribution. LDA works with a bag-of-words (BOW) representation of the data. Each text is represented as a set of words and their cardinality, neglecting the sentence structure and context. Commonly, the BOW representation is translated into term frequency (TF) or TF-inverse document frequency (TF-IDF) matrix representation. K-Means is a common clustering technique used in topic modeling [18], [10], [19], [30], [17]. While LDA works on probability distributions of topics on the document, K-Means uses the distance between clusters. Similarly to LDA, K-Means commonly [10] uses the TF or TF-IDF matrix representation of the data. The algorithm simultaneously divides the dataset into a number of T_n clusters. The number of clusters is predefined, and the algorithm repeats two steps: an assignment and an update step. While in the assignment step, each data point is assigned to the cluster centroid based on the least squared Euclidean distance, the update step recalculates the centroids. HDBSCAN is a hierarchical and density-based clustering technique which creates a minimum spanning tree and condenses it into smaller trees to create clusters, stopping at C_{min}. Unlike K-Means, HDBSCAN allows for outliers.

Parameter	Values
Number of topics (T_n)	[4; 20]
Document-topic density (α)	[0.1, 0.4, 0.7, 1.0, 1/T_n]
Word-topic density (β)	[0.1, 0.4, 0.7, 1.0]
Weighting strategy	[TF, TF-IDF]
Minimum cluster size(C_{min})	[5; 30]
Edge-connectivity (k)	[1, 2, 3]
Edge weight threshold (p_t)	[0.50, 0.60, 0.70, 0.80, 0.90, 0.95]

TABLE I: Parameter settings of the models

D. Measures

The different topic modeling approaches are measured by: (1) a coherence score, (2) intra-topic assessment, and (3) a user study.

a) (1) Coherence score: Topic coherence measures the degree of semantic similarity between topic representatives, the topic's ten most eminent words. A model's coherence score is the average of all topic scores. This study uses the c_v coherence score [44]. It is based on a sliding window approach that uses normalized pointwise mutual information (NPMI) and cosine similarity. Röder et al. [44] studied the correlation between numerous coherence scores and human judgement and found that c_v correlates best with human ratings.

b) (2) Intra-topic assessment: As coherence scores only capture the similarity between topic representatives, the intra-topic assessment compares the inferred topics with the dataset topic labels (the gold topics) [29]. It includes two measures:

- Topic coverage (T_c): how many gold topics are inferred? This is the proportion of gold topics that are included in the model's topics. A larger number indicates better gold topic representation.
- Topic overlap (T_o): how much do the topics overlap? Each topic is given a label based on its representative, we compare these labels to find the proportion of duplicates. A small overlap indicates unique semantic structures.

c) (3) User study: Furthermore, a user study is conducted on MuSe-CaR models to measure the human interpretability of the inferred topics. Although topic coherence is measured, the interpretability of topics does not always align with coherence scores [45]. Our user study consists of the word intrusion task [46], [47], [45]. Each task is composed of six words, five representatives of a single topic, and a *not sure* option. The task is to find the word that represents a different topic, i. e. , the intruder. Given the following intrusion task: {system, screen, diesel, menus, voice, entertainment, *not sure*}, all words besides "diesel" represent the same topic (infotainment). In this example, "diesel" is the intruder.

A models precision defines how well the intruder detected by the participants corresponds to the true intruder. We define the Word Intrusion Precision (WIP) by the fraction of subjects that find the correct intruders,

$$WIP_k^m = \sum_s \mathbb{1}(i_{k,s}^m = w_k^m)/S. \quad (1)$$

Let w_k^m be the intruder from the k^{th} topic inferred by model m and let $i_{k,s}^m$ be the intruder selected by participant s on the

Topic Models	T_n	c_v	T_C	T_O	WIP	NSF
LDA (α = 0.10, β = 0.70)	8	.51	.60	.25	.43	.13
K-Means (TF-weighted)	8	.73	**.60**	.25	.61	.15
HDBSCAN (C_{min}=6)	11	.63	.60	.4	-	-
GraphTMT (k = 1, p_t= 0.80)	6	.76	.50	**.17**	**.63**	**.08**
GraphTMT (k = 2, p_t= 0.80)	5	**.85**	.40	.20	-	-
GraphTMT (k = 3, p_t= 0.80)	2	-	-	0	-	-

TABLE II: Results on MuSe-CaR for the different topic models and five different evaluation metrics: coherence score (c_v), topic coverage (T_c), topic overlap (T_o), WIP, and overall *not sure* fraction. Note, HDBSCAN was not included in the user study and only one GraphTMTmodel was assessed by the participants.

k^{th} topic. Let S denote the number of participants in the user study. Furthermore, the fraction of subjects that chose the *not sure* option (NSF) is captured.

To reduce study complexity, each model is assessed by half of its inferred topics (chosen at random) and each topic is assessed by a single word intrusion task. Overall the study includes 31 participants, each having an upper-intermediate English level (minimum of B2 in the Common European Framework of Language Reference).

V. MuSe-CaR Evaluation

We first present the results on MuSe-CaR followed by the performance on Citysearch. All model parameters are optimized to maximize the topic model coherence. During the experimental process in this paper, adjustable parameters are set uniformly as shown in Table I. Any model inferring less than four topics and any topic with less than 5 representatives is not considered in our evaluation.

A. Coherence Score Comparison

In the first set of experiments, we compare our four models (LDA, K-Means, HDBSCAN, GraphTMT) on MuSe-CaR based on their coherence score. Table II shows the results of the best performing hyperparameters. Although the corpus has 10 gold topics, LDA and K-Means perform best with eight topics. The clustering-based model gets better scores using TF instead of TF-IDF. K-Means scores better than HDBSCAN but the hierarchical clustering techniques results in more topics. Our graph-based approach results in the highest coherence score (c_v = .85), achieving significant average topic coherence without specifying the number of topics (T_n) or the minimum size of a topic (C_{min}).

Furthermore, Table II shows the impact of k on GraphTMT. Increasing the edge connectivity parameter positively impacts the coherence score but at the expense of fewer topics. By increasing k, lower-weighted edges are removed from the graph, splitting or removing previously existing subgraphs. The new subgraphs only include the highest-weighted edges and most semantically related words. We note that GraphTMT (k = 3) results in only two topics, with $\geq$ 5 representatives, so it is not assessed in our experiments.

From these results, we can make the following observations: (1) the best performing approaches do not include 10 topics; (2) baseline approaches can be used on MuSe-CaR to infer coherent topics; (3) clustering-based topic modeling achieves higher scores than probability-based LDA; (4) GraphTMT infers the most coherent topics without the need to specify the number of topics; and (5) by increasing k, the overall topic coherence of GraphTMT increases but T_n decreases.

B. Word Intrusion

As described in Section IV-D, the word intrusion task measures how well the inferred topics are interpretable by humans. Table II lists the precision results for the three best performing models (LDA, K-Means, GraphTMT) on MuSe-CaR. In our case, the c_v score aligns well with human judgement [44]. The best scoring topic model (GraphTMT) has the highest precision and the worst scoring model (LDA) has the lowest precision. Furthermore, GraphTMT has the lowest NSF score. These findings suggest that GraphTMT results in the most interpretable topics, underlining previous coherence results.

C. Intra-Topic Assessment

The previous two sections show K-Means and GraphTMT having the best topic coherence and WIP. This section looks at these two models' topic coverage and overlap (cf. Table II). K-Means has higher topic coverage than GraphTMT, but GraphTMT has a lower overlap between its topics. The overlap between topics reduces when we increase the edge connectivity constraint (k) but at the expense of topic coverage.

The eight topics inferred by K-Means (TF-weighted) are listed in Table III. Each topic is given a label, based on its topic representatives, and assigned to a gold topic. Overall, six unique gold topics can be matched (T_c = 6/10) but two topics are duplicates (T_o = 2/8).

Table III (middle) lists the six GraphTMT (k=1) topics. The topics include five gold topics (T_c = 5/10) and one overlap (T_o = 1/6). These topics can be compared to GraphTMT (k=2) in Table III. By increasing k, one of the two inferred *Infotainment* topics is removed from the graph, while *Performance* is split into two separate topics. Furthermore, the *Handling* topic was removed. As the coherence score increases with k, topics remaining in Table III (GraphTMT, k = 2) have a higher topic coherence score than the ones removed.

VI. Citysearch Evaluation

In the second part of our evaluation, we compare the performance of all four models on the Citysearch to show GraphTMT's applicability outside of YouTube transcripts. The models are compared on their coherence score, topic coverage, and topic overlap.

A. Coherence Score Comparison

Table IV lists the results of the best performing models based on their coherence scores. K-Means and GraphTMT (k=3) result in the highest coherence score, and LDA has the lowest. Similar to MuSe-CaR, K-Means gets better scores using TF instead of TF-IDF and increasing k has a positive effect

Inferred Topic	Topic Representatives	Gold Topic
K-Means		
Handling	suspension, handling, dampers, corners, chassis	Handling
Infotainment	menus, satnav, swivel, commands, entertainment	User Experience
Interior Features	dash, design, events, wood, plastic	Interior Features
Performance	engine, turbo, litre, cylinder, engines	Performance
Safety	detection, assist, safety, collision, airbags	Safety
Storage	storage, items, space, boot, hooks	General Information
YouTube	please, enjoy, click, share, wow	General Information
Miscellaneous	cars, guys, opportunity, brand, tomorrow	General Information
GraphTMT (K= 1)		
Infotainment	navigation, controls, touch, apple, buttons	User Experience
Infotainment	hand, pop, screen, entertainment, information	User Experience
Passenger Space	area, head, roof, room, headroom	Interior Features
Handling	suspension, corners, steering, gear, response	Handling
Performance	seconds, turbo, twin, acceleration, cylinder	Performance
YouTube	channel, dot, please, thanks, share	General Information
GraphTMT (k = 2)		
Infotainment	hand, pop, screen, entertainment, information	Infotainment
Passenger Space	seat, back, headroom, room, head	Handling
Performance	seconds petrol miles diesel economy gallon fuel	Performance
Performance	seconds, turbo, acceleration, twin, cylinder	Performance
YouTube	dot, channel, please, wow, share	General Information

TABLE III: List of topics extracted on MuSe-CaR where K-Means uses TF-weighted; GraphTMT uses p_t = 0.8.

Topic Models	T_n	c_v	T_c	T_o
LDA(α = 1/T_n, β = 0.40)	8	.48	.67	.50
K-Means(TF-weighted)	8	**.64**	**.83**	.38
HDBSCAN(C_{min}=5)	3	.61	.33	.33
GraphTMT (k = 1, p_t= 0.80)	9	.40	.67	.56
GraphTMT (k = 2, p_t= 0.80)	6	.60	.67	.33
GraphTMT (k = 3, p_t= 0.80)	5	**.64**	.67	**.20**

TABLE IV: Results on the Citysearch for four different topic models (LDA, K-Means, HDBSCAN, GraphTMT) and three metrics: coherence score (c_v), topic coverage (T_c), and topic overlap (T_o).

on the coherence score of GraphTMT but reduces the number of topics. Citysearch has six gold topics, but K-Means infers eight and GraphTMT (k=3) results in five topics. At k = 1 our approach infers nine topics but has a lower score than LDA. HDBSCAN performed similar to K-Means but infers only three topics.

These scores show that our approach is applicable outside of YouTube transcripts, achieving the highest c_v score. Furthermore, they confirm a previous finding, increasing k results in a better score but fewer topics.

B. Intra-Topic Assessment

The previous scores show that K-Means and GraphTMT (k=3) have the best overall topic coherence. In the following, we look at their topic coverage and overlap (cf. Table II). Table V lists all K-Means topics, their inferred labels, and the model's gold topic coverage. The table shows that K-Means covers five of the six gold topics (T_c = .83): *Ambience*, *Anecdotes*, *Food*, *Miscellaneous*,*Price*, but *Anecdotes*, and *Food* are captured twice (T_o = .375).

All GraphTMT models cover four of the six gold topics but as k increases, the topic overlap decreases. Table V lists the nine GraphTMT (k=1) topics, their inferred labels, and the topic coverage. Comparing these topics with the topics at k=3 shows the effect of k on GraphTMT. Increasing the edge connectivity parameter lowers the number of topics but can also let new topics turn up (i. e. , *Ambient* is in GraphTMT

Inferred Topic	Topic Representatives	Gold Topic
K-Means		
Ambience	comfy, spacious, calm, sleek, couch	Ambience
Miscellanous	appear, control, clue, sight, fooled	Miscellanous
Anecdotes	yesterday, today, tonight, march, celebrate	Anecdotes
Anecdotes	refused, proceeded, busboy, ignored, annoyed	Anecdotes
Price	normal, pay, normally, expensive, afford	Price
Location	south, astoria, williamsburg, ues, houston	Miscellaneous
Food	yogurt, pear, pate, walnut, cinnamon	Food
Food	sliced, char, pate, prawn, chorizo	Food
GraphTMT (k = 1)		
Food	pickled, seed, puree, fennel, curried	Food
Food	poivre, hanger, hangar, flank, frites	Food
Service (neg.)	unhelpful, unattentive, unapologetic, arrogant, unfriendly	Staff
Service (pos.)	responsive, cordial, polite, gracious, professional	Staff
Location	washington, seaport, murray, madison, greene	Miscellaneous
Location	chelsea, downtown, soho, meatpacking, tribeca	Miscellaneous
Location	brand, england, yorker, orleans, yorkers	Miscellanous
Anecdotes	incredible, outstanding, terrific, excellent, fantastic	Anedcote
Time	tuesday, wednesday, monday, friday, thursday	Anedcote
GraphTMT (k = 3)		**Gold Topic**
Food	pickled, seed, puree, fennel, curried	Food
Service (neg.)	unhelpful, unattentive, unapologetic, arrogant, unfriendly	Staff
Service (pos.)	responsive, engaging, sincere, caring, hospitable	Staff
Anecdotes	flavorless, tasteless, overcooked, undercooked, inedible	Anecdotes
Ambient	painted, tile, lantern, banquette, chandelier	Ambient

TABLE V: List of topics extracted on Citysearch where K-Means uses TF-weighted; GraphTMT uses p_t = 0.8.

(k=3) but not in GraphTMT (k=1)). This shows that topics can hold more semantics than indicated by their representatives, and increasing k can split an existing topic into semantically different topics, showing the hierarchical structure of our graph-based approach.

VII. Discussion

We evaluated the competitiveness of our novel graph-based topic modeling approach to common alternatives (LDA, K-Means, HDBSCAN) on two different datasets (MuSe-CaR, Citysearch). Our experiments have shown that GraphTMT achieves the highest coherence scores on MuSe-CaR and Citysearch. Furthermore, the model's edge-connectivity parameter (k) positivly affects the coherence score but decreases the number of topics. These findings suggest that by varying k we can remove incoherent topics and words that do not semantically align with a topic. We should note that K-Means had the same coherence score on Citysearch but with more topics. All other models (LDA, HDBSCAN) scored less on both datasets. Although K-Means achieved a comparable score on Citysearch with more topics, the model requires one to predefine the number of topics. Since GraphTMT does not require a specification of the (true) number of topics, it is a good alternative if this information is not available, should not be predetermined, or a search for a suitable parameter k can not be performed. Moreover, the automatic retrieval of k by techniques such as the elbow method is controversial and rarely optimal [48].

In addition to comparing the semantic coherence of topics, we conducted a user study to assess the human interpretability of the MuSe-CaR topics. The study included the models with the highest coherence scores (LDA, K-Means, GraphTMT). As in previous studies, the resulting coherence scores align with the coherence scores [44], GraphTMT topics were more interpretable than topics from K-Means and LDA.

The intra-topic assessment allowed us to compare topics

from K-Means and GraphTMT, the two highest scoring models on both datasets. K-Means covered more gold topics, but GraphTMT resulted in topics with less overlap. Note that varying k revealed the hierarchical structure of GraphTMT, increasing the parameter can split a topic into two semantically different topics.

These findings suggest that GraphTMT provides a valid alternative to common topic model techniques as users can interpret the topics better, more unique topics are extracted, and the approach does not require the true number of topics. Overall, this study has shown the relevance of graph connectivity in topic modeling on two different datasets (YouTube transcripts and online restaurant reviews).

In our experiments, GraphTMT has proven to be very robust on a spoken dataset with a high word error rate. We want to validate these findings on other datasets in future work. Furthermore, we want to evaluate different preprocessing approaches for transcript. Another future aim is to compare different graph connectivity algorithms (e. g. , clique percolation method) to find and develop even more effective approaches for topic extraction.

VIII. Conclusion

In this paper, we demonstrated the capability of graph-based topic modeling on real-world YouTube transcribed data (MuSe-CaR) and textual reviews (Citysearch). On the MuSe-CaR dataset, our proposed novel GraphTMT outperforms all three baseline models in terms of cluster coherence, uniqueness, and interpretability. An accompanying user study assessed the last one. On the Citysearch dataset, our method achieves competitive results to K-Means. However, the clusters produced by GraphTMT have less semantic overlap. We conclude that graph-based clustering is a valid alternative for topic modeling on transcripts and provides meaningful results on real-world text datasets. For the future, we will focus on an integrated approach of several modalities, such as, vision, audio and metadata as any attempt at drawing meaning from YouTube must consider all aspects.

References

[1] Corey Basch, Anthony Menafro, Jen Mongiovi, Grace Clarke Hillyer, and Charles Basch. A content analysis of youtube videos related to prostate cancer. *American Journal of Men's Health (AJMH)*, 2017.

[2] Vinaya Manchaiah, Monica L Bellon-Harn, Marcella Michaels, and Eldré W Beukes. A content analysis of youtube videos related to hearing aids. *Journal of the American Academy of Audiology*, 2020.

[3] Vassia Gueorguieva. Voters, myspace, and youtube: The impact of alternative communication channels on the 2006 election cycle and beyond. *Social Science Computer Review*, 2008.

[4] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 2013.

[5] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*. ACM, 2011.

[6] Lukas Stappen, Alice Baird, Lea Schumann, and Björn Schuller. The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements. *IEEE Transactions on Affective Computing*, 2021.

[7] Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva-Maria Meßner, Erik Cambria, Guoying Zhao, and Björn W. Schuller. The muse 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, page 5–14, New York, NY, USA, 2021. Association for Computing Machinery.

[8] Lukas Stappen, Alice Baird, Michelle Lienhart, Annalena Bätz, and Björn Schuller. An estimation of online video user engagement from features of continuous emotions. *arXiv preprint arXiv:2105.01633*, 2021.

[9] Ken Harrenstien. Automatic captions in youtube, Nov 2009. accessed on 29. April 2021.

[10] Stephan A Curiskis, Barry Drake, Thomas R Osborn, and Paul J Kennedy. An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. *Information Processing & Management*, 57(2):102034, 2020.

[11] Paolo Missier, Alexander Romanovsky, Tudor Miu, Atinder Pal, Michael Daniilakis, Alessandro Garcia, Diego Cedrim, and Leonardo da Silva Sousa. Tracking dengue epidemics using twitter content classification and topic modelling. In *Current Trends in Web Engineering*. Springer International Publishing, 2016.

[12] Dr. Rajesh Prabhakar Kaila and Dr. A. V. Krishna Prasad. Informational flow on twitter–corona virus outbreak–topic modelling approach. *International Journal of Advanced Research in Engineering and Technology (IJARET)*, 2020.

[13] Carina Jacobi, Wouter Atteveldt, and Kasper Welbers. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 2015.

[14] Subhasree Basu, Yi Yu, and Roger Zimmermann. Fuzzy clustering of lecture videos based on topic modeling. In *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2016.

[15] Mohamed Morchid and Georges Linarès. A lda-based method for automatic tagging of youtube videos. In *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*. IEEE, 2013.

[16] Priyanka Das, Asit Kumar Das, Janmenjoy Nayak, Danilo Pelusi, and Weiping Ding. A graph based clustering approach for relation extraction from crime data. *IEEE Access*, 2019.

[17] M. Tarik Altuncu, Sophia N. Yaliraki, and Mauricio Barahona. Graph-based topic extraction from vector embeddings of text documents: Application to a corpus of news articles. In *Complex Networks & Their Applications IX*. Springer International Publishing, 2021.

[18] Suzanna Sia, Ayush Dalmia, and Sabrina Mielke. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2020.

[19] Robert-George Radu, Iulia-Maria Rădulescu, Ciprian-Octavian Truică, Elena-Simona Apostol, and Mariana Mocanu. Clustering documents using the document to vector model for dimensionality reduction. In *2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR)*. IEEE, 2020.

[20] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, 2019.

[21] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[22] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*. ACM, 2013.

[23] Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchen Du, Felix Hafner, Lea Schumann, Adria Mallol-Ragolta, Bjoern W. Schuller, Iulia Lefter, Erik Cambria, and Ioannis Kompatsiaris. Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: Emotional car reviews in-the-wild. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop*. ACM, 2020.

[24] Lukas Stappen, Alice Baird, Erik Cambria, and Björn W Schuller. Sentiment analysis and topic recognition in video transcriptions. *IEEE Intelligent Systems*, 2021.

[25] Lukas Stappen, Lea Schumann, Benjamin Sertolli, Alice Baird, Benjamin Weigell, Erik Cambria, and Björn W. Schuller. ¡span class="smallcaps smallercapital"¿muse-toolbox:¡/span¿ the multimodal sentiment analysis

continuous annotation fusion and discrete class transformation toolbox. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, MuSe '21, page 75–82, New York, NY, USA, 2021. Association for Computing Machinery.

[26] Samuel Brody and Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. ACL, 2010.

[27] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. ACL, 2017.

[28] Stephan Curiskis, Barry Drake, Thomas Osborn, and Paul Kennedy. An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. *Information Processing & Management*, 2019.

[29] Magnus Sahlgren. Rethinking topic modelling: From document-space to term-space. In *Findings of the Association for Computational Linguistics: Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020*. ACL, 2020.

[30] Lili Wang, Chongyang Gao, Jason Wei, Weicheng Ma, Ruibo Liu, and Soroush Vosoughi. An empirical survey of unsupervised text representation methods on Twitter data. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*. ACL, 2020.

[31] David W. Matula. k-components, clusters, and slicings in graphs. *SIAM Journal on Applied Mathematics*, 1972.

[32] Rong-Hua Li, Lu Qin, Jeffrey Xu Yu, and Rui Mao. Influential community search in large networks. *Proceedings of the VLDB Endowment (PVLDB)*, 2015.

[33] Diana Iulia Bleoancă, Stella Heras, Javier Palanca, Vicente Julian, and Marian Cristian Mihăescu. Lsi based mechanism for educational videos retrieval by transcripts processing. In Cesar Analide, Paulo Novais, David Camacho, and Hujun Yin, editors, *Intelligent Data Engineering and Automated Learning – IDEAL 2020*, pages 88–100, Cham, 2020. Springer International Publishing.

[34] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics (TACL)*, 2015.

[35] Jordi Torrents and Fabrizio Ferraro. Structural cohesion: Visualization and heuristics for fast computation. *Journal of Social Structure (JoSS)*, 2015.

[36] Douglas R White and Mark Newman. Fast approximation algorithms for finding node-independent paths in networks. *SSRN Electronic Journal*, 2001.

[37] Gayatree Ganu, Noemie Elhadad, and Amélie Marian. Beyond the stars: Improving rating predictions using review text content. In *12th International Workshop on the Web and Databases*. ACM, 2009.

[38] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., 2009.

[39] Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Third IEEE International Conference on Data Mining (ICDM)*. IEEE, 2003.

[40] Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations (ICLR)*. ACL, 2013.

[41] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 2003.

[42] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1967.

[43] Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2017.

[44] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 2015.

[45] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates Inc., 2009.

[46] Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. ACL, 2014.

[47] Fiona Martin and Mark Johnson. More efficient topic modelling through a noun only approach. In *Proceedings of the Australasian Language Technology Association Workshop (ALTA)*. Australasian Language Technology Association (ATLA), 2015.

[48] David J Ketchen and Christopher L Shook. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal*, 1996.