Technische Universität München
TUM School of Social Sciences and Technology

Structures and Dynamics in Communication Conflicts in Social Media Networks

Wienke Christine Strathern

Vollständiger Abdruck der von der TUM School of Social Sciences and Technology der Technischen Universität München zur Erlangung einer

Doktorin der Philosophie (Dr. phil.)

genehmigten Dissertation.

Vorsitz:		Prof. Dr. Janina Steinert
Prüfer*innen der Dissertation:	1.	Prof. Dr. Jürgen Pfeffer
	2.	Priv.-Doz. Dr. Angela Oster

Die Dissertation wurde am 23.05.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Social Sciences and Technology am 31.07.2023 angenommen.

Abstract

Monitoring online hate speech based on user-generated content has become increasingly important for various entities as a response to the alarming rise of online abuse and hateful language that target vulnerable groups on social media platforms. This kind of abusive language encompasses a variety of hostile messages that are meant to intimidate or provoke violence and animosity against particular communities and can also be present in other types of online text. Despite numerous proposed methods in recent years to detect and supervise hateful content, the issue continues due to the intricacy of abusive language and its implicit forms. Our research focuses on exploring different methodological approaches to automatically recognize abusive language in various circumstances to investigate how hate speech or negative communication is conveyed and handled. The central question of this work is how opinions are spread through communication in online social networks. Insights into the structure and dynamics of interaction networks are presented based on empirical data collected from social media networks. Methods for capturing communication behavior in conflict situations are discussed, with a focus on linguistic markers and patterns. The theoretical part of the work deals with the concepts for addressing questions related to networks and communication. Different methods of quantitative text and content analysis, experiments, and network analysis techniques for measuring opinion processes are compared. The primary emphasis of the thesis lies in comparing various approaches to identify distinct linguistic situations. Additionally, the study highlights the differentiation of text data in terms of content and language, along with its preprocessing for quantitative analysis. We discuss how properties from text analysis are suitable for developing models that automatically detect change processes. In addition, linguistic metrics in the context of polarizing situations are experimentally tested. A classification scheme for capturing implicit and explicit hate will be presented, which is matched with automated methods. As a result of this thesis, the importance of similarity in behavior and its possible differentiation through language is discussed.

Zusammenfassung

Überwachung von Online-Hassrede ist aufgrund des besorgniserregenden Anstiegs von Online-Missbrauch und Hasssprache, die auf sozialen Medienplattformen auf schutzbedürftige Gruppen abzielen, für verschiedene Organisationen zu einer immer wichtigeren Angelegenheit geworden. Diese Art beleidigender Sprache umfasst eine Vielzahl feindseliger Botschaften, die dazu dienen, bestimmte Gemeinschaften einzuschüchtern oder Gewalt und Hass gegen sie zu provozieren und kann auch in anderen Formen von Online-Texten gefunden werden. Trotz der in den letzten Jahren vorgeschlagenen verschiedenen Ansätze zur Identifizierung und Überwachung von Hassinhalten besteht das Problem aufgrund der Komplexität beleidigender Sprache und ihrer impliziten Formen weiterhin. Unsere Untersuchung konzentrierte sich auf die methodischen Ansätze zur automatischen Identifizierung von beleidigender Sprache in verschiedenen Situationen, um zu untersuchen, wie Hassrede oder negative Kommunikation ausgedrückt und verarbeitet werden. Im Zentrum der vorliegenden Arbeit steht daher die Frage, wie sich Meinungen in online sozialen Netzwerken durch Kommunikation verbreiten. Anhand empirischer Daten erhoben aus sozialen Mediennetzwerken werden Einblicke in die Struktur und Dynamik von Interaktionsnetzwerken präsentiert. Methoden werden diskutiert, mit denen Kommunikation in Konflikten erfasst werden können. Dabei liegt der Fokus auf sprachlichen Markern und Mustern. Der theoretische Teil behandelt die für die Bearbeitung der Fragestellungen erforderlichen Konzepte zu Netzwerken und Sprachverhalten. In der Arbeit werden verschiedene Verfahren der quantitativen Text- und Inhaltsanalyse, Experimente sowie Techniken der Netzwerkanalyse zur Messung von Meinungsprozessen gegenübergestellt. Der Schwerpunkt der Arbeit liegt in der Gegenüberstellung verschiedener Methoden zur Identifikation verschiedener Sprachsituationen. Ein weiterer Schwerpunkt liegt auf der inhaltlichen und sprachlichen Differenzierung von Textdaten und ihrer Aufbereitung für quantitative Analysen. Es wird gezeigt, dass sich Eigenschaften aus der Textanalyse eignen, um Modelle zu entwickeln, die automatisiert Veränderungsprozesse

erkennen. Außerdem werden linguistische Metriken im Kontext polarisierender Sprachsituationen experimentell getestet und diskutiert. Ein Klassifikationsschema zur Erfassung impliziten und expliziten Hasses wird vorgestellt und mit automatisierten Methoden abgeglichen. Als Ergebnis der Arbeit werden die Bedeutung von Ähnlichkeit in Verhalten, sowie ihrer möglichen Differenzierung über Sprache präsentiert.

Publications

Table 1 shows the list of peer-reviewed papers including the journal and conference and the years at which they were accepted and presented. Table 2 shows the list of other publications thematically relevant to the thesis including the publisher and the year of publication. Table 3 shows the list of presentations including venue, the date and location. **Core publications** of the thesis as listed in Table 1 are relevant for examination in accordance with Exhibit 6 of the regulations for the award of doctoral degree. **Other publications** as in Table 2 are not formally relevant for examination in accordance with Exhibit 6 of the regulations for the award of doctoral degree. However, all publications are thematically relevant for the thesis and will be presented and discussed in Chapter 5 (except for #5).

Table 1: Core publications of the thesis

No.	Year	Authors	Paper	Venue
#1	2022	Wienke Strathern, Raji Ghawi, Mirco Schönfeld, Jürgen Pfeffer	Identifying Lexical Change in Negative Word-of-Mouth on Social Media Networks	Social Network Analysis and Mining 12, Article Number 59, Springer (2022).
#2	2022	Wienke Strathern, Angelina Mooseder, Jürgen Pfeffer	The Polarizing Impact of Continuous Presence on Users' Behavior	Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media, 5-9 July, Atlanta, USA (2022).
#3	2022	Wienke Strathern, Jürgen Pfeffer	Identifying Different Layers of Online Misogyny	Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media, 5-9 June, Cyprus, Greece (2023).

Table 2: Other publications

No.	Year	Authors	Paper	Venue
#3	2020	Wienke Strathern, Moritz Issig, Kati Mozygemba, Jürgen Pfeffer	QualiAnon – The Qualiservice Tool for anonymizing Text Data	Technical Report, TUM-I2087, Technical University of Munich, Department of Informatics (2020).
#4	2020	Wienke Strathern, Mirco Schönfeld, Raji Ghawi, Jürgen Pfeffer	Against the Others! Detecting Moral Outrage in Social Media Networks	The 2020 IEEE/ACM International Conference in Social Networks Analysis and Mining, December 8-10, The Hague, Netherlands, virtual (2020).
#5	2020	Wienke Strathern and Jürgen Pfeffer	Negative Dynamics on Social Media and their Ethical Challenges for AI	Research Brief, Technical University of Munich, Munich Center for Technology and Society, Institute for Ethics in Artificial Intelligence (2020).
#6	2021	Wienke Strathern, Raji Ghawi, Jürgen Pfeffer	Advanced Statistical Analysis of large-scale Web-based Data	Data in Economics and Finance for Decision Makers. Per Nymand-Andersen (Ed.), (p. 43-72) Risk Books. London (2021).
#7	2021	Maximilian Wich, Melissa Breitinger, Wienke Strathern, Marlena Naimarevic, Georg Groh and Jürgen Pfeffer	Are your Friends also Haters? Identification of Hater Networks on Social Media	WWW'21: Companion Proceedings of the Web Conference (Data Paper) (p. 481-485), Ljubljana, Slovenia (virtual), (2021).

Presentations

Table 3: List of presentations

Date	Title and Venue
2020	Against the Others! Detecting Moral Outrage in Social Media Networks. The 2020 International Conference on Advances in Social Network Analysis and Mining, 10 December (The Hague, Netherlands), Virtual
2020	Polarization on Reddit? Understanding dynamics of user interactions in social, ", Sunbelt 2020, International Network for Social Network Analysis media networks, July 13 2020 (Paris, France), Virtual
2021	Examining the role of conspiracy videos related to COVID19 in YouTube's video, Networks 2021, A joint Sunbelt and NetSci Conference, July 8 2021 (Bloomington, Indiana, USA), Virtual
2021	Exploring linguistic patterns of negative opinion dynamics in social media networks, Responsible AI Forum 2021, December 8 2021 (Munich, Germany), Virtual
2022	The Polarizing Impact of Continuous Presence on Users' Behavior, MEDIATE Workshop at the 16th International AAAI Conference on Web and Social Media (ICWSM 2022), July 5 2022 (Atlanta, USA)
2022	Measuring Dimensions of Gendered Hate in Social Media Networks, Sunbelt 2022, International Network for Social Network Analysis, July 13 2022 (Cairns, Australia)
2023	Identifying Different Layers of Online Misogyny, NEATCLasS Workshop at the 17th International AAAI Conference on Web and Social Media (ICWSM2023), 5-8 June 2023, Cyprus (Greece)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> "If the world were drained of every individual and we were left only with the messages that passed between them, we would still be in possession of the information needed to construct our discipline. For every truly socio-psychological phenomenon is rooted in *communication*." (Milgram, 1977, p. 317)

1.1 Motivation

Since its inception in 2009 (Lazer et al., 2009), the interdisciplinary research area of Computational Social Science has been in use for over a decade, and it is still progressing and developing (Lazer et al., 2020). Computational Social Scientists aim to answer critical social science research questions through the use of computational methods that blend computer science and statistics. This interdisciplinary approach offers many benefits, but also presents challenges that I address in this dissertation. With a background in literary scholarship and linguistics, my research predominantly focuses on the application and discussion of computational methods rather than developing computational methods. In this thesis, I present my contributions to the growing field of Computational Social Science. The overarching focus of this dissertation is a multiple perspective on the phenomenon of online communication conflicts. We hereby classify approaches for analysis of hate speech and situations of conflict. Analytically, the focus is on the analysis of communication patterns and dynamics within social media networks, which is a social science research area concerned with understanding the interactions among actors within a network. The overarching question for this dissertation is: How can we use computational methods to study negative human behavior? How can we use social media data to map human-to-

human online communication conflicts? The practice of analyzing large datasets with computers has a long history in studying societies and people. However, the vast amount of real-time and individual-level information available today is unparalleled as a resource for tracking trends, making predictions, and as an aid to decision-making (Lazer et al., 2021). This information is now accessible to almost every social science discipline, allowing researchers in fields ranging from psychology to economics and political science to utilize data in their investigations of critical societal questions. According to their observations, algorithms, which are ubiquitous in our society, influence both individual and group behavior. This means that any analysis of human behavior needs to consider the impact of algorithms on behavior. Studies suggest that social science theories must be updated to account for these influences, because without a clear understanding of how algorithms affect the data available for analysis, researchers will be unable to draw accurate conclusions (Wagner et al., 2021). In addition, the private ownership of large datasets by commercial entities presents another challenge for computational social science. The methods for assessing behavior that originated during the early stages of quantitative social science were (1) essential due to the limitations of measurement techniques at that time, and (2) based on a social context that was significantly distinct (Lazer et al., 2021). The aim of utilizing behavioral trace data now for measurement is to derive meaning from the initial data source. Every scientific data source faces this challenge, "but the leap from raw data to meaningful measures is often particularly large when we use data recycled from systems designed for other purposes often more substantial when utilizing data repurposed from systems intended for other uses" (Lazer et al., 2021, p. 190). The critical issue is whether the measurements accurately represent the construct that is intended to be investigated, hence they need to align with definitions of the relevant concepts. The process of measuring observed phenomena relies on identifying pertinent research questions, arising from behavioral norms, theories, or empiry. According to Lazer et al. (2021) measures could be constructed from data, whether generated by research instrumentation or repurposed from non-research data. The task at hand is to create metrics that offer a certain level of consistency over time or across different systems, pertaining to a specific research query (Lazer et al., 2021). The flexibility of human language and expression presents broad difficulties when it comes to drawing conclusions about attitudes and opinions based on language data. Deciphering sentiment on Twitter is a well-known challenge for computers due to their strug-

gle with identifying sarcasm, irony, and exaggeration. The severity of this issue varies depending on the nature of the data and the specific research question at hand (Lazer et al., 2021). Social media platforms are part of our daily lives and the platforms provide millions of users worldwide with information and enable communication and community building. Positive aspects of social media communication are strong network effects, enabling people to get in touch all over the world, and share information on every existing topic. Social media consumption and participation have become an essential part of many societies, politics, culture, lifestyle, music. They have completely permeated societies. Hopes were and still are high, that interconnected social media usage would scale up political participation, enhance community building, and open doors for business-making. It has become a low-threshold medium that guarantees access to information and contacts. Social media is increasingly prevalent in today's world and has the potential to both positively and negatively impact human interactions and relationships. Studies have shown that excessive social media use can lead to feelings of anxiety, depression, and low self-esteem (Karim, 2020). Negative social media dynamics refer to the negative effects that social media platforms can have on individuals, brands, and society. This can be due to the comparison trap, where users compare their lives to the curated, idealized versions of other people's lives they see on social media, leading to feelings of inadequacy. Furthermore, social media has been linked to a decrease in face-to-face communication and empathy, as well as an increase in cyberbullying, hate speech, and the spread of misinformation. Additionally, research has shown that social media algorithms can contribute to the spread of negativity by amplifying and reinforcing harmful content (Gonzales et al., 2010). This can result in a negative feedback loop, where users are exposed to increasingly negative and divisive content, further fueling negative emotions and behavior. They are, on the one hand, a highly beneficial environment for the propagation of new ideas (Strathern and Pfeffer, 2020), building communities, and sharing experiences or products as influencers do. On the other hand, individuals, companies and their brands, politicians, governmental institutions, and celebrities have increasingly been facing the impact of complaint behavior (Pfeffer et al., 2013; Strathern and Pfeffer, 2020). In times of conflict, social media is a medium for expressing grievances. Social media platforms offer space for uncivil behavior. Alongside from the positive effects of social processes, the negative ones are strongly on the rise. Negativity is an integral part of human behavior. The emergence of social media has led to

the amplification of emotions, attitudes, and arguments through the expression of provocative opinions, with some individuals resorting to inflammatory language when discussing morally charged topics. The impact of cyberbullying and virtual abuse has been researched, but there is still limited knowledge regarding the different types, motivations, and purposes of uncivil, aggressive, and abusive comments posted online. Many comments include uncivil language, which often affects users mental health. By examining the emotional discussions on social media, we can gain a better understanding of these dynamics and how the unique communication style of social media contributes to shaping arguments. It is important to consider how the online format affects conversations, as the escalating nature of online commenting can exacerbate negative emotions and responses, ultimately harming the victims of these outrages. Negative interactions online can lead to the intimidation and silencing of social media victims (Deavours et al., 2022). Hate speech, polarization, radicalization, and the marginalization of underrepresented groups are all phenomena that can be traced back to strong network effects. An increasing amount of studies focus on the problems caused by social media in terms of mental health, political participation, stigmatization of groups, children and teen health, and media consumption. Following recent and ongoing public discussions on the negative aspects of social media, many tools have been developed to track, monitor and capture negativity or uncivil behavior in terms of inappropriate speech and images. Uncivil behavior can be expressed in various forms and its perception as such depends on a variety of factors such as cultural background, one's own perception of incivility, a common understanding of language and intentions, and also on legal and societal definitions of incivility or hate. For this reason, statistical quantification of hate speech in text is a valid approach, but it is limited in the extent to which it can successfully capture the broad range of offensive behavior. And precisely because the number of violations and infringements is increasing and more automated methods and techniques are being applied, validation and recalibration of data and methods from a social science and linguistic point of view become relevant (Radford and Joseph, 2020). Language plays a crucial role in conflicts and uncivil behavior, as it is the primary means of communication and the vehicle through which people express their thoughts, opinions, and emotions. In online contexts, language is used to express disagreement, criticism, or aggression, and it can also be used to defend one's own position or to attack others. One way language can contribute to online conflicts is through the use of inflammatory or divisive language, which

can escalate tensions and incite negative reactions from others. The tone and language used in online interactions can also have a significant impact on how the message is received and interpreted. Using a sarcastic or condescending tone can be perceived as aggressive or dismissive, which can lead to further conflict (Corsevski, 1998). Another aspect of speech that can contribute to online conflicts is the use of language that is perceived as derogatory or offensive. This can include hate speech, slurs, and other forms of speech intended to demean or marginalize certain groups of people. The use of this type of speech can create a toxic and unwelcoming environment for certain individuals and can also lead to further conflicts. This can lead to the exclusion of certain individuals, and the marginalization of particular groups. Such dynamics often manifest themselves in a more aggressive tone. While dictionary-based approaches are frequently used to measure emotions in text, these methods struggle to capture the nuances of language used in context. Additionally, while measuring polarization based on opinions and attitudes can provide insights into the level of polarization around a given topic, it has difficulty in registering the polarizing effects that emerge from heated debates. Furthermore, studies show that hateful speech is expressed and perceived differently from everyday verbal communication in ways that require further distinctions and differentiations. These issues are connected methodologically, and share common challenges. However, traditional methods such as experimental research designs and survey data suffer from low external validity, limiting their generalizability. Communication is a critical component of human interaction, encompassing both verbal and nonverbal behaviors, such as speaking, writing, facial expressions, gestures, and body language. In the online sphere, communication involves a dynamic process of actions such as mentioning, hashtag usage, sharing, liking, retweeting, and following, all of which reflect various forms of communicative behavior (Marres, 2017; Crystal, 2001). Words are used to express attitudes, opinions, and sentiments, including the negative and contentious attitudes that characterize negative dynamics. This thesis focuses on impolite or negative opinions expressed towards individuals, companies, or political topics and seeks to understand the language used when taking a stance for or against something or someone. Valence, sentiment, and linguistic style are all factors that must be considered. Identifying linguistic features in social media dynamics poses a significant methodological challenge, as language is adaptive, flexible, and easily molded into new forms. Simply searching for hate speech in a list of hateful words is not sufficient, as people often circumvent these lists with

linguistic creativity (Strathern et al., 2022b; Frenda et al., 2022; Udupa, 2020). Moreover, aggression in language can take various forms, and hate speech lists may not reflect the complexity of human hatred and its verbal attacks. There are several approaches to identifying linguistic style as a marker of behavioral processes. Smith et al. (2020) conducted a pilot study with a sample of Hong Kong residents during the 2019 protests to study how individuals' psychological changes can be detected by analyzing their mobilizing interactions on social media. Focusing on short texts about the political situation, the researchers used methods of detecting mobilizing actions in order to identify collective action intentions. Using software such as the LIWC, Cohen et al. identified several linguistic markers that were more prevalent among individuals who had been arrested or convicted of acts of radical violence. These markers included the use of negative emotions, words associated with violence, and words related to politics and religion (Cohen et al., 2014). In a recent investigation, Park and Conway (2017) examined the impact of digital health communities on individuals who suffer from depression. They sought to determine whether involvement in such communities can alleviate depression symptoms, or whether continued engagement with others who are also depressed could exacerbate symptoms. The aim of their research was to observe how the psychological states of depressed community members change over time, as indicated by linguistic alterations in their communication with other depressed individuals. The study's findings indicate that joining an online depression community may produce positive emotional outcomes in participants, as determined by various LIWC dictionaries. O'Dea et al. (2021) have also employed comparable methods to explore the connections between linguistic characteristics in blog posts and the manifestation of depression, anxiety, and suicidal thoughts in individuals. Stevens et al. (2021) conduct a longitudinal study using natural language processing (NLP) to examine the impact of the COVID-19 crisis on LGBTQ+ youth, a group that had already experienced a high prevalence of adverse mental health outcomes prior to the pandemic (Stevens et al., 2021). Araque et al. (2022) provide a comprehensible overview on drivers and factors of radicalization, including the level type, language signals, data resources, and references to the literature on this theme. Based on the investigations mentioned above and other recent findings, this thesis seeks to provide insights into the structure of human-to-human communication and into social processes within these networks. The sole purpose of social media is communication and this means that (social) interaction happens continuously

on social media platforms. The question is, when does such a system move in one direction or the other? As a starting point and the main motive of this work, the core scientific question can be formulated as follows: *"How can we map structures and dynamics of networks between actors that emerge through human-to-human communication?"*

1.2 Questions

In this thesis, we are interested in the structure and dynamics of human-to-human interaction on social media in situations of conflict. We are particularly interested in approaching this topic from multiple methodological perspectives. We concentrate on online situations of users who share a common behavior: On the structural level this includes the social interactions of users participating in an online firestorm, actively interacting in polarizing communities, joining a firestorm against a female celebrity. On the content level it includes the comments user share online in these cases. The overall questions are:

Structural Level

- **How can we map change in communication conflicts? What are the properties of negative word-of-mouth** What are the properties of human-to-human networks at the micro-level? What is the structure of the network? What are the co-occurring topics and associations? How can we detect change? How can we distinguish different conflict stages? How do the contributions of users in negative word of mouth differ from normal times?

- **What are the structural properties of polarized communities?** How can we capture interactions? How does activity effect interaction? How can we identify polarizing communities?

Content Level

- **What are the distinct properties of online communicative conflicts?** What are the properties of the content shared? What are the distinct features of communicative aggression? What are the distinct features of emotional expressions?

- **How can we identify interpersonal differences and similarities in language use and style?** How can we analyze speech in conflict situations? How do interpersonal language differences manifest themselves? How can we identify their distinct language properties? How can the effects of polarization be interrupted? Does absence have an effect on language? What are the distinct language properties in conflicts?

- **How can different linguistic layers of hate be identified?** How are they connected? What are the semantic properties of hate text? How are they semantically related?

Linguistic style is identifiable through speech use that is specific to certain situations and can be quantitatively measured using various techniques. By applying computational methods to large-scale text data, typical language use can be classified and segmented through the identification of statistically significant occurrences that are characteristic of a particular situation. This approach allows for an exploratory analysis of the data and the development of methods, without requiring an a priori definition of the specific patterns that will be found. Based on our previous considerations, the following hypotheses are examined in this work:

- Variation in language use and style can be an indicator of social processes.

- Methodologically, this makes it possible to analyze social media data for their exemplarity.

- Theoretically, this results in a better understanding of linguistic styles that make up the "how" of speaking in times of conflict or aggression.

1.3 Structure

Figure 1 shows the structure of the present work and roughly reflects the chapter structure. The first, more extensive part of the work represents an engagement with the theoretical concepts used. Chapter 2 serves as an introduction to the conceptual background. Starting with a structural overview of social media networks, the necessary terms and characteristics at the network level required for this work are defined in subsequent sections. The interaction forms typical of social networks are explained, particularly focusing on social processes and concepts

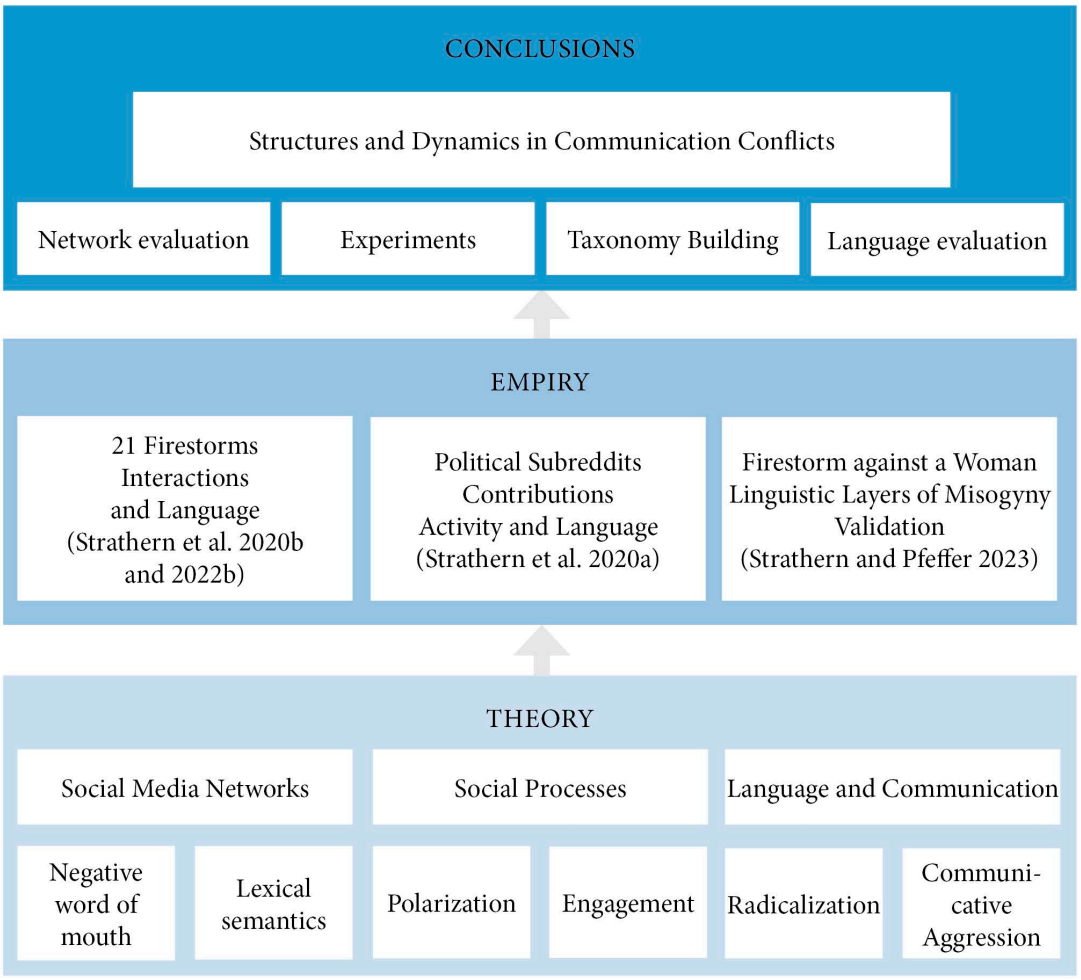


Figure 1: Structure of the present work

in human-to-human interaction. An elucidation is given of the forces that come into play in the formation of negative word-of-mouth, how polarization develops, and how it can be measured. In order to understand the differentiation between normative and non-normative behavior, this section is supplemented with an overview of radicalization tendencies. To enrich data with information, certain properties must be defined. Chapter 2 also includes an introduction to the most important properties of language for this work. The conclusion of this chapter provides a consideration of communicative aggression and hate speech. Language is the medium through which communication happens/is realized. Therefore, both the properties of language and the attributes of aggressive communication are discussed here. In order to be able to conduct language analysis with computational methods, data must be processed in a quantifiable way. An introduction to the three main methods and techniques for these purposes can be found in

Chapter 3. The conclusion of this chapter is a consideration of the strengths and weaknesses of these methods. While the chapter on the conceptual background lays out the theoretical framework, the following chapter is devoted to methodological approaches to analyzing social media text and network data. Chapter 3 deals with the operationalization of the research questions, and begins with a focus on quantitative text analysis. This includes both statistical measurement procedures and techniques for emotional analysis. It presents techniques from lexical research which are well suited to capturing emotions in text data. Following this, quasi-experimental techniques and metrics from linguistics are introduced for complexity analysis. These make it possible to draw inferences about diversity in language data with reference to interaction patterns in social media. Network text analysis is another of the quantitative methods used, and its background and procedures are explained in relation to the research questions of this work. In addition to quantitative text analysis, quantitative content analysis has also proven to be a reliable method for analyzing content from communication situations. In this connection, Lazarsfeld's communication-theoretical approach and Krippendorff's techniques of content analysis have played a particularly important analytic role in the thesis. Content analysis is primarily used for model building and the classification of knowledge. An explanation of taxonomy building rounds off this section in Chapter 3. The next section gives an introduction to social network analysis as a means of representing the structures and dynamics of interaction networks. Starting with a historical and methodological classification, the terms required for this work from the fields of graph theory and network analysis are defined. In addition, this section is supplemented with an account of change detection procedures. Chapter 3 ends with final remarks on the methods used. Chapter 4 reflects on the semi-structured nature of social media data and the challenges presented by this format. Each social media platform has technical and structural characteristics tailored to a specific communication format and designed to faciliate the distinctive types of communication interactions on these platforms. Chapter 5 presents the published papers and articles relating to our central questions. The first article deals with the statistical analysis of large-scale web-based data with a focus on the methodological requirements for the statistical evaluation of this new form of data. The second article is a technical report in which we present, on the one hand, a concept for anonymizing text data and, on the other hand, the technical aspects of the tool itself. The next two papers are thematically linked to the preceding discussion. These papers

investigate and evaluate communication structures and behavioral changes in negative word-of-mouth communication. The fifth paper analyzes and evaluates user interaction behavior in political subreddits, as part of the theoretical section on polarization. Papers six and seven address the linguistic evaluation of hate speech, including considerations of radicalization and communicative aggression. The focus in these two papers is on developing a taxonomy to capture hate in its various linguistic facets.

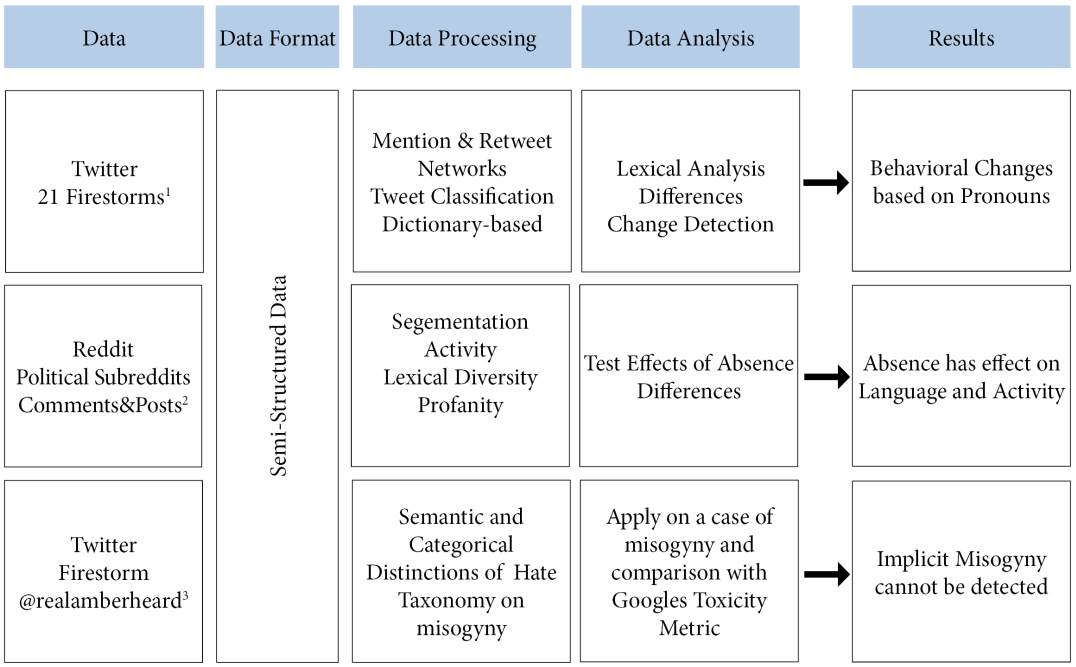


Figure 2: Data research design

Figure 2 summarizes the steps of analysis. The final research questions regarding communicative structures are very comprehensive and to some extent involve implications that go far beyond the scope of this study, which means these questions cannot be addressed conclusively. Instead, new hypotheses should be formulated based on the findings of this study. Chapter 5 concludes with a summary response to the research questions listed at the beginning of this study. Chapter 6 addresses the methodological issues arising in a study of this kind. In addition, questions are mentioned that raised during this thesis but have not been addressed.

1.4 Innovation

The overarching focus of this dissertation is a multiple perspective on the phenomenon of online communication conflicts. We hereby classify approaches for analysis of hate speech and communication conflicts.

Insights for the automated Detection of Negative Word-of-Mouth

The first two papers that have been integrated in this thesis (cf. Chapter 5) are concerned with the detection and prediction of online firestorms on Twitter (Strathern et al., 2020b) and (Strathern et al., 2022b). We ask, how can we detect and predict online firestorms? In Strathern et al. (2020b) we define different points in time, examine language use with a psycho-linguistic dictionary for these points, and look for differences. As a result, we conclude that negativity in words increases and that users switch from using the pronoun "I' to the third-person pronoun. In a nutshell, we observe a sudden change in user perspective. To account for linguistic cues, we construct a change detection model. In Strathern et al. (2022b), we combine network and linguistic features for machine learning to build a model that predicts the start of firestorms. The models allow for automated detection and prediction of firestorms.

Insights for Content Moderation on Platforms

The effect of absence on activity in political communities on Reddit is studied in Strathern et al. (2022a). For this study, we collected data from three subreddits, two of which were banned due to continuous violation of community rules. To examine the effects, we ask if a temporal absence from these subreddits changes the level of activity, the level of diversity, and the level of profanity in language use. We develop an experimental setup to test the effects. Results show that users who are continuously active increase in activity, whereas users who are absent for a while decrease in activity. For each group we test for changes in the linguistic style. We apply a metric for lexical diversity and count for profanity words. The results show that users who are continuously active decrease in diversity whereas users who are absent for a while increase in diversity. In accordance with ongoing research on the impact of social media on polarization and radicalization, these initial findings open the door to more work on experimentally testing effects that could slow down or interrupt escalation. Further research could investigate intervention techniques on social media platforms in more detail.

Insights for the automated Detection of Hate Speech

The last paper deals with misogynistic speech on Twitter (Strathern and Pfeffer, 2023). For this study, we ask if automated methods to detect toxic language can identify misogyny. The actress Amber Heard was the target of a firestorm on Twitter after accusing her then-husband Johnny Depp of domestic violence. We collected Twitter data containing the mention realamberheard. To better understand misogyny, its detection and modeling, we conduct a literature review in which we perform a content analysis of the top 1000 hateful retweets. In concordance with existing classifications and taxonomies on misogyny and hate speech we develop a schema that identifies explicit and implicit misogyny. We apply the schema to the top 5000 retweets from this dataset and annotate it manually. In the last step, we compare our manual coding with a toxicity measurement from Google, the Perspective API. Results show that Google does well in detecting explicit hate but performs badly in detecting implicit misogyny. In another paper (Wich et al., 2021) we develop a classification schema to determine different categories of hate from German right-wing users on Twitter. This dataset is the basis for a structural network analysis of abusive online behavior.

Chapter 2

Conceptual Background

Hate speech is often considered and defined as unambiguous as long as it contains commonly accepted swear words. However, conflicts in communication and the violence contained therein are expressed in various forms and result from different social processes. To understand this range, we explain here the conceptual background as applied in Strathern and Pfeffer (2020); Strathern et al. (2020b, 2022b,a); Strathern and Pfeffer (2023).

2.1 Social Media Networks

Social network sites are "a networked communication platform in which participants 1) have uniquely identifiable profiles that consist of user-supplied content, content provided by other users, and/or system-provided data; 2) can publicly articulate connections that can be viewed and traversed by others; and 3) can consume, produce, and/or interact with streams of user-generated content provided by their connections on the site" (Ellison and Boyd, 2013, p. 158). The authors further elaborate that the main reason people use social network sites (SNSs) is to communicate and share content, which is made possible by various communication-oriented features. SNSs are particularly good at facilitating communication and sharing because they lower the barriers to doing so, making it easier for people to build and maintain networks of connections. This is especially true for weak tie relationships, which would be likely to fade away if it were not for the ease of communication and sharing on SNSs. SNSs are primarily communication platforms, but the authors also emphasize the importance of sharing content, usually in the form of a stream. SNSs support a range of communication modes, including one-to-many and one-to-one, synchronous and asynchronous, textual and media-based. These features can be private or public,

depending on the site. SNSs provide opportunities for users to communicate not only with their own networks of friends, but also with their friends' network, which can provide access to novel information and diverse perspectives. SNSs offer a semi-public forum for communication which can be more productive than other online forums where accountability and motivation are lower. Most SNSs are organized around a stream of recently updated content, with spaces for media sharing being nearly universal. Profiles are less significant in the user experience than they used to be, but they still serve as spaces for self-presentation and content distribution. Communication and information sharing are now the main motivators for participating in SNSs (Ellison and Boyd, 2013).

2.1.1 Structural Features

According to Labianca et al. (2013) social network research sought to create a framework based on two axes: "explanatory goals (social homogeneity or performance variation) and explanatory mechanisms (network content or structure)" (Labianca et al., 2013, p. 4). One of these types is environmental shaping, which examines how the network surrounding can exert a predictable influence on its associates. The spread of resources and their impact on nodes within a network is known as contagion. Structural capital refers to the advantageous or constraining connections between people, while resource access relates to how nodes acquire and derive advantages from the resources accessible within a network. Scientists endeavor to clarify the significance of social networks by examining either the material that traverses the networks or the configuration of the networks, which enables them to obtain or control significant resources. The term "content" refers to the resources that are present in a network such as knowledge, rumors, finances, or even infections. Meanwhile, the term "structure" refers to the recognizable arrangements of nodes and links within a network (Labianca et al., 2013). Regarding social media, network content is the content generated by users offering information, influence, or social support (Labianca et al., 2013). By representing connections or interactions in a structured data format through a computerized platform, social media platforms measure and systematize relationships among nodes. This formalization enables social media networks to possess relational functionalities that are absent in face-to-face social networks, such as the effortless visualization and analysis of connections (Labianca et al., 2013). The system's features enable and limit its users in certain ways, leading to similar behaviors among those who use the system; within the realm of so-

cial media, these features may be technical (such as the platform's capabilities), normative (such as its rules and guidelines), or economic (such as incentives for certain usage patterns). However, users may choose to utilize or react to these functionalities in diverse ways, leading to variations in performance among users of the same platform (Labianca et al., 2013). What is even more intriguing is that the functionalities of the social media platform, as an environmental element, create social consistency (which explains why users of the same platform behave in similar ways), whereas user behavior is the primary factor responsible for variations in individual performance (which clarifies why users of the same platform perform differently) (Labianca et al., 2013). Labianca further elaborates that social media networks are characterized by four crucial features. Firstly, users create a distinct user profile. The second feature of social media networks, according to Labianca et al. (2013), is that users can access digital content and safeguard it from search tools provided by platforms. The third feature of social media networks is that users can create a list of other users with whom they have a relationship, while the fourth feature is that users can explore and navigate their connections, as well as those formed by other users on the platform (Labianca et al., 2013). o handle their social connections on social media networks, users must understand and manipulate the network's structure, which includes how connections are formed and sustained within it. Digital profile creation, access, and security of platform-generated content primarily pertain to network content, which refers to how digital resources are shared and accessed within the network (Labianca et al., 2013). The purpose of social media network research is to uncover the inner workings of how connections are formed on these platforms, commonly referred to as the "black box." The objective is to scrutinize how the different design decisions taken by social media platforms influence and constrain user actions in predictable manners, eventually affecting outcomes of social media networks (Labianca et al., 2013). As per Labianca, connections between individual members, known as ties, can comprise different types of links between nodes, and Social Network Analysis research identifies four fundamental tie types that could be relevant to the design of social media networks: proximities, relations, interactions, and flows. Proximities denote familiar physical or social contexts, such as physical proximity or group membership, that create opportunities for ties to emerge; Relations involve enduring social connections between nodes, such as those based on roles (e.g., friends, family) or affective ties (e.g., likes, dislikes); Interactions are distinct, momentary relational episodes,

such as sharing a meal or signing a contract with another node, that can generate or transform relations. Last but not least, flows indicate concrete and abstract substances, such as currency, commodities, data, or convictions, that can transfer from one node to another as nodes engage with each other (Borgatti et al., 2018).

2.1.2 Interaction and Interactivity

To better understand the network implications of user behavior and performance variation, we need to take closer look at driving factors for relational connections. Social media platforms do not just offer a place for building social relations but follow a business model whose purpose is to keep users interacting with brands, people, and topics. Trunfio and Rossi (2021) summarize different categories of social media engagement metrics from a broad literature review. According to their study, social media engagement can stem from various sources, such as interactions with a community of other users in the network and with brands. Researchers have also studied the underlying factors and consequences of social media engagement, with a view to understanding the motivations behind user interactions on social media and the potential outcomes, such as increased loyalty, satisfaction, trust, and commitment to brands and communities (Trunfio and Rossi, 2021). Continuing the tradition of customer engagement research, social media engagement is also said to be composed of affective, cognitive, and behavioral elements. Most studies focus on the behavioral aspect as it can be demonstrated through actions such as liking, commenting, sharing, and viewing content. According to Trunfio and Rossi (2021) social media engagement can be broken down into three dimensions: consumption, contribution, and creation. Consumption is the most basic form of engagement, where users simply consume brand- or topic-related media such as videos, audio, or images. Contribution refers to user participation in peer-to-peer interactions with brands, people, topics; for example, by liking, commenting, or sharing brand content. Creation is the most advanced form of engagement, where users actively participate in discussions by publishing their own content, videos, audio, images, or articles related to a specific topic. Several methods have been developed for measuring social media engagement such as scales, indexes, and metrics.

2.2 Social Processes

The idea of social processes pertains to the recurring forms of social interaction. The interaction between individuals and groups is essential to social life, and social processes refer to the repeated forms of this interaction. Social interaction involves reciprocal relationships that affect not only the individuals involved but also the quality of the relationships. Social interaction encompasses all types of dynamic social relations between individuals or groups (Gillin and Gillin, 1950). Social interaction can be defined as the process in which meaningful contact between two or more people results in behavioral modifications. Mere physical proximity is not enough to create a social unit or group. Social interaction is the dynamic interplay of forces resulting in behavioral and attitudinal modifications among the participants. The two fundamental conditions of social interaction are social contact and communication, with social contact being the initial phase of interaction (Gillin and Gillin, 1950).

2.2.1 Negative Word-of-Mouth

According to the definition proposed by Gordon Allport and Leo Postman, a rumor can be defined as a statement or claim that is transmitted from one individual to another, often verbally, without any solid evidence to back it up (Allport and Postman, 1947). While online firestorms may share certain similarities with rumors, such as being based on hearsay and uncertainty, they present distinct challenges due to the rapid pace and widespread reach of social media interactions (Strathern et al., 2022b). The distinguishing feature of firestorms is the high level of aggression, which gives them their name. Even though customer criticism can sometimes spark negative comments, the later stages of the conversation often contain language that is intended to be insulting without any meaningful substance or logical reasoning (Pfeffer et al., 2013). Firestorms may be based on either unconfirmed rumors or confirmed events, but they are predominantly opinion-based rather than fact-based, making them highly affective (Pfeffer et al., 2013). Within a socio-technical system, social media users engage through AI-based algorithms that mediate and propel the system; the primary objective of social media platforms is to maintain user engagement and maximize their platform use time, as this is crucial to the platform's business model that relies on selling ads, which is accomplished most effectively with highly-engaged users who spend a significant amount of time on the platform

(Strathern et al., 2022b). Nonetheless, the crucial question remains: What content would capture a particular user's interest and sustain their engagement? To this end, recommender systems are engineered to heighten the probability of a user clicking on suggested content and interacting with it; these algorithms consider sociodemographic details, a user's prior actions, and behavioral data from their acquaintances or "alters" (Strathern and Pfeffer, 2020; Strathern et al., 2022b). Social scientists have studied the underpinnings of social connections for a considerable period, including why people form social ties (Strathern et al., 2022b). Network formation hinges on two critical factors: homophily, as highlighted by McPherson et al. (2001), which implies that friends share similarities and interests, and transitivity, which denotes that a person's friends are typically connected. Consequently, most individuals are enmeshed in personal networks comprising like-minded, well-connected individuals. According to Pfeffer et al. (2013), there are several observations and generalized factors that contribute to negative Word-of-Mouth dynamics: A constant flow of information with a short information half-life, which affects the speed and volume of communication (Burton and Kebler, 1960). Absence of discursive interactions results in binary choices and no gradualist opinions (Schelling, 1973). As posited by Heider (1946), clustering within social media networks augments the spread of epidemics, resulting in increasingly dense network clusters. Weak and strong ties are blurred and the practice of having hundreds of "friends' creates information, resulting in an unrestrained flow of information (Granovetter, 1973). Limited information caused by homophily, where friends act as a filter, results in a lack of diversity (Simon, 1972; McPherson et al., 2001). There exist cross-media dynamics as offline and online media mutually reinforce each other. Ultimately, this leads to alterations in the opinion adoption process, whereby network-triggered decision processes are propelled by dominant network effects, as Rogers (1995) posited. Regrettably, the forces that drive human network formation, combined with AI-powered recommender systems, have troubling implications (Strathern and Pfeffer, 2020). Recommender algorithms used on social media platforms influence the content and connections suggested to users, giving rise to filter bubbles that envelop individuals with familiar content and like-minded people, as Pariser (2011) noted, which can lead to polarization. The ephemeral nature of social media communication further exacerbates this issue, leaving little room for nuanced discourse and creating an ideal setting for online firestorms. For instance, a small group of people expressing dissatisfaction with a politician, celebrity, or private individual

may find agreement from their like-minded peers, who can quickly disseminate the negative sentiment through reposting or retweeting, potentially reaching tens of thousands of users within a matter of hours (Strathern and Pfeffer, 2020).

2.2.2 Polarization

2.2.2.1 Types of Polarization

Polarization often refers to a form of "splitting (into two camps or similar), in which the differences stand out clearly" or the "formation of opposition" (DiMaggio et al., 1996). From a communication studies point of view, for example, two central forms can be distinguished. Both concepts share the approach that polarization represents the advocacy of a strong political position (Fiorina et al., 2010). Topic-based polarization focuses on a person's attitude or stance towards a specific political topic or a specific issue in general. When this attitude is or becomes extreme, it is referred to as political polarization. Thus, polarization can be illustrated here as a process in which one's own opinion changes from initially moderate positions to more extreme positions, or as a state that describes an already polarized opinion. It is important at this point to differentiate topic-based polarization from extremism. The latter not only involves extreme political attitudes or ideologies, but also the rejection of the democratic state and the willingness to abolish it (Gaspar et al., 2020). Group-based polarization, sometimes referred to as affective polarization, deals with the evaluation of entire political groups such as parties or other issue-based groups by individuals (Bail et al., 2017). Specifically, it involves the tendency to have sympathy for the political ingroup and, at the same time, a strong antipathy towards the political outgroup. To capture people's attitudes towards entire groups, questions are asked not about personal opinions on political issues, as in topic-based polarization, but about attitudes towards different social groups. This can be done, for example, with a so-called "Feeling Thermometer", in which voters indicate how strongly they feel positive or negative emotions towards the (own) ingroup compared to the (foreign) outgroup. The evaluation of this information is used to determine how benevolent or antagonistic two or more political groups are towards each other. Selective Exposure refers to Festinger's theory of cognitive dissonance (Festinger, 1957). According to the theory, people strive for a positive and consistent self-image. This self-image is achieved when a person acts in accordance with their attitudes. When two cognitions contra-

dict or conflict with each other, the state of cognitive dissonance occurs, which is perceived as unpleasant by humans. In such situations, the person is motivated to end this state and restore the desired consonance between cognitions. The theory of cognitive dissonance explains the phenomenon of people perceiving information that aligns with their pre-existing attitudes. The echo chamber theory sees polarization as a result of a lack of confrontation with dissonant information. In the context of political communication, this human tendency can manifest itself in people surrounding themselves only with other individuals or exposing themselves to media that aligns with their political views, while avoiding people or information that contradict their own cognitions. Political face-to-face communication refers mainly to non-medial context, as a significant proportion of political communication takes place face-to-face - at work, in clubs, or at home. People therefore surround themselves with others who are similar to them, as this is conducive to cognitive consonance. Homophilic mechanisms may lead to the formation of politically homogeneous groups instead of the exchange of different ideas and opinions. The members of a group reinforce each other through sharing and exchanging the same opinions and information. This is accompanied by the concept of group polarization. On social media platforms even the smallest interest groups can connect, not only with minimal effort but also across geographical distances. Homophilic associations can therefore lead to the emergence of echo chambers. The possibility of social networking on social network systems can possibly intensify polarization. The concept of echo chambers aims to describe virtual spaces where only like-minded individuals in terms of behavior and opinions are present. There are different types of political polarization: ideological polarization and affective polarization (Hohmann et al., 2023). Ideological polarization occurs when political adversaries hold differing opinions, beliefs, attitudes, and stances. Affective polarization is driven by the role of identity in politics and the impact of in-group identity on animosity towards out-groups (Iyengar and Hahn, 2009). This type of polarization measures the degree to which individuals "feel warmth" toward their political allies and a "lack of warmth" toward their political opponents (Iyengar and Hahn, 2009). Both types of polarization can be observed in social media platforms and have implications for the functioning of democratic societies (Sunstein, 2002; Tsfati and Chotiner, 2016). Although high levels of polarization can be beneficial in promoting political participation and electoral choice, political polarization can also have detrimental effects on democracy. These negative effects include the

centralization of power, congressional gridlock, and decreased citizen satisfaction (Wagner et al., 2017). Previous studies have shown that polarization can lead to negative interpersonal consequences, such as an unwillingness to interact with political opponents and the dehumanization of them (Mason, 2015). The media is increasingly important in shaping people's impressions of opposing groups, which can significantly impact their perception of the political landscape. As groups become additionally divided, and subgroups emerge, the system becomes more polarized. Polarization is a dynamic process that typically starts with individuals discussing their opinions with those in their immediate social circles. Individuals may become more polarized and distance themselves from others when encountering disagreement and opposing viewpoints resulting in behavior changes, such as switching political affiliations and forming new social groups that reinforce pre-existing beliefs. The phenomenon of polarization has been studied in relation to social tension, rebellion, and unrest (Esteban and Ray, 1994). Conflict and inter-group dynamics are fundamental to understanding societies (Simmel, 1955; Tilly, 1987). Social media is an influential tool in shaping people's beliefs in various aspects of life, including marketing (de Vries et al., 2012; Chang et al., 2015), entertainment (Williams and Ho, 2016), and politics (Benkler, 2006; Chadwick et al., 2015; van Dijck and Poell, 2015). In recent political events, social media has played a significant role (Howard et al., 2011; Davis, 2017; Groshek and Koc-Michalska, 2017). At the individual level, polarization is associated with strengthening existing predispositions and attitudes (DiMaggio et al., 1996; Tsfati and Chotiner, 2016; Stroud, 2010; Harris et al., 2014). Polarization was identified as a factor in elections (Allcott and Gentzkow, 2017). Though the role of social media platforms in polarization is unclear, there is an ongoing debate about whether social media platforms increase or decrease polarization among its members or society in general (ElSherief et al., 2018; Kleiner, 2018; Barberá, 2014; Du and Gregory, 2016). One view suggests that social media is increasing polarization by putting users in echo chambers (Bail et al., 2018a) and filter bubbles (Hong and Kim, 2016) through their recommendation algorithms. The counter-view is that social media decreases polarization by allowing users to express their views freely, thus exposing them to cross-cutting content (Garrett et al., 2013; Mutz and Mondak, 2006). Many researchers have focused on discovering echo chambers in news media use (Iyengar and Hahn, 2009), blog readership (Lawrence et al., 2010), Twitter (Garimella et al., 2018), and Facebook (Bakshy et al., 2015), thus showing how social media contributes

to increasing polarization. However, the past work has yet to concentrate on establishing the exact role of social media in polarization, comparing two contrasting processes - comparing echo chamber processes that increase polarization with cross-cutting discussions that decrease polarization.

2.2.2.2 Measuring Polarization

The model of opinion formation studies the issue of polarization in society (Dandekar et al., 2013). An opinion formation process is polarizing if it leads to an increased divergence of opinions. Bramson et al. (2016) proposed an improved approach to measure polarization in attitudes by distinguishing between nine different types of polarization and developing specific metrics for each one. A study of public opinion polarization shows that it encourages individuals to attend lawful demonstrations (Kleiner, 2018). Researchers study polarization to address whether Americans' opinions have become more dispersed due to increased ideological constraints and whether groups change their opinions over time (DiMaggio et al., 1996). Multiple research studies have confirmed the significant role played by social media in shaping political discourse (Sunstein, 2007; Aday et al., 2010; Tumasjan et al., 2010). Adamic and Glance (2005) found that political blogs tend to link more frequently to other blogs sharing the same ideology. As previously mentioned, there are two competing hypotheses related to polarization and social media's contribution. The first hypothesis suggests that social media increases polarization: Conover et al. (2011) studied political polarization on Twitter by analyzing retweets and mention networks. The authors find that retweet networks are ideologically separating, while the mention networks are not. This result indicates that users who agree with the same ideology retweet, but users who disagree with the ideology of the original tweet might only mention the tweet, thus creating a cross between the two different ideologies. Hong and Kim (2016) find that politicians with extreme ideologies had a much greater readership on Twitter, suggesting the presence of echo chambers on the platform. Bail et al. (2018b) conducted experiment to study how exposure to opposing views on social media can affect political polarization, finding that exposure to opposing views increased polarization among Republicans and Democrats. On the other hand, Gruzd and Roy (2014) and Hong (2013) present a competing hypothesis that social media decreases polarization by facilitating exposure to diverse viewpoints. Barberá (2014) also argues that social media can reduce mass political polarization by providing exposure to political diver-

sity. Findings are supported by research on the effect of cross-cutting exposure by Allport et al. (1954); Mutz and Mondak (2006), which suggests that exposure to political diversity positively affects political moderation. However, it is still unclear whether political polarization arises as a consequence of discussion or already exists beforehand (Cheng et al., 2014).

2.2.3 Radicalization

The concept of radicalization is often associated with terrorism and violence, which makes it challenging to track individual stages of radicalization systematically. However, Gaspar et al. (2020) proposes an approach to address this issue by differentiating radicalization from extremism and terrorism and separating it from violent actions. This broader definition of radicalization, distinct from violence, is necessary to understand the process better. The concept of extremism pertains to a state, whereas radicalization is a process. Extremism is characterized by rejecting democratic values and the constitutional state. Radicalization, on the other hand, involves an increased questioning of established norms. Recent research has explored three perspectives on radicalization: radicalization leading to violence (Porta and LaFree, 2012; Crossett and Spitaletta, 2010; McCauley and Moskalenko, 2008; Moghaddam, 2005), radicalization through violence (Schmid, 2013; Morrow, 2017), and radicalization without violence. The focus of interest for this work is on the last approach—radicalization without violence. People and groups pursue their objectives through non-violent means but growing inclined to challenge the current system. Current research on radicalization tends to focus primarily on radicalization involving violence, making it challenging to understand radicalization that includes non-violent forms (Gaspar et al., 2020). In some cases, non-violent radicalization acknowledges a distinction between attitudes and actions (Gaspar et al., 2020). The authors provide additional examples of this distinction, such as the differentiation between "cognitive radicalization" and "violent radicalization" (Vidino, 2013, p. 11-12) or "behavioral radicalization" (Neumann, 2013, p. 873). They emphasize the separation between radical thought and radical action (Fishman, 2010), highlighting that subscribing to radical beliefs does not necessarily lead to engaging in extreme actions (Bartlett et al., 2010). Another approach to distinguishing between these concepts is to separate them into attitude and action levels, as radicalization refers to the development of extremist beliefs. In contrast action pathways refer to engaging in violent actions. (Borum, 2011).

2.3 Language and Communication

Language is the medium through which communication happens. This section is concerned with the structure of language in its various levels and its statistical structure. Additionally, it discusses lexical semantics, a subfield in the study of the language system that is particularly relevant to the studies in Strathern et al. (2020b, 2022b). This is followed by a discussion on the topic of communicative aggression. Verbal violence constitutes its own area of communication research and linguistics and is especially relevant in Strathern and Pfeffer (2023).

2.3.1 The Structure of Language

2.3.1.1 Linguistic Levels

To approach linguistic data methodologically, we briefly explain linguistic levels and how they are interconnected according to David Crystal's Cambridge Encyclopedia of the English Language. Language models often recognize a distinction between the physical forms of a language (such as sounds, letters, signs, and words) and the abstract meanings conveyed by those forms (Crystal, 2010). However, this distinction is often further divided to differentiate various types of abstractness. The study of pronunciation, involving the processes of articulation, acoustic transmission, and audition, is known as phonetics. Phonology refers to how different languages use sounds to convey meaning. Grammar is the study of how meaningful units are arranged to convey broader and more varied patterns of meaning, while the study of the patterns of meaning themselves is known as semantics. While four-level language models are often used (phonetics, phonology, grammar, semantics), further divisions within and between these levels are expected. For instance, morphology (the study of word structure) is often distinguished from syntax (the study of word sequence within sentences) within grammar. Segmental phonology (the study of vowels, consonants, and syllables) is often separated from suprasegmental phonology (the study of prosody and other stress and intonation patterns) within phonology. Similarly, vocabulary (or lexicon) is sometimes studied separately from more extensive patterns of meaning (such as text or discourse) within semantics. These divisions are commonly referred to as levels of structure (Crystal, 2010). It is impossible to analyze one language level without considering the assumptions made about other levels. For example, when examining phonetics, one must take into account the importance of certain sounds within a language (phonology) and how

they are used to distinguish words (grammar) and convey meaning (semantics). Likewise, when examining the structure of sentences in grammar, it is crucial to take into account the connections of significance (semantics) and the characteristics of pitch and tone (phonology) that aid in identifying sentence components in speech (Crystal, 2010).

2.3.1.2 Statistical Structure

Linguistic structure can be analyzed by counting the units that occur and examining the statistical regularities in their use (Crystal, 2010). Research into various aspects of language, such as grammar, vocabulary, sound systems, and writing systems, has uncovered a range of patterns. Some of these patterns, known as "universals" (Crystal, 2010), are statistically common across all languages and independent of the speaker, writer, or topic. Linguistic behavior generally adheres closely to statistical expectations. For instance, in English, it is highly probable that a "q" will be followed by a "u," and that consonants make up around 60% of all spoken language, while vowels make up about 40%.Furthermore, it has been observed that roughly 50% of the syllables we employ in everyday speech follow the structure of consonant + vowel + consonant, as exemplified by the word "cat". Moreover, the top 50 most frequently used words in a language constitute about 45% of all written material (Crystal, 2010). The study of linguistic structure at different levels, such as grammar, vocabulary, sound system, and writing system, can reveal statistical regularities and patterns in language usage. Statistical linguistics is the study of statistical properties present in any large sample of speech or writing and the factors that influence them. One of the earliest discoveries of significant statistical patterns in language was made by George Kingsley Zipf, an American linguist. He is particularly known for his "Zipf's Law," which posits a fixed relationship between the position of a word in a list of frequently used words and the frequency with which it appears in a given text. Zipf discovered that there is a correlation between the frequency of a word's usage and its length. He found that shorter words are used more frequently than longer words in languages like English and German. This pattern can also be seen in the tendency to shorten words when they are frequently used, such as the transformation of "microphone" to "mike." Zipf believed that this pattern occurs because shorter and simpler words are easier to communicate, and this principle of "least effort" (Zipf, 1949) explains the balance between diversity and uniformity in language. Despite the explanation that simpler sounds and

shorter words are more common in human communication, limitations such as the difficulty in measuring the effort involved in speaking and the existence of exceptions to the rule have been identified. Consequently, researchers typically rely on a conventional explanation based on probability theory (Crystal, 2010).

2.3.2 Lexical Semantics

Segmentation in linguistics refers to the process of breaking down complex units (such as sentences or words) into their constituent elements (segments), which are then classified according to specific criteria - their meaning and/or function. Segmentation is used for the analysis of linguistic units. Classification of the components of linguistic units, i.e. their assignment to specific categories, is based on segmentation. Segmentation and classification presuppose that language is an organized system whose elements are related to each other in certain ways. According to Ferdinand de Saussure, the founder of structural linguistics, segmentation and classification are methods that characterize structuralist linguistics (Bußmann, 2008). In linguistics, semantics is the study of meaning in language. Its focus lies on the usage of words, phrases, and sentences to communicate meaning and the multiple factors that affect the comprehension of these meanings. According to Crystal (2010), semantics is a complex and multifaceted field that encompasses a wide range of sub-disciplines, including lexical semantics (the study of word meaning), conceptual semantics (the study of the relationships between words and concepts), and discourse semantics (the study of the meaning of texts and conversations). Words can have multiple meanings depending on the context in which they are used, and the interpretation of a word's meaning is influenced by a variety of factors, including the words that come before and after it, the tone of voice in which it is spoken, and the cultural and social context in which it is used. Another important aspect of semantics is the concept of reference, which is the relationship between a word or phrase and the thing it refers to. Words can be used to refer to concrete objects (such as "tree" or "book"), abstract concepts (such as "happiness" or "justice"), or even other words (such as "synonym" or "antonym"). Semantics focuses on how words and phrases are used to convey different types of meaning, such as literal meaning (the direct meaning of a word or phrase), figurative meaning (the implied or symbolic meaning of a word or phrase), and contextual meaning (the meaning of a word or phrase as it is used within a specific context) (Bußmann, 2008). Grammar can be compared to something like a 'framework'. However, these physical

comparisons do not fully convey the various types of formal patterns and abstract connections that are revealed through grammatical analysis. The study of grammar typically involves two phases. The first identifies the building blocks of language, such as words and sentences. The second phase involves analyzing the ways in which these building blocks are arranged and the meanings conveyed by these arrangements. The way in which grammar is defined can vary depending on the units that are recognized at the start of the study. Most linguistic approaches start by acknowledging the fundamental unit of language, namely the sentence, and grammar is typically defined as the investigation of how sentences are structured. According to this perspective, grammar for a language is a description of the possible sentence structures in that language, which are organized based on general principles. "Grammar" can be used in different ways, with one being more specific and traditional, while the other is more general. In a specific sense, grammar is considered to be just one aspect of language structure, separate from the study of phonetics and semantics. Chomsky popularized the broad meaning of the term *linguistics* which includes all aspects of sentence structure such as phonology and semantics, and he also introduced the more specific term "syntax" to refer to the study of sentence patterning [1]. Syntax - the system of rules that govern the structure of language - is a key component of grammar. Syntax is concerned with the way words are combined to form phrases and sentences, and with the rules that govern the arrangement of these words. Syntax helps to understand the meaning of a sentence and to communicate effectively (Crystal, 2010). There are several key concepts that are central to the study of syntax. One of these is the concept of a clause, which is a group of words that contains a subject and a verb. A clause can be either independent (able to stand alone as a sentence) or dependent (needing another clause to complete its meaning). Another important concept in syntax is the concept of a phrase, which is a group of words that functions as a single unit within a clause or sentence. Phrases can be either noun phrases (which function as the subject or object of a verb), verb phrases (which contain the verb and any other words that are necessary to complete the verb's meaning), or prepositional phrases (which contain a preposition and the noun or pronoun that follows it). In addition to clauses and phrases, syntax also involves the use of word order and punctuation to convey meaning. This is a limitation to the study of written language. The order in which words are placed within a sentence can affect the meaning of the sentence, and punc-

[1] https://www.britannica.com/science/linguistics/Dialectology-and-linguistic-geography

tuation is used to indicate the structure of a sentence and to clarify its meaning. Morphology is the study of the structure of words and involves the examination of the ways in which words are formed and the rules that govern this process. It is a key branch of linguistics that helps us to understand the way in which words are built up from smaller units and to analyze the internal structure of words in different languages. Morphology is concerned with the structure of words and the way in which they are formed from smaller units called roots and affixes. Roots are the core part of a word that carries its main meaning, while affixes are added to the front or back of a root to change its meaning or grammatical function (Crystal, 2010). In simpler terms, morphology is the study of the way words are formed and changed to convey different meanings in a sentence. This field is traditionally divided into two parts: inflectional morphology, which examines how words change to indicate grammatical contrasts, such as singular and plural forms, and derivational morphology, which looks at how new words are created without considering their grammatical roles in a sentence. Morphology is also concerned with the study of inflection, which is the way in which words change their form to indicate tense, number, gender, or case. For example, in English, the verb "walk" changes to "walked" to indicate the past tense, and the noun "child" changes to "children" to indicate the plural form. In addition to these concepts, morphology is also concerned with the processes of derivation and compounding, which are used to create new words. Derivation involves the addition of affixes to existing roots, while compounding involves the combination of two or more existing words. Words are typically the simplest units to locate in written language because they are usually separated by spaces in most writing systems. Since the beginning of the study of grammar, words have been classified into different categories, commonly known as the "parts of speech". Typically, eight categories are recognized, such as nouns, pronouns, verbs, adjectives, prepositions, conjunctions, adverbs, and interjections in the case of the English language (Crystal, 2010). The study of grammar focuses on identifying structural patterns and relationships in language, rather than relying on definitions based on meaning. In particular, word classes are determined on the basis of how words behave and interact within a language.

2.4 Communicative Aggression

According to Vangelisti and Hampel (2010) communicative behaviors refer to psychological constructs that impact how individuals express their feelings, needs, and thoughts using indirect messages instead of direct and open communication. Non-verbal communication is a significant aspect of communicative behaviors. Essentially, any behavior or its absence can be considered communicative if it intends to convey a message. For instance, a particular hairstyle, a display of emotions, or actions such as doing or not doing the dishes can be used to convey messages (Dailey et al., 2007). People vary in their preferred communication styles, with some being more inclined towards indirect or behavioral communication, even when verbal options are available. This behavior can be conscious or unconscious, significantly impacting an individual's verbal and nonverbal communication patterns. People usually use a mix of behavioral communication styles rather than all of them. Self-awareness is crucial to understanding one's behavioral style (Platt et al., 2016). Communication behavior can be classified into four types: aggressive, assertive, passive, and passive-aggressive. Aggressive communication involves random acts of anger with the intention to harm someone or something. Aggressive communicators often engage in personal attacks and put-downs, which creates avoidable conflicts and leads to a win-lose situation. They use intimidation in their dealings with others; they lack empathy and believe that power and control are the only ways to achieve their goals. Close-mindedness, poor listening skills, and monopolizing others are typical characteristics of aggressive communicators (Dailey et al., 2007). Aggressive communication is characterized by several typical behaviors, such as belittling others, dominating or exerting power over them, failing to express gratitude or acknowledgment, using coercion or pressure to get what one wants, disregarding others' emotions, speaking condescendingly, instilling fear or intimidation. Nonverbal cues like frowning, critical glares, and speaking loudly are commonly used during aggressive communication. This type of communication usually leads to resistance, counter-aggression, and estrangement, with the receiver often experiencing feelings of hurt, fear, defensiveness, humiliation, and resentment. However, there may be certain circumstances where aggressive communication is necessary, such as during emergencies or when prompt decisions need to be made (Platt et al., 2016). Hate against certain groups such as racism, sexism, homophobia, transphobia, misogyny, antisemitism, and Islamophobia manifests in different ways, but the act itself is a boundary violation. According to ongoing

research hate speech has a purely signal-functional character: it represents verbal violence whose objective is to hurt, wound, and express negative evaluations in the form of vulgar and offensive language (Gagliardone et al., 2015). These are linguistic expressions whose purpose is to harm, apply violence, and establish a power structure. Generally, from a legal perspective, hate speech is any form of expression in which the speaker intends to defame, degrade, or incite hate against a group, class of individuals, or person based on race, religion, skin color, sexual identity, gender identity, ethnic affiliation, disability, or national origin. Hate speech can be categorized as statements that are intended to incite harm (specifically discrimination, hostility, or violence) and are made on the basis of the target person's membership in a particular social or demographic group. Another category encompasses statements that promote a climate of prejudice and intolerance, under the assumption that this can lead to targeted discrimination, hostility, and acts of violence (Gagliardone et al., 2015). In linguistics, hate speech is discussed on different levels. Scharloth (2017) argues that the concept of invective is a key concept in understanding the various forms of language use that can be considered hate speech. Invective refers to the use of verbal or non-verbal communication that is intended to negatively evaluate an individual or group and can lead to discrimination or exclusion. This concept encompasses a wide range of everyday phenomena, such as rudeness, insult, verbal aggression, and hate speech. Scharloth notes that invective can take many forms, including the use of swear words, pejorative expressions, and accusatory intonation, and it is often subtle and disguised (Scharloth, 2017). He also emphasizes that invective can have serious consequences for the individuals and groups who are targeted, and is a complex and multi-faceted phenomenon that cannot be reduced to a simple definition. Linguistic research has engaged with the concept of invective from a variety of perspectives, including those of lexical-semantics, speech act theory, conversation linguistics, and post-structural discourse analysis. An invective statement has also been understood as a speech act in the sense of Austin (1962) in linguistic research. Linguistic demeaning and exclusion is then a speech act of attributing a negatively evaluated social category. (Scharloth, 2017) summarizes that other invectives have been conceptualized by linguistics as rudeness, modeled as a face-threatening act - as a mirror image to the politeness theory by Penelope Brown and Stephen C. Levinson. The central anchor of these theories is the concept of the face from face from Erving Goffman. In interaction, those involved are normally trying to approve and support the behavioral strategies of

the other interactants and assign them a consistent self-image and therefore a positive social value. In this perspective, linguistic violence consists of crossing or explicitly negating the behavioral strategies of the interaction partners. The numerous language actions for realizing invective that are available in a language community can be partially identified using communication verbs; for example, offend, insult, defame, discredit, hurt, attack, humiliate, annoy, provoke, harass, intimidate, humiliate, mock, tease, gossip, demean, degrade, disrespect, mock, laugh at, heckle, swear or intimidate. Speech acts set conditions under which future actions are considered adequate and thus construct social reality (Scharloth, 2017). From what we can observe, conversational maxims are often disregarded, or perhaps there are different mindsets about the idea of an acceptable conversation. In this regard, the following definitions compiled by (Culpeper, 2011, p. 19-20) appear in what can be considered impolite from the linguistic pragmatics literature. We quote Culpepers schema[2] and as we can see here by the various approaches, a clear definition of what is impolite, rude, or abusive is still pending:

- *The lowest common denominator [underlying definitions of impoliteness in Bousfield and Locher 2008] can be summarized like this: Impoliteness is behaviour that is face-aggravating in a particular context. (Locher and Bousfield 2008: 3)*

- *rude behaviour does not utilise politeness strategies where they would be expected, in such a way that the utterance can only almost plausibly be interpreted as intentionally and negatively confrontational. (Lakoff 1989: 103)*

- *rudeness is defined as a face threatening act (FTA – or feature of an FTA such as intonation – which violates a socially sanctioned norm of interaction of the social context in which it occurs. (Beebe 1995:159)*

- *impoliteness, communicative strategies designed to attack face, and thereby cause social conflict and disharmony [...] (omission in the original) (Culpeper et al. 2003: 1546) Impoliteness comes about when: (1) the speaker communicates face-attack intentionally, or (2) the hearer perceives and/or constructs behaviour as intentionally face-attacking, or a combination of (1) and (2). (Culpeper 2005a: 38)*

[2]https://www.lancaster.ac.uk/fass/projects/impoliteness/definitions.htm

- *marked rudeness or rudeness proper occurs when the expression used is not conventionalised relative to the context of occurrence; following recognition of the speaker's face-threatening intention by the hearer, marked rudeness threatens the addressee's face [...] (omission in the original) impoliteness occurs when the expression used is not conventionalised relative to the context of occurrence; it threatens the addressee's face [...] (omission in the original) but no face-threatening intention is attributed to the speaker by the hearer. (Terkourafi 2008: 70)*

- *impoliteness constitutes the communication of intentionally gratuitous and conflictive verbal face-threatening acts (FTAs) which are purposefully delivered: (1) unmitigated, in contexts where mitigation is required, and/or, (2) with deliberate aggression, that is, with the face threat exacerbated, 'boosted', or maximised in some way to heighten the face damage inflicted. (Bousfield 2008: 72)*

- *verbal impoliteness [is] linguistic behaviour assessed by the hearer as threatening her or his face or social identity, and infringing the norms of appropriate behaviour that prevail in particular context and among particular interlocutors, whether intentionally or not" (Holmes et al 2008: 196)*

- *Rudeness is a kind of prototypically non-cooperative or competitive communicative behaviour which destabilises the personal relationships of the interacting individuals [...] (omission in the original) creates all maintains an emotional atmosphere of mutual reverence and antipathy, which primarily serves egocentric interests [...] (omission in the original) (Kienpointner 1997: 259; see also Kienpointner 2008)*

In addition to the linguistic perspective, there are various definitions in the social psychology literature:

- *"aggression", "social harm" or "hurt" - all of which overlap with impoliteness. Aggression may be defined as any form of behaviour directed towards the goal of harming or injuring another living being who is motivated to avoid such treatment. (Baron and Richardson 1994: 37; original emphasis)*

- *Communicative aggression is defined as any recurring set of messages that function to impair a person's enduring preferred self image [...] (omission in the original) (Dailey et al. 2007: 303; original emphasis)*

- *Social harm involves damage to the social identity of target persons and a lowering of their power or status. Social harm may be imposed by insults, reproaches, sarcasm, and various types of impolite behaviour. (Tedeschi and Felson 1994: 171)*

- *People feel hurt when they believe someone said or did something that caused them emotional pain. (Vangelisti 2007: 122)*

There are various forms of rudefulness expressed in different manners. These actions are mostly transmitted by language that can cause offence (Culpeper, 2011).

Chapter 3

Methods for Measuring Communication Behavior

We approach the online phenomenon of negative communication from multiple perspectives, trying to classify the various methods according to their applicability. Assuming that communication is characterized by ongoing change, the focus is on quantitative methods for text and communication analysis, which we use to capture structures, dynamics, and expression in language and communication. In this chapter, we discuss the various applied methods and techniques for analyzing different situations of communication (Strathern et al., 2020b, 2022b,a; Strathern and Pfeffer, 2023)

3.1 Quantitative Text Analysis

Quantitative text analysis pertains to analyzing textual information through statistical methods. In conducting quantitative text analyses, automated and systematic methods are used to process large amounts of text. Regardless of whether it is representational or instrumental, the process of conducting a quantitative text analysis always involves creating a data matrix. Words are represented in a two-dimensional matrix that is suitable for statistical analysis. By specifying the columns (or variables) and rows (or units of analysis) of a data matrix, the scope of inquiry for quantitative text analysis can be determined. This helps to identify the types of questions that can be answered through such analysis. As the analysis progresses, a theoretical map emerges, which helps to locate the research question and determine which text analysis techniques are most suitable for the study (Roberts, 2000).

3.1.1 Statistical Measurements

This chapter gives a systematic overview of how to access speech data statistically. For this purpose, we refer to Lemnitzer and Zinsmeister (2015) and give a brief summary of statistical methods and tests. Any corpus linguistic research that aims to quantitatively evaluate the results of a corpus search starts by determining the frequencies of the search results (whether they are single word types or more complex expressions).

Measuring: There are a number of different frequency measures that can be used to represent and analyze the search results in different ways. The two most commonly used measures are absolute frequency and relative frequency. Absolute frequency is the number of occurrences of a searched word or phrase in the corpus. This is the simplest frequency measure and is used when creating frequency distributions within a single corpus, but it is not useful for comparing frequencies between corpora of different sizes. Relative frequency is the normalization of the absolute frequency with respect to corpus size by dividing the absolute frequency by the total number of tokens in the corpus (N). This is written as fr = fa / N, where fa is the absolute frequency of the search expression and N is the total number of tokens in the corpus. Relative frequency is useful for comparing the frequencies of the same units (e.g., words or more complex expressions) between different corpora (or different parts of a corpus) that are not the same size, and for drawing further statistical conclusions. Proportional frequency is the normalization of the absolute frequency of a search expression with respect to the sum of the absolute frequencies of relevant related search expressions. This measure is useful for comparing the ratios of the frequencies of different variants of a lemma. In lexicography and linguistics, a lemma is the base form of a word, that is, the word form under which a term can be found in a reference work (citation form, dictionary form). between different corpora. The difference coefficient is a measure that provides information on whether the number of hits in a search in a subcorpus is as large as expected or larger or smaller than expected (with respect to the entire corpus). It "normalizes" the hit frequency to the range -1 to $+1$ by comparing the actual frequency with the expected frequency of the associated corpus subsection. A value close to 0 means that the hit occurs as statistically expected. The more the value decreases toward -1, the less frequently the hit occurs.

Measures of frequency distributions: Tables and graphs can be used to represent frequency distributions by showing all the determined frequencies. For

further quantitative evaluation, it is often useful to determine specific ratios from the total frequencies that characterize the distribution in various ways. These ratios are called statistics of the distribution. The word "statistics" has two meanings: on the one hand, it refers to a figure of a distribution, and on the other hand, it refers to the principles and procedures for the determination, presentation, and analysis of such figures. Statistics are classified based on which aspects of the shape of frequency distributions they measure, quantify, and illuminate. Two important categories of statistics are location measures and dispersion measures. Which measure to use depends on the scale level or the type of characteristic that the distribution represents (Lemnitzer and Zinsmeister, 2015).

Scale Levels: An essential part of any statistical investigation involves determining the frequencies of expressions of a characteristic (or several characteristics) of the objects of investigation (called statistical units). For quantitative analysis, it is useful to assign numerical values to expressions of a characteristic, which is referred to as measuring the characteristic (Lemnitzer and Zinsmeister, 2015). In statistics, there are four types of characteristics or variables: nominal, ordinal, interval, and ratio. Nominal characteristics have purely qualitative values, such as gender or word type. Ordinal characteristics can be ranked, but the intervals between values may not be equal. Interval characteristics have equal intervals between values, but ratios between values may not be meaningful. Ratio characteristics have equal intervals and meaningful ratios between values. For frequency distributions of nominal variables, the only measure of position that can be calculated is the modal value, which is the value that occurs most frequently. For ordinal variables, the median and the interquartile range are the best measures of position and dispersion, respectively. For ratio variables, the arithmetic mean and standard deviation are typically the best measures of central tendency and dispersion, respectively. Symmetrical frequency distributions will have similar values for the arithmetic mean, median, and modal values. Asymmetrical or skewed distributions will have diverging values for these measures, with the median generally lying in between the other two. For skewed distributions with extreme values or outliers, the median and modal values are less affected compared to the arithmetic mean and range. For more symmetrical distributions, the arithmetic mean and standard deviation are usually the best measures, they reflect both the ranking and absolute values of the variables (Lemnitzer and Zinsmeister, 2015).

- Nominal scale: Purely qualitative characteristics, where an object either

has a characteristic or does not. The corresponding mathematical property is equality: y = x or y ≠ x. Examples of nominal features in linguistics: numerus, with the characteristics (values) singular and plural; genus, with the values feminine, masculine, and neutral; word types, with the values noun, verb, adjective, preposition, etc.; text class, with values such as fiction, popular literature, science, and newspaper; corpus vocabulary, with the different word forms in the corpus as values.

- Ordinal scale: The expressions can be ranked, and the differences or intervals between the ranks need not be equal. The corresponding mathematical property is order: y ≤ x or y ≥ x. Examples of ordinal features in linguistics: grammaticality judgments, with values such as (completely) ungrammatical, very questionable, questionable, (completely) grammatical; complexity (e.g., syntactic or morphological), with values such as simple, somewhat complex, complex, very complex (or with more specific values for given syntactic or morphological constructions); publication time by decade.

- Interval scale: Equal intervals lie between expressions, but ratios between expressions cannot be formed. The corresponding mathematical property is additivity (or linearity): y = ax + b or y = ax - b. Linguistic features are usually not purely interval-scaled, but they also satisfy the definition of the ratio scale. Examples from other domains: temperature in Celsius or Fahrenheit (but not Kelvin); calendar dates according to the Gregorian or Islamic calendar; intelligence according to IQ scales.

- Ratio scale: Equal intervals lie between expressions, and ratios between expressions can also be formed because there is a natural initial quantity (a zero point) for the characteristic. The corresponding mathematical property is multiplicativity (or similarity): y = ax or y = x/a. Examples of ratio-scaled features in linguistics: length (in letters, words, phrases, etc.); distance (between words, phrases, etc.); duration of an utterance (in milliseconds, seconds, etc.).

The scales of characteristics are arranged hierarchically, with each scale having different mathematical properties. A characteristic can be downscaled to a lower scale, but not upscaled to a higher scale. For instance, a ratio-scaled characteristic like length can be described using an ordinal scale (e.g. longer or shorter) or a

nominal scale (e.g. has a certain length or not). However, it is not possible to do the opposite. For example, there is no inherent hierarchy between feminine, masculine and neuter, and no meaningful interval or ratio can be derived from these values. There are four scales that are used to classify characteristics in statistics: nominal, ordinal, interval, and ratio. Categorical characteristics are those that are either nominal or ordinal, while metric characteristics are those that are either interval or ratio. A characteristic is considered statistically interesting if it has multiple expressions that vary between individual objects (statistical units) and their frequencies vary between different groups of objects. In statistics, a characteristic is also referred to as a variable and a characteristic expression is called a variable value or simply a value (Lemnitzer and Zinsmeister, 2015).

Measures of location (measures of central tendency): A location measure is a characteristic value that is selected or calculated from observed variable values of a frequency distribution and is considered characteristic or typical of that frequency distribution. Because in many distributions this characteristic value corresponds in some way to the center of the distribution, such measures are also called measures of central tendency. The most common measures of position are the arithmetic mean, the median, and the modal value.

The modal value: The modal value, or mode, of a frequency distribution is the value that appears most frequently. If there are two or more values that occur with equal frequency and significantly more often than the other values in the distribution, the distribution is called bimodal or multimodal. A distribution in which all values occur approximately equally often has no modal value. The modal value can often be easily identified in a graphical representation of the frequency distribution, such as the longest column in a bar chart, the rightmost point in a dot plot, the largest "slice" in a pie chart, or the highest point in a line chart (Lemnitzer and Zinsmeister, 2015).

The median: If all the variable values that form a frequency distribution can be ranked, then the variable value "in the middle" of the distribution can be determined. This value is called the median: it divides the distribution into two equal groups in such a way that all values of one group do not rank higher than the median and all values of the other group do not rank lower than the median. The number of values makes a difference: If the total number of values is odd, then the median is the actual value in the middle of the ranking because there are just as many values before and after it. If the total number of values is even, then the median is the value that results when the highest value of the lower

half of the ranking and the lowest value of the upper half are added together and divided by two. The value so determined may or may not be an actual occurring variable value of the distribution (Lemnitzer and Zinsmeister, 2015).

The arithmetic mean: The median of a frequency distribution is the value that divides the distribution into two equal groups, with all values in one group ranking lower than the median and all values in the other group ranking higher. If the total number of values is odd, the median is the middle value in the ranking. If the total number of values is even, the median is the average of the highest value in the lower half of the ranking and the lowest value in the upper half. The median may or may not be an actual occurring value in the distribution (Lemnitzer and Zinsmeister, 2015).

Range: The range of a frequency distribution is the difference between the largest and the smallest value of the distribution.

Quartiles and quantiles: The first quartile, also known as the lower quartile, is the value that divides the lowest 25% of the values from the rest of the distribution. The second quartile, also known as the median, is the value that divides the distribution into two equal halves. The third quartile, also known as the upper quartile, is the value that divides the highest 25% of the values from the rest of the distribution. The quartiles can be used to summarize the spread of a distribution and to identify outliers. The quartiles of a frequency distribution are values that divide the distribution into four equal sections. The first quartile, also called the 25% quartile, separates the lowest quarter of values from the three higher quarters. The second quartile, also called the 50% quartile, divides the values in half and is the same as the median. The third quartile, also called the 75% quartile, separates the highest quarter of values from the three lower quarters. The interquartile range is the difference between the 75% and 25% quartiles. Other divisions, such as deciles and percentiles, may also be used. The quantiles of a distribution include the smallest value, called the 0% quantile, and the largest value, called the 100% quantile, which gives the range of the distribution (Lemnitzer and Zinsmeister, 2015).

The standard deviation: The standard deviation is a measure of the dispersion of a frequency distribution, calculated by taking into account the deviation of each value from the arithmetic mean of the distribution. It is commonly used for ratio variables and takes into account all of the values that make up the frequency distribution. The standard deviation provides a way to understand the average deviation of the individual values from the mean of the distribution.

Deviation in this context refers to the arithmetic difference between values. To determine the central tendency or dispersion of a given frequency distribution, the choice of measure depends on both the scale level of the variable of the distribution and the shape of the distribution (Lemnitzer and Zinsmeister, 2015).

Frequency measures for corpora: Absolute frequency is the number of occurrences of a particular word or expression in a corpus. It is a simple measure that is used to create frequency distributions within a single corpus, but it is not useful for comparing frequencies between corpora of different sizes. Relative frequency is the normalization of the absolute frequency with respect to corpus size. It is calculated by dividing the absolute frequency of a word or expression by the total number of tokens in the corpus. This measure is useful for comparing the frequencies of the same units between different corpora or different parts of a corpus that are not the same size. It can also be used to draw statistical conclusions. Proportional frequency is the normalization of the absolute frequency of a word or expression with respect to the sum of the absolute frequencies of related words or expressions. It is used to compare the ratios of different variants of a word between different corpora. The difference coefficient is a measure that provides information on whether the number of occurrences of a word or expression in a subcorpus is as expected or higher or lower than expected. It is calculated by comparing the actual frequency of the word or expression to the expected frequency in the associated corpus section and normalizing the result to a range from -1 to 1. A value close to 0 indicates that the frequency is as expected, while a value approaching -1 or 1 indicates that the frequency is lower or higher than expected, respectively (Lemnitzer and Zinsmeister, 2015).

Statistics and probability: Statistical investigations typically take place in four phases: In the first phase, the research project, e.g., a particular empirical question, is operationalized, i.e., put into a quantifiable form, which thus enables the determination and ascertainment of frequencies of the relevant characteristics (i.e., the variables of the investigation). In the second phase, data are collected, e.g., in corpus linguistics by means of search queries in a corpus or in different corpora (or sub-corpora), and from this a data set is formed, which contains the statistical units and their feature expressions (e.g., the hits of the search query as well as their metadata and possibly other features). In the third phrase, we quantitatively describe the data in terms of the features of interest by, among other things, calculating various frequency measures and forming frequency distributions, tabulating and graphing them, and calculating

their location and dispersion measures. Finally, in the fourth phrase, we attempt to draw more general conclusions from the quantitative evaluation of the data at hand; in other words, in corpus linguistics, to make statements that apply not only to the corpora used but also to the language domain as a whole or even to the language itself from which the corpus data originate. The procedures of the third phase belong to the so-called descriptive statistics (also called descriptive or empirical statistics). They allow comparisons between different data sources, in corpus linguistics, for example, between different corpora or subcorpora. The methods of the fourth phase, which allow conclusions and generalizations to be drawn from the data studied, belong to what is known as inferential statistics (also called inferential, inductive, or analytical statistics). These methods are essentially based on mathematical probability theory. In order to use such methods of inferential statistics sensibly and correctly, we do not extend on the mathematical details of probability theory, but it is useful and helpful to know the most important basic concepts and features of this theory, so they are presented below (Lemnitzer and Zinsmeister, 2015).

Population and Sample: First of all, it is important to distinguish between the actual data under study and the existing but unstudied data on which one wants to draw conclusions. In a statistical study, the set of all units that can have the expressions of the studied characteristics (i.e., the values of the variables) is called the population of the study (also called the population). The set of statistical units from a population that is actually considered in detail is called a sample. In this sense, a corpus is a sample from the population of a particular language or language domain. Inferential statistics thus provides methods and procedures by which one can draw conclusions about the corresponding population from the results of a sample (or several samples). The ratios of a population that corresponds to the statistics of a sample are called parameters of the population. Thus, inferential statistics are used to draw conclusions from the values of the determined statistics of the samples to the values of the corresponding parameters of the population (Lemnitzer and Zinsmeister, 2015).

Sampling Distribution: To determine the statistic for all possible samples from the population, we have a distribution of all possible values of the parameter of the population, the so-called sampling distribution of the statistic. One of the values of a sampling distribution is the actual value of the parameter of the population. Not all the values of the population are known, just the values of the sample. The best we can do is to determine how likely it is that a

given value from the sampling distribution is or is not the value of the parameter. Using probability theory, we can assign a probability to each value in the sampling distribution that it is the value of the parameter in the population. Thus, the sampling distribution is also a probability distribution of the parameter. For realistic (especially linguistic) populations it is impossible to form an actual sampling distribution because we cannot collect all possible samples. But experience from many observations and experiments shows that for many natural (including linguistic) phenomena, the probability distribution of the possible expressions of a feature of interest (i.e., the possible values of a variable) has a shape that can be approximately characterized by a (more or less complicated) mathematical formula. In such cases, there is no need to actually form a sampling distribution; instead, we use the formula as a model of the distribution and draw inferences about the population based on this model (i.e., assuming that the model represents the relevant parameters of the population with sufficient precision) (Lemnitzer and Zinsmeister, 2015).

Random experiment and random variable: In order for the conclusions about the population drawn from a sample to be as reliable and convincing as possible, the sample should satisfy certain properties. In probability theory, the model of such a sample is called a random experiment. The repeated execution of a random experiment yields a probability distribution (where one usually replaces the actual execution of random experiments by mathematical - and usually computer-implemented - simulations) (Lemnitzer and Zinsmeister, 2015).

The Binomial Distribution: Binomial distributions are discrete probability distributions; they result from random experiments whereas the normal distribution is a continuous version of symmetric binomial distributions. Its development goes back to the attempt to find a computationally manageable estimation of the binomial coefficients because the factorials occurring in it are very computationally intensive (even for modern computers) except for small numbers. Many naturally occurring phenomena have approximately normal distributions (e.g., the distribution of human body size). In addition, there is a theorem of probability theory, the central limit theorem, which shows that the distributions of the means of samples approach a normal distribution as the size (or the number of samples) increases, even if the distributions of the samples themselves are not normally distributed (Lemnitzer and Zinsmeister, 2015).

The Chi-Square Distribution: Another important family of continuous probability distributions is the chi-square distribution, which is based on the

standard normal distribution. The chi-squared distribution (named after the Greek letter χ and often written as χ^2 -distribution; the corresponding random variable is usually written as X2) is defined as a sum of independent squared standard normal distributed random variables $Z_i \sim \mathcal{N}(0, 1)$:

Confidence intervals: A statistic from a random sample represents an estimate of the unknown value of the corresponding parameter of the population. We construct the sampling distribution of the statistic and determine the probability that the value of the statistic will agree with that of the parameter. However, using the sampling distribution, we can also calculate an interval of values around the statistic that encloses the parameter value with a given probability. The interval is called a confidence interval and the given probability is called the confidence level. For example, a confidence level of 95% means that in 95% of all samples, the parameter value of the population lies within the confidence interval; this is then called the 95% confidence interval. Since the probability of agreement between the value of the point estimator and the parameter value is always relatively small, a confidence interval usually gives a more accurate idea of the probable parameter value. The probability expressed by the confidence level refers only to the totality of the samples. The larger the confidence level, the more values of the sampling distribution it contains. Consequently, with a confidence level of 100%, we would have absolute certainty that the value of the population lies within the confidence interval - but only because the confidence interval then contains all possible values of the random variable. But with this, we would not know more than before the sample. 95% is considered a good compromise between certainty and accuracy (Lemnitzer and Zinsmeister, 2015).

Hypothesis Testing: The determination and measurement of statistics, the construction of sampling distributions, and confidence intervals are based on research projects of an empirical and theoretical nature. The analysis of samples serves the purpose of drawing conclusions about the population from which the samples originate. One of the most common methods for this is hypothesis testing. This is used to determine whether the statistics obtained are significant in a statistical sense, hence they are also called significance tests. A working hypothesis is formulated about one or more characteristics of a population. Usually, the working hypothesis asserts the occurrence of a certain effect, which would result in a change in the current state of knowledge. The working hypothesis is tested using a sample statistic (or several sample statistics). For a statistical study, we quantitatively evaluate expressions of features, which requires the quantitative

operationalization of a qualitative hypothesis. The working hypothesis is called the alternative hypothesis. Usually, the notation H0 is used for the null hypothesis and H1 for the alternative hypothesis. One takes a random sample from the population, determines the relevant statistic and the corresponding sampling distribution according to the null hypothesis. In this context, the statistic is called the test statistic. Using the sampling distribution, one calculates the probability of the value of the test statistic observed in the sample, as well as the probabilities of all values of the distribution that are even more extreme, i.e., even further away from the expected value of the test statistic according to the null hypothesis, than the observed value. The sum of all these probabilities is called the p-value of the hypothesis test: it represents the risk of falsely rejecting the null hypothesis (Lemnitzer and Zinsmeister, 2015). **Statistical Testing:** Before selecting the appropriate test for mean differences, the hypothesis being tested should be clear and unambiguous. The chosen hypothesis determines which test procedure is suitable. Hypotheses can be either non-directional or directional. The following decision criteria are relevant for selecting the appropriate analysis procedure: Are the measured samples independent or related? How many variables were examined? Are the variables being examined normally distributed? Samples are independent when the means of two independent groups are compared. The tests are differentiated based on the number of variables and the necessity of a normal distribution of the dependent variables. The one-sample t-test tests whether the mean of a sample differs from a given value based on the mean of a population. The one-sample t-test is a (para-)metric test procedure. Prerequisite is that the dependent variable is normally distributed and has at least one metric scale level. The test compare the means of a measured characteristic (dependent variable; such as learning performance or school grades) of two independent samples (independent variable with 2 levels; for example, school class A and B, women and men). The independent samples t-test tests for mean differences of a normally distributed, at least interval-scaled variable between two independent samples. The independent samples t-test is an extension of the one-sample t-test and also a (para-)metric test procedure. Prerequisite is that the dependent variable is normally distributed and has at least one metric scale level. The independent samples come from populations with approximately identical variances of the dependent variable, meaning that variance homogeneity should be checked (Lemnitzer and Zinsmeister, 2015). The following tests compare the means of a measured characteristic (dependent variable; e.g. academic perfor-

mance, grades) between two dependent samples (independent variable with 2 levels; e.g. class A is measured before and after a teaching unit). A paired sample t-test (also: t-test for dependent samples) examines mean differences of a normally distributed, at least interval-scaled variable between two dependent samples. The paired sample t-test is a (para-)metric test procedure. Prerequisite is that the dependent variable is normally distributed and has at least a metric scale level (Kühnel and Krebs, 2014). The WILCOXON Test examines differences in means of at least an ordinal scaled variable between two dependent samples. Normal distribution is not a requirement. The WILCOXON Test is a non-parametric testing method. It requires that the dependent variable is at least ordinal scaled. Distributions can also differ from each other in terms of their variability. The decision criterion for choosing the appropriate test procedure is also the scale level of the dependent variable. The Chi² test can be used to examine the variance of dependent variables of any scale level between a sample and a population. The Chi² test only tests undirected hypotheses. The Chi² test is a non-parametric test procedure. Prerequisite is that the dependent variable is normally distributed. The variance of the population must be known (Kühnel and Krebs, 2014). In statistics, in addition to differences, relationships between two or more variables are incrementelly analyzed (e.g. a relationship between gender and pay in a study on gender-sensitive pay). The aim is to determine whether the measured variables are related, how strong this possible relationship is, and, if applicable, what direction the relationship has (positive or negative): Positive relationship: The more – the more (and vice versa). Negative relationship: The more – the less (and vice versa). Relationships between metric variables can be visualized graphically in a scatterplot, in which individual values are represented as points. A relationship can be suspected when the points approximate a diagonal line, such as in a linear regression. If no specific point pattern is discernible, it is likely that there is no relationship. A scatterplot does not replace statistical analysis on the existence of a relationship, its size, and direction. When speaking of correlation in general, we often mean the Pearson correlation coefficient r (also known as product-moment correlation or Pearson correlation). The correlation coefficient is the result of standardizing covariance (cov) and is interpretable, unlike covariance (Kühnel and Krebs, 2014). The Pearson correlation coefficient can be calculated when a possible linear relationship between two at least interval-scaled variables is undirected. Undirected means that the measured variables vary together, but it is unclear

whether one variable causes the other or not. For this reasin, one cannot speak of a dependent and an independent variable. Causal statements are also not possible. The correlation coefficient can take values between −1 (perfect negative relationship) and +1 (perfect positive relationship). If the value is close to 0, there is no relationship. To estimate the size of a relationship, one can use Cohen's classification: weak from 0.1, moderate from 0.3, and strong from 0.5. Requirements are that the variables are at least interval-scaled, and the variables are normally distributed. The suspected relationship between the variables is linear. The correlation coefficient can be obtained by squaring the correlation coefficient. The coefficient of determination can take values between 0 and +1. It is traditionally multiplied by 100 and expressed as a percentage. It indicates the proportion of variance in both variables that is explained by common sources of variance (Kühnel and Krebs, 2014). Relationships between at least ordinal scaled variables can be determined using the rank correlation coefficient rs by Spearman (also known as Spearman correlation or Spearman's rho). In addition to linear, non-linear undirected relationships can also be examined. An undirected relationship exists when the measured variables vary with each other, but it is unclear whether one variable causes the other or not. It is therefore not possible to speak of a dependent and an independent variable. Causal statements are also not possible. The rank correlation by Spearman is the non-parametric alternative when the data to be examined do not meet the requirements for a correlation according to Pearson. The rank correlation coefficient can take values between -1 (perfect negative correlation) and +1 (perfect positive correlation). If the value is close to 0, there is no correlation. To estimate the size of a correlation, one can use the classification by Cohen: from 0.1 as weak, from 0.3 as moderate and from 0.5 as strong. Requirements are that the variables are at least ordinal scaled. The correlation coefficient can be obtained from the rank correlation coefficient by squaring it. The correlation coefficient is traditionally multiplied by 100 and expressed as a percentage. It indicates what proportion of variance in both variables is explained by common sources of variance (Kühnel and Krebs, 2014).

3.1.2 Sentiment Analysis

Sentiment analysis, also referred to as opinion mining or emotion analysis, uses natural language processing, computational linguistics, and text analysis to systematically detect, extract, measure, and analyze emotional states and personal

information (Hamborg and Donnay, 2021). This technique is widely implemented in analyzing the feedback of customer on product reviews. Sentiment analysis plays a crucial role in categorizing the polarity of a given text by determining whether the opinion expressed in the text is positive, negative, or neutral. This analysis can be performed at the document, sentence, or feature/aspect levels. There are two approaches to use a neutral class: one is to have the algorithm identify the neutral language first, remove it, and then evaluate the remaining text for positive and negative sentiments; the other is to create a three-way classification in one step, typically by estimating a probability distribution across all categories (Taboada et al., 2011). Including a neutral class depends on the data's characteristics. If the language in the data falls into the categories of neutral, negative, and positive, it is recommended to remove the neutral part and focus on the polarity between positive and negative sentiments. On the other hand, if the data mostly consist of neutral language and only have minor occurrences of positive or negative expressions, using a neutral class may not be the most effective approach. Another approach is to use a scaling system, where words are assigned numerical values within a specific range, such as -10 to $+10$ or 0 to a maximum positive value like $+4$, based on their negative, neutral, or positive sentiment (Taboada et al., 2011; Mehmood and Balakrishnan, 2020). This enables sentiment adjustment of a specific term in relation to its context, often at the sentence level. Natural language processing is used to score each concept in the specified context based on its association with sentiment words and their corresponding scores. This approach allows for a more nuanced understanding of sentiment, as the value of a concept can be adjusted based on the modifying words that surround it, such as intensifiers, relaxers, or negators. In cases where the objective is to assess the sentiment of a text, as opposed to its polarity and intensity, positive and negative sentiment strength scores can be allocated to the text (Thelwall et al., 2010).

3.1.2.1 Classification with Psycho-Linguistic Dictionary

Rather than count words that occur, dictionary-based approaches pre-define words associated with specific meanings. This involves no human decision-making as part of the text analysis procedure. A psycho-linguistic dictionary is a type of dictionary that relates language and psychology, specifically the psychological processes involved in language use and language learning (Lehmann et al., 2017). Psycho-linguistic dictionaries include information about the psychological

factors that influence language acquisition, processing, use, and change. They contain information about how language and psychological processes interact with each other, as well as how language can be used to manipulate psychological states or processes. Sentiment analysis analyzes text data to determine the sentiment or emotion expressed in the text. According to psychologist James W. Pennebaker, the use of psycholinguistic dictionaries is important for sentiment analysis because it allows for a more accurate and nuanced analysis of the text (Pennebaker et al., 2015). Pennebaker argues that words carry a lot of information about a person's thoughts and emotions and that the use of psycholinguistic dictionaries can provide insight into the underlying psychological processes that are reflected in the language used. Psycholinguistic dictionaries are designed to capture the specific psychological and emotional meanings of words and phrases. By using these, the sentiment analysis can be more fine-grained and can detect subtle variations in emotions and sentiment that may not be captured by more traditional sentiment analysis methods. Additionally, psycholinguistic dictionaries can help to overcome some of the limitations of traditional sentiment analysis methods, such as the use of pre-determined sentiment categories (yes/no), which may not fully capture the complexity of human emotion (Lehmann et al., 2017). Basically, the methodological approach to quantifying emotional words with the help of emotion lexicons is similar to sentiment analysis: texts are matched with emotion dictionaries; the latter consists of lists of words as well as additional columns in which a word is assigned to an emotion category or the values for valence and arousal are recorded. Emotion analysis is currently carried out as quantification of words associated with emotions. While highlighting those words in the text that are recognized as emotional words also allows for a close reading, longer texts are evaluated with regard to quantities and trends. Usually, emotion lexicons list the words in their basic form; English emotion lexicons only list them in lowercase. Before matching with an emotion lexicon, the source texts must usually be lemmatized and provided with lowercase letters in order to find as many matches as possible in the emotion lexicons. Using lexicons does not enable consideration of the context of emotional words within the sentence and thus, for example, identification of the speaker of emotional words (Lehmann et al., 2017). The classification of emotional words does not deal with the objects to which the emotions refer, and more complex structures such as figurative language or indicators of irrealis (moods, conditional constructions, intentional verbs) are not taken into account. The orientation/polarity of emotion

(positive/negative) can be context-dependent (domain-specific), which is also not taken into account (Lehmann et al., 2017). But methodologically, applying dictionaries that also count for word categories offers the chance to discover linguistic peculiarities. The quantification of written text with categories that reflect content with numbers makes use of an annotation process and enables a quantitative analysis. The LIWC Tool can analyze individual text files, groups of files, or texts within a spreadsheet, one at a time. The program reads every word in the text, searches its dictionary for a match, and then increments the appropriate category scale if a match is found. While processing the text, LIWC2015 also tracks various structural elements, such as word count and sentence punctuation. The tool records around 90 output variables for each text file and saves them in an output file. These variables include file name, word count, information on language, descriptors, linguistic dimensions, psychological constructs, personal concerns, informal language markers, and punctuation categories. The dictionary is the central component of the text analysis process and consists of nearly 6,400 words, word stems, and emoticons (Pennebaker et al., 2015; Tausczik and Pennebaker, 2009). Each entry in the dictionary also defines one or more word categories or subcategories. For example, the word "cried" is associated with five different categories: sadness, negative emotion, overall affect, verbs, and past focus. If the word "cried" is found in the text, the scores for these five subcategories will be incremented. Many categories are organized hierarchically, with certain words belonging to broader categories. Additionally, the system can also recognize word stems, like the stem "hungr*" which captures words like "hungry" or "hungrier". To calculate statistics in the LIWC2015 dictionary, each word in a category list is measured as a percentage of the total words in the text. These scores are then used as an "item" in a standard Cronbach's alpha calculation, which provides raw alpha scores for each word category separately for each corpus. The uncorrected alphas are calculated as an average of each corpus's alpha score, but this method can underestimate reliability due to the variable usage rates of words within a category. To address this issue, corrected alphas are calculated using the Spearman-Brown prediction formula, considered to be a more accurate approximation of a category's proper internal consistency (Pennebaker et al., 2015).

3.1.2.2 Feature Extraction and Selection

Sentiment analysis, and dictionary-based approaches are methods from corpus linguistics. Before we dive deeper into the single features, we focus on some background considerations when working with large-scale speech data and corpus data. Scharloth (2017) argues that current methods in natural language processing and corpus linguistics cannot simply supplement or replace traditional, mainly qualitative cultural analysis methods, but must instead develop new categories and strategies for understanding the social and cultural meanings of language. According to his considerations, one such category is found in linguistic pragmatics, specifically, in the theory of idiomatic imprinting, which suggests that pragmatic information is expressed not just through speech acts, but also through the "use value" of linguistic structures. This use value is evident in linguistic patterns on the surface of language, and frequently occurring patterns may be the result of recurrent language use that reflects typical contexts, goals, and frameworks for interpretation. Instead, recent pragmatics suggests that language use helps create the context ("contextualizes"), and that routine language formulas can be understood as contextualization clues (Scharloth, 2017). Idiomatic imprinting can be seen as the result of conventionalized interpretations that are reflected in language. Frequently occurring language patterns can be understood as the result of repeated language use by speakers, which includes typical contexts, goals, and frameworks for interpretation. Corpus linguistic techniques are very effective for identifying patterns in language that occur repeatedly. These patterns can be quantified on the basis of the frequency of specific linguistic elements in a corpus, and their recurrence can be identified through induction (Scharloth, 2017). In mainstream linguistics, corpus linguistics is used to identify patterns in language use and interpret them as regularities or norms. In branches of linguistics focused on cultural and social issues, recurrent linguistic patterns are often linked to cultural or social phenomena. Depending on the linguistic theoretical perspective, these patterns may be seen as symptoms of these phenomena or as contributing to their formation. Scharloth (2017) makes the concept of "collocation" central of linguistics. The term refers to the way certain words regularly occur together, even though the combination of those words may not follow usual grammatical rules or be easily explained by the meanings of the individual words themselves. Collocations must be learned through exposure to and experience with the language. He states that collocations are understood as frequent concurrences of linguistic units that can be easily identified through

statistical analysis. This knowledge of how words tend to occur together in language use is inscribed in collocations and it has been further developed in the field of phraseology and in computational linguistics. The category of collocation is just one example of a series of analytical categories such as "keyword" or "n-gram" that have become standard analysis categories in linguistics through the use of corpus linguistic methods (Scharloth, 2017). Researching corpus pragmatics, therefore, means inductively searching for significantly frequent patterns in large text corpora and interpreting these patterns as expressions of recurrent language acts, in other words as patterns with sociocultural salience (Lemnitzer and Zinsmeister, 2015). According to the authors, quantitative evaluations involve determining frequencies in the corpus and the possible comparison of results. In a purely quantitative approach, such data are extracted and evaluated from raw, i.e. not linguistically annotated, corpora using statistical methods. Qualitative evaluations are concerned with the identification, classification, categorization, and interpretation of particular phenomena. Accordingly, corpus data in linguistics is considered an additional source of evidence. Research into these phenomena involves a targeted search in the corpora for relevant (mostly syntactic) constructions in order to confirm or refute predictions of a theory (Lemnitzer and Zinsmeister, 2015). In this context, the linguistic preparation of the corpus can play a crucial role. Quantitative-qualitative evaluations combine methods from both approaches: statistical procedures are applied but the data derived from corpora do not remain uninterpreted; e.g., annotations such as part of speech, and syntactic function can be taken into account (Lemnitzer and Zinsmeister, 2015). Bubenhofer and Scharloth (2016) claim that there are limitations to this data-driven paradigm. Although it refrains from formulating hypotheses and from specifying certain analysis categories, it is obvious that prior knowledge also flows into the research process in data-driven procedures. They name the following issues:

- through the choice of corpora,
- in terms of the design of algorithms for pattern calculation,
- in determining what should be considered as a linguistic unit of investigation (token), and
- in determining which unit types should actually be considered as potential components of a pattern.

- Finally, the categorization of data following pattern calculation is also an interpretive process that can be partly objectified through statistical methods; however, the amount of data is often so extensive that further reduction and weighting in the sense of the research interest is necessary (Bubenhofer and Scharloth, 2016).

Often, however, both approaches complement each other: the first access to the data is data-driven to identify linguistic peculiarities that are then used for the operationalization of central concepts. The actual analysis is then data-based using the data-driven measurement instrument. Data from Twitter is semi-structured, which requires structuring in order to conduct studies and analyses. One approach to achieving this is to add information to the text, resulting in a structured text corpus that can be readily analyzed. According to Adamzik (2016) the structure of a corpus consists of speech data, annotation, and metadata. A preprocessed corpus has three layers: the language data, the analytical annotations, and the descriptive metadata. The core of a corpus is the language data, which consists of texts, speech recordings, or their written transcriptions, and is stored in digital form. They can be based on linguistic primary data, such as a sound recording or text publication, that exists independently of the corpus. Depending on the type of primary data, one distinguishes between text corpora and spoken language corpora. Textual primary data may already exist in digital form, or it may only be available as printed texts, handwriting, or the like. If the primary data exist as concrete publications, they have words and an external form: the distribution of the text on one or more pages, the size, the color, the font of the letters (Adamzik, 2016).

Speech Data: At the second level, the language data can be analyzed at various annotation levels. The first analysis level consists of segmentation, the decomposition of the speech signal or the string of characters into linguistically defined units such as phonemes, words, or sentences. In text corpora, the segmentation can also delimit text structural units such as paragraphs, chapters, headings, or footnotes. Segmentation can be indirectly encoded, for example, by the convention that token boundaries are marked by spaces, sentence boundaries by line breaks, and paragraph boundaries by blank lines. However, an annotation that explicitly separates the language data from the analysis by naming the text structural units and language data with annotation labels is preferable (Adamzik, 2016).

Annotation: On the basis of the segmentation, there may be further lin-

guistic and non-linguistic annotation levels. Corpora often contain annotations are added for part of speech and lemma. Syntactic annotations are available for constituent structure and grammatical functions and dependencies. Semantic annotations include word senses, semantic roles, semantic frames, tense and aspect. Discourse-related annotations include coreference phenomena, information status, information structure, discourse relations, and dialog acts. Beyond the purely linguistic analysis, annotations of emotions and opinions, as well as the analysis of facial expressions and gestures, are included. To achieve consistent annotation and for any later use, it is very important that the annotations are thoroughly documented. The meaning of annotation labels (tags) is clearly defined in a tagset and the annotation criteria are demonstrated with examples in guidelines. To document the quality of annotations, the agreement among annotators is recorded (inter-annotator agreement) (Adamzik, 2016).

Metadata: Metadata is also referred to as data about data. It describes primary data, the language data contained in the corpus, and the annotations. For example, it captures the text genres the data belongs to, the size of the data set, and how the language data is encoded. Furthermore, contextual aspects relating to the creation of the corpus are documented, such as the time and place of primary data creation and publication, involved parties, the time of annotation creation, and the names of annotators. In addition, there are references to external sources such as the definitions of the annotation labels (tagset), annotation guidelines, and publications that describe the corpus. Indications of the corpus and its primary data's copyright properties are also significant information. In addition to information about the data and annotations, information about the metadata itself is also given, such as whether the metadata was created manually or automatically and if it follows a certain standard (Adamzik, 2016).

Pronouns: Specifically, pronouns are a class of words in grammar that, according to the literal meaning of the term, "stand in place of a noun (name word)". Examples include "he" (a personal pronoun), "my" (a possessive pronoun), or "which" (an interrogative or relative pronoun). However, a pronoun cannot always be replaced by a noun in the same position in a sentence (e.g. only a relative pronoun can introduce a relative clause, not a noun) (Bußmann, 2008). Nonetheless, a pronoun establishes a reference to an individual, just as it can be done alternatively with a noun and article. Therefore, pronouns can have the same grammatical features as articles and nouns in German: gender, number, and case. In contrast to nouns, pronouns are not content words. Rather, they

designate people or things only through their grammatical features. These features are then used to refer to the context of utterance (deictic, as in the first and second person of personal and possessive pronouns, and in other ways, demonstrative pronouns), or they refer to the linguistic context (anaphoric, usually in the third person of personal and possessive pronouns, as well as reflexive and relative pronouns) (Bußmann, 2008). Additionally, they can act as placeholders for individuals newly introduced in the text (as with indefinite and interrogative pronouns). In traditional linguistics, expressions that stand alone without a noun (e.g. "I", "you", "this" in "but this one said") and those that precede a noun (e.g. "his" in "his house", "this" in "this man") are both considered pronouns. While pronouns have been traditionally considered a part of speech, some contemporary theorists argue that they do not constitute a single class due to the range of functions they can serve across different languages (Alexiadou et al., 2008).

Netspeak: In his book Language and the Internet, Linguist David Crystal describes the phenomena of Netspeak (Crystal, 2001) as slang, an informal mode of communication due to its ability to incorporate aspects of both oral and written communication. We summarize his findings on Netspeak's structure and functional characteristics, specifically the language in chat groups. According to David Crystal, there are five areas of the Internet where a specific slang is used: the Web itself, email, asynchronous chat (such as mailing lists), synchronous chat (such as Internet Relay Chat), and virtual worlds. As Crystal states, the Internet is a worldwide, interactive electronic platform that significantly impacts the language used on it. Its electronic nature profoundly affects the communication options available to users. The kind of hardware required to access the Internet determines what can be communicated, while the size and shape of the screen determine what can be seen. The Internet software and hardware properties also limit the language the sender and receiver use. Therefore, certain traditional linguistic practices are well-suited to this medium, while others are impossible. It is crucial to understand the limitations and advantages that come with using the Internet. A well-known communication principle suggests that users must be aware of the capabilities and constraints of their chosen medium based on their intended purposes and usage. However, this relationship is only sometimes straightforward. The development of Netspeak is an example of the tension that "exists between the nature of the medium and the aims and expectations of its users" (Crystal, 2001). The central issue appears to be how the Internet

language relates to spoken and written communication forms. Some scholars have referred to Internet language as "written speech," and people "write how people talk"(Crystal, 2001). The electronic discourse often resembles spoken language, as if the sender were writing a conversation. However, to what extent is it feasible to "write speech" on a keyboard that only allows letters, numbers, and a few other symbols and in a medium lacking some essential conversational speech? The language of geeks has significantly influenced Netspeak in the past, using jargon that caters to a younger, tech-savvy audience. However, as the user base expands to include people with varying language preferences, Netspeak is evolving. What are the characteristics of verbal and written communication and the various elements that set them apart? These components have been heavily studied in the field of linguistics. Time constraints commonly limit verbal expression; it is often impromptu, conducted in person, involves social interaction, has a flexible framework, allows for immediate modifications, and includes prosody, stress, and rhythm variations. In contrast, written communication is commonly restricted by space, premeditated, lacks social cues, primarily communicates information, has a complex structure, can be revised multiple times, and employs visual elements. The internet represents one end of this spectrum, where many features are similar to traditional writing-based situations. Most forms of written language can be found on the internet with only minor alterations in style to suit the electronic medium. Various types of text, including legal, religious, literary, scientific, and journalistic writing, are all available on the internet, just as they are in their non-digital form. Visual and graphic elements need to be considered to identify the distinctive stylistic features of web pages. In this sense, web pages embody the general characteristics of writing. For instance, web page authors typically need to gain knowledge of their readership and engage in traditional authors' and organizations' exact targeting and feedback-seeking behaviors. However, certain web functions, such as e-sales, bring it closer to the type of interaction commonly associated with speech, which has implications for the language used (Crystal, 2001). As a result, interactive features like email, microblogs, and chats are commonly integrated into websites. In contrast to the web, communication modes such as email, chat groups, and virtual worlds share similarities with oral communication despite being reliant on written expression. These methods of communication are time-sensitive, require prompt responses, and are ephemeral, as messages can be deleted or lost. The language used in these settings is often similar to face-to-face conversations. While chat groups

and virtual worlds are primarily intended for informal conversation and are more "spoken" in nature, people still "write" emails to communicate. There are two primary reasons why Internet messages may not conform to typical speech patterns. To begin with, the technology used in this medium needs immediate feedback. Messages sent through a computer are one-way, complete, and final. Unlike traditional teleprinters, the recipient does not receive the message keystroke-by-keystroke but rather as a whole message. Secondly, this medium has unique characteristics, such as asynchronous communication, distinguishing it from face-to-face conversation and may impact the language style used. Once a message is typed and sent, it is transmitted as a whole and arrives on the recipient's screen simultaneously due to the one-way nature of the medium. As a result, recipients cannot respond to the message while it is being typed since they do not know when it will appear on their screen (Crystal, 2001). As a consequence, participants cannot assess the effectiveness of their message in real-time, such as whether it has been understood or requires clarification. The medium does not allow for the receiver to provide the simultaneous audio-visual reactions, that are critical in face-to-face interaction. Additionally, messages cannot overlap and there is a waiting period before the text appears on the recipient's screen, which differs greatly from the complex realities of everyday conversation. This lack of simultaneous feedback is the first major difference between Netspeak and face-to-face conversation (Crystal, 2001). Crystal points out that the second significant difference between the internet and face-to-face interaction is the slower pace of online interactions caused by the technology itself. This slower pace prevents some of the most significant aspects of conversation that are usually present in face-to-face communication. For instance, e-mails and asynchronous chat groups may take a long time to receive a response, ranging from a few seconds to several months. The exchange speed depends on various factors, such as the recipient's computer setup, the sender's communication habits, and the situation of the participants, including their access to computers. This time lag is a critical factor in many situations, as there is always uncertainty about how long it will take to receive a response after sending a message. Due to a delay in communication, the tempo and regularity of an exchange, even in the swiftest online conversations like real-time chats and virtual environments, are not as quick and foreseeable as those observed in phone calls or in-person dialogues. Every type of delay can create issues, but some are more problematic than others. A short delay usually lasts around 2-3 seconds, which most participants can

handle, but some people may still find it challenging since it is longer than the average duration in most conversations. This affects both parties involved in the communication. From the sender's perspective, the ideal moment to speak may be missed since the related topic may have already scrolled off the screen and quickly disappeared from the group's collective memory. Looking at it from the recipient's viewpoint, the absence of an expected response is uncertain, as it is hard to tell if the delay is caused by transmission issues or some "attitude" of the sender (Crystal, 2001). The language techniques used in chat groups are less reliable than face-to-face interactions. When the number of participants increases, the situation becomes more unpredictable. Delays between two individuals can be inconvenient and unclear, but they can typically be managed since each person only speaks with one person. In contrast, electronic exchanges, such as email, may experience significant delays in response. In interactions with multiple participants, such as chat groups, virtual worlds, and forwarded emails, these delays create a distinct situation that disrupts a crucial aspect of face-to-face communication, conversational turn-taking (Crystal, 2001). Taking turns speaking is a fundamental aspect of the conversation and essential for successful interactions. Individuals naturally follow this practice to prevent talking over each other randomly or excessively. Furthermore, specific expected adjacency pairs occur, such as questions followed by answers or complaints followed by excuses or apologies. These basic strategies form the framework of a typical conversation. When orderly turn-taking is disrupted, and adjacency pairs are frequently interrupted, it can lead to confusion. As the number of participants increases, the potential for overlapping interactions grows, and keeping track of a topic or thread on a screen becomes increasingly tricky (Crystal, 2001). One of the primary characteristics crucial for communication is the realm of prosody and paralanguage. These linguistic concepts encapsulate how one expresses oneself vocally, such as through "variations in pitch, loudness, speed, rhythm, pause, and tone of voice" (Crystal, 2001), which is just as important as the words themselves. Virtual worlds allow individuals to express their emotions through text, often accompanied by synthesized sounds and visual effects. This relates to how Netspeak "lacks the facial expressions, gestures, and conventions of body posture and distance" (Crystal, 2001), crucial for expressing personal opinions and attitudes and regulating social interactions. This limitation was recognized early on with the use of Netspeak, creating emoticons (Crystal, 2001). In the early days of Netspeak, emoticons were created to fill the gap left by the absence of facial

expressions, body postures, and gestures in virtual communication. These are created by combining keyboard characters to form emotional facial expressions. They are usually typed in a single line and inserted at the end of a sentence after the final punctuation mark. Emoticons, which are used in virtual communication, are divided into two main categories. The first category represents positive feelings, while the second category represents negative feelings. Positive emoticons are formed by typing either :-) or :), while negative emoticons are represented by :-(or :(. Usage guides often warn against the ambiguity of humor and irony in virtual communication due to the absence of the prosody of speech (Crystal, 2001). Using a smiley can prevent a significant misunderstanding of the speaker's meaning. However, even with a smiley, it is still possible to interpret it in various ways, such as happiness, humor, and sympathy. The only way to clarify the meaning of a smiley is to refer to the surrounding words. Some observers have even criticized smileys as being "useless". Smileys, despite their restricted utility, are a distinctive characteristic of the language used in e-mails and chat groups. In addition to the absence of traditional conversational features, internet communication also lacks other characteristics of spoken language, making it more difficult to use language online in a truly conversational manner. These limitations arise from the need for the medium to depend on typing speed and proficiency (Crystal, 2001). The vocabulary of the internet is a highly creative and constantly evolving field in modern English, incorporating all significant word methods formation. To create new words on the Internet, a common technique is to combine two distinct words to form a compound. Netspeak is characterized by distinct graphology that includes a variety of styles and fonts found on websites, as well as simpler systems with minimal typographical distinctions, such as those used in email and chat group conversations. Orthographic features, including capitalization, have all been affected by this. While most Internet is not case-sensitive, capitalization is arbitrary, and lowercase is generally preferred. (Crystal, 2001). In electronic communication platforms such as e-mails, chat groups, and virtual worlds, people often try to save keystrokes by avoiding capitalization and punctuation in their messages. This habit has become so prevalent that using lowercase letters has become the default mode of communication. Capital letters are now considered a separate form of communication, and using them in excess can come across as aggressive or rude, commonly known as "shouting." Netspeak incorporates unconventional spellings, minimal punctuation, and a unique vocabulary, which have become widely accepted and

facilitate innovative communication methods. Punctuation is particularly important in bridging the gap between written and spoken language, conveying intonation and grammatical structure. While changes to grammar are less common, there are specific situations or groups in which variations may occur. A phenomenon in chat groups is verb reduplication, where repeating a verb twice quickly can convey different meanings, such as pleasure, pain, sarcasm, or the end of an utterance (Crystal, 2001).

3.1.3 Experimental Testing

Kubin and von Sikorski (2021) summarize research studies that employ experiments that manipulate media to explore how media can shape political polarization. All the experiments find that social media can increase ideological polarization. In particular, the studies find that exposure to negative Tweets about candidates, uncivil Facebook comments, and counter-attitudinal Twitter posts make people more ideologically polarized. Some studies also explore ideological differences. The study indicates that exposure to counter-attitudinal content resulted in greater ideological polarization among Republicans but not Democrats. However, no insights were found into how social media can reduce or have no impact on ideological polarization. As for affective polarization, nearly all experiments show that social media can further polarize individuals. Specifically, YouTube algorithm recommendations and exposure to derogatory social media comments about political adversaries can increase affective polarization. Moreover, deactivating Facebook before the 2018 United States midterm election has decreased affective polarization. Nonetheless, no insights have been provided into how social media can decrease or have no effect on affective polarization. In summary, the impact of social media on polarization is consistent, predicting both ideological and affective polarization in experimental settings. Furthermore, media coverage of polarization tends to exacerbate polarization, although this effect is not always observed. These findings emphasize the need for further research into the factors that contribute to media coverage of polarization and how media can be used to mitigate political polarization.

3.1.3.1 Quasi-Experimental Setup

A quasi-experiment is a research method that seeks to evaluate the impact of an intervention on a particular population but does not involve the random as-

signment of subjects to treatment and control groups (DiNardo, 2016). This approach shares similarities with conventional experimental designs and randomized controlled trials but without the randomization aspect. Instead of randomization, the researcher has control over the assignment of participants to the treatment group based on specific criteria, such as eligibility cutoffs. Quasi-experiments face challenges when it comes to establishing causality, as the treatment and control groups may not be evenly matched from the outset (Rossi et al., 2004). This makes it difficult to establish a clear connection between the intervention and the resulting outcomes. The absence of randomization further exacerbates this issue, as uncontrolled variables may impact the results. In contrast, when randomization is used, participants are chosen for either the treatment or control group randomly, giving both groups an equal chance of being selected. This helps to balance out any existing differences between the groups, reducing the risk of confounding variables affecting the results. As a result, any changes in characteristics after the intervention can be attributed to the intervention itself. Quasi-experimental designs are a practical alternative to accurate experimental designs that require random assignment and can be challenging to implement due to ethical or practical constraints (DiNardo, 2016). Unlike laboratory-controlled experiments, quasi-experiments have better ecological validity because they take place in natural settings. Moreover, the quasi-experiment results can be generalizable to other subjects and contexts, making them ideal for longitudinal research. The first step in constructing a quasi-experimental design is to identify the variables involved, including the quasi-independent variable (x-variable), which is manipulated to affect the dependent variable (y-variable). The x-variable can take different forms, such as a grouping variable with two or more levels, and the dependent variable is observed over time using a time series analysis. Once the variables have been defined, the procedure is carried out and differences between groups are analyzed (Gribbons and Herman, 1996). In traditional experimental designs, study units are randomly assigned to treatment conditions, ensuring equivalent experimental and control groups. However, in a quasi-experiment, the assignment is based on factors other than randomness, such as cost, feasibility, or convenience. In a quasi-experiment, the researcher may have some or no control over the assignment of participants, and the criteria used for assignment may not be known. This lack of randomization can raise concerns about the internal validity and the ability to establish cause-and-effect relationships from the experiment's

findings. To address potential confounds or biases, quasi-experiments typically involve pre-post testing, where participants are tested before and after the intervention. The pre- and post-test results are then compared or analyzed to explain the experimental outcomes. In a quasi-experiment, naturally occurring variables such as age, gender, or eye color are often measured and can be continuous or categorical (Morgan et al., 2000). Although some may be hesitant to accept quasi-experimental designs (Campbell, 1988), they can still be precious in situations where it is not feasible or desirable to conduct a randomized control trial or a traditional experiment. These scenarios can include evaluating the impact of policy changes, educational initiatives, and large-scale healthcare interventions. The primary limitation of quasi-experimental designs is their inability to remove the possibility of confounding bias, making it challenging to establish causal relationships. However, this issue can be addressed through statistical methods such as propensity score matching or multiple regression. These methods can help isolate confounding variables' impact, leading to more accurate results from quasi-experiments. Quasi-experimental studies have been known to yield results consistent with experimental studies, despite using different approaches to data collection (Armstrong and Patnaik, 2009). Additionally, quasi-experimental studies offer certain benefits over natural experiments, such as greater control over the manipulations being carried out by the researcher. Self-selection is another advantage of quasi-experimental studies, as it eliminates concerns related to ethics or other issues during the study (DeRue et al., 2012). However, confounding variables may reduce the causal link between an intervention and its outcome in quasi-experimental studies. Factors that could affect the results and the validity of the findings are difficult to rule out in such studies. Internal validity is crucial in quasi-experiments as they are designed to examine cause-effect relationships, and maintaining internal validity involves controlling all factors that could influence the outcome of the study. However, factors such as statistical regression, past events, and participant characteristics can compromise the internal validity of the results. To ensure high internal validity, it is important to consider alternative explanations for the observed outcome besides the intended cause. If there are other explanations, the internal validity of the study may be weakened. The concept of external validity pertains to the degree to which the findings of a study can be extended to a broader population of interest, as well as across different subgroups, times, settings, and research methods (Cook and Campbell, 1979). Although it may only sometimes be feasible to

generalize the results to a population, the critical concern is whether the effects of the treatments can be applied across diverse subgroups. This depends on the findings' consistency and the interventions' impact across various subsets of people, contexts, periods, and research methods. The external validity of a study is influenced by the extent to which treatment effects remain consistent or vary across different subgroups and by the researchers' awareness and understanding of these variations (Cronbach, 1975).

3.1.3.2 Activity and Complexity Metrics

We hypothesize that activity is a significant factor in the process of radicalization. The more involved users are in a particular topic or group, the more active they become, leading to continued interaction with the same content and individuals. We also postulate that homogeneous groups are less diverse regarding language and topics. These mechanisms are familiar to us from marketing and behavioral research. The primary objective of social media platforms is to maintain user activity, which necessitates constant interaction with content and other users. There are numerous methods of achieving this, and contribution is the aspect that has received the most attention from researchers (Trunfio and Rossi, 2021). Contribution encompasses activities like commenting, sharing, or liking pre-existing content and is popular due to the interactive nature of these behaviors. Other studies focus on creation, such as publishing content, uploading multimedia, or writing articles, as a measure of social media engagement. Researchers have categorized metrics for measuring social media engagement into various groups. The most extensive group is "quantitative metrics," which provide a straightforward assessment of the effects of social media engagement based on metrics such as the number of comments, likes, shares, and followers. The second category is "normalized indexes", which aim to quantitatively measure the level of engagement generated by a specific content relative to the number of people who have viewed it (Trunfio and Rossi, 2021). **Lexical Diversity** In quantitative text analysis, lexical diversity is a metric to determine a text's complexity level. It is a standard metric for assessing the readability of texts. Lexical diversity measures the number of unique words used in a text or speech as a proportion of the total number of words (Jarvis, 2013). It is often used to assess the vocabulary size and richness of a piece of writing or speech. A text with high lexical diversity will use a wide range of different words, while a text with low lexical diversity will use a smaller number of words more frequently.

Lexical diversity is usually calculated by dividing the number of unique words by the total number of words in a text and expressing the result as a percentage. For example, if a text contains 100 words and uses 50 unique words, the lexical diversity of the text would be 50%. This indicates that the writer or speaker has a relatively small vocabulary and uses a limited range of words. On the other hand, if the exact text contained 100 words and used 80 unique words, the lexical diversity would be 80%, indicating a more extensive vocabulary and a greater range of words being used. Lexical diversity is an essential concept in linguistics. It can be used to compare the complexity and richness of different texts or to assess the language abilities of writers or speakers. It is often used in conjunction with other measures of language complexity, such as word frequency and sentence length, to provide a complete picture of the language used in a text. The study of lexical diversity (or lexical variety) refers to the richness of the vocabulary used in a corpus. In applied disciplines such as language acquisition, speech pathology, and stylometry, measuring this richness is an essential method for, for example, measuring the growing vocabulary of a child or the reduced vocabulary of people with speech disorders (Bonvin and Lambelet, 2017). It was also used to measure the effect of team performance (Shi et al., 2019). Jarvis (2013) distinguishes six components that influence how we perceive lexical diversity: variability, volume, evenness, rarity, dispersion, and disparity. It is believed that the diversity of words in language use indicates the complexity of vocabulary knowledge and the person's language proficiency level. The methods used to measure lexical diversity are helpful in terms of overcoming the effects of task difficulty and predicting language knowledge and behavior but less so in terms of actually measuring lexical diversity. A common problem with various existing methods for measuring the richness of a vocabulary is that they need to consider the length of a piece of writing when determining the number of distinct words used (Bonvin and Lambelet, 2017). This can pose a challenge in comparing texts of varying lengths. Bonvin and Lambelet (2017) review the various relative measures and indices researchers have devised to tackle this issue. Consequently, the capacity of a program to gauge the lexical diversity of a text without the result being skewed by the length of the text has become a crucial criterion for its adoption. Johnson introduced a method called type-token ratio (TTR) to measure lexical diversity. TTR is determined by dividing the total number of unique words (types) by the overall number of words (tokens) in a given text. TTR is an uncomplicated means of measuring lexical diversity. However, it does

not consider the text's length, making it challenging to compare TTR values across texts of differing lengths. When the length of a text increases, the frequency of already established unique words (types) and prevalent grammatical words like "the" and "and" typically increases. In contrast, the number of new unique words (types) diminishes. As a result, the lexical diversity may decline as the text length grows.

3.1.4 Network Text Analysis

In the last few decades, classical text analysis methods that relied on word and phrase counting have been enhanced by statistical analysis techniques that consider other variables, such as those used in semantic and network text analysis. Traditional text analysis involves identifying a text's themes, topics, and concepts, while semantic analysis focuses on analyzing the relationships between sentences or clauses that contain themes. On the other hand, network text analysis examines the position of themes and sentences within networks of related themes (Roberts, 2000). Roberts suggests these three approaches are not mutually exclusive and can be integrated into a single analysis. To analyze texts in a network, text blocks can be transformed into networks of interconnected themes, generating variables to assess the positions of themes and theme relations. For instance, the network of themes representing causal relations can be used to measure the "causal salience" of theme A on another theme B by determining the proportion of all causal linkages in which A is the cause and B is the effect. This measure can be calculated by assigning theme-A and theme-B labels to any pair of themes in the network (Roberts, 2000). These measures can be used to generate a data matrix. Additionally, other measures, such as a theme's "conductivity," can be used to characterize networks, which refers to the number of linkages that a theme provides between other pairs of themes, as Carley (2000) described. Therefore, network text analysis is based on the idea that encoded statements can form networks, which can be analyzed using different variables and measures to understand the interrelationships between themes in the text. According to Kroeger (2005) co-occurrence is a linguistic concept that refers to the frequency with which two or more words or linguistic elements appear together in a text or a corpus. It can be used to measure semantic similarity or idiomatic usage. We can identify typical word combinations for lexical items in a language by analyzing co-occurrence patterns through corpus linguistics and statistical analysis. This analysis expands word frequency analysis into higher di-

mensions and can be quantitatively described using measures such as correlation or mutual information (Kroeger, 2005). To compare words for similarity, the first step is to observe their co-occurrence frequencies, denoted as $n_{A,B}$ for any word A and B. These frequencies are then interpreted using a co-occurrence significance measure, taking into account the individual word frequencies n_A and n_B, as well as the corpus size n. This measure can quantify the degree of association between two words in a corpus (Bordag, 2008).

3.2 Quantitative Content Analysis

3.2.1 Content Analysis

Content analysis is a research method that enables researchers to make accurate and dependable inferences from textual data to their usage contexts. The procedure involves specialized techniques that researchers can learn and utilize regardless of their technical expertise. Replicability is the critical component of reliability. Content analysis adheres to the methodological principles of reliability and validity, which are not exclusive to this approach but place specific requirements on it (Krippendorff, 2013). In the view of Berelson (1952), content analysis is a research technique that involves a methodical, objective, and quantitative description of the clear message of the communication. Krippendorff's definition encompasses the concepts of objectivity and systematization, as stipulated by Berelson, within the more extensive principles of replicability and validity. Replicability necessitates that a procedure is guided by specific regulations that are evenly applied to all units of analysis. Berelson emphasized the importance of systematization to counteract people's natural tendency to read texts, confirming preexisting expectations rather than considering contrary evidence (Lasswell, 1948; Berelson, 1952). Content analysis was introduced as a scientific method at the beginning of the last century. Theoretically, it was most heavily influenced by the then-prevailing positivist-behaviorist tradition focusing on quantitative research. Since its scientific foundation, content analysis has systematically evaluated much larger samples. The explicit orientation towards existing theories (with the formulation of densely formulated questions and hypotheses developed from the literature), as well as measurable quality criteria, is very similar to the approach of quantitative methods. Consequently, the calculation of inter-coder reliability for securing the quality of the results (avoiding excessive variances in the category system) corresponds to the inter-

rater reliability of quantitative research and is mandatory in content analysis research. There are different types of calculation. Mayring (2019) recommends calculating with the Kappa coefficient, which should be at least .70. Measuring the inter-rater agreement ensures the results' reliability. Quantitative analysis steps can also be integrated into the analysis process, e.g., to indicate cross-case agreements. In Germany, Phillip Mayring especially has presented four methods of Qualitative Content Analysis since the 1980s, which are based on three basic techniques: (1) summary content analysis and (2) inductive ("concept-driven" (Kuckartz, 2019) category formation, using the technique of summarization, (3) explicative content analysis using explanations and (4) the structural (deductive or "data-driven" (Kuckartz, 2019)) content analysis uses structuring. The inductive category formation follows a material reduction by summarization. That is the initial material is paraphrased (reduced to the factual part of the statement) and then selected on the basis of overlaps. The selection is followed by bundling the categories developed through integration into other categories (subsumption). This step is followed by another selection and deletion and a renewed bundling and integration of categories. Before the evaluation begins, the object and goal of the analysis must be precisely defined. This is usually achieved by formulating a specific and theoretically justified research question. The material (the cases) on which the category system is to be developed is defined. The specific formulation of a research question, which guides the entire analysis, determines the selection criteria. As a result, all text passages that do not provide information about the research question are disregarded. Furthermore, before the analysis begins, the level of abstraction of the categories to be formed must be regulated once the level of abstraction is maintained throughout the entire analysis. When the above formal requirements have been specified, the material is reviewed line by line until a selection criterion is met. Considering the level of abstraction, a category (defined with a single term or a short sentence) is formulated. The review is continued until the next time a text passage is found that meets the selection criterion. Since the formulation of categories is based solely on specific text passages, the categorization considers whether the text passage can be subsumed under an existing category or whether a new category should be formed. The subsumption is carried out at the previously established level of abstraction rather than by forming a higher-order category. The (relatively quickly performed) method makes it possible to code the entire selected transcript excerpt. The material is processed using the procedure mentioned above

until, in the cases consulted, no new categories need to be formulated, and the existing categories are judged to be sufficient. With a considerable sample size, this may be possible as soon as 10% of the material has been processed. When this point is reached, it is necessary to check whether the categories help answer the research question and whether the level of abstraction has been reasonably selected for the research goal. If changes are necessary, the entire material analyzed so far must be reviewed again. Once the revision is completed or has been found to be unnecessary, the analysis of the entire material can continue. If a is necessary, the entire category system (again) must be adjusted. Further analysis can be done in different ways after the categorization of the complete material. It is possible to interpret the entire category system concerning the research question. In order to achieve further reduction, main categories can also be formed, which either follow an inductive (from the text) or a deductive (based on theoretical knowledge) approach. Quantitative analysis, such as frequency counts of categories, can be accomplished.

3.2.1.1 Syntactical and Semantic Distinctions

Syntactical and semantical distinctions are made based on the words of a selected text. Distinctive speech features are determined through considerations of lexical semantics. For the role of the lexicon in the communication process, we refer here to Lasswell (Lasswell, 1948). In communication-oriented linguistics, language is understood as a practical means of interaction in which the speaker's utterances are intentional forms of action that affect the social environment (Bühler, 2011). Performance phenomena in this respect are of great linguistic interest:

- who is using language
- to whom language is spoken
- what language is used
- where language is used
- how language is used
- what language is used for

In addition to Lasswell "Who says what in which channel to whom with what effect?", this approach considers the various roles that language plays in communi-

cation. Bühler (2011) identifies three functional aspects of language: representational functional, appeal-functional, and expressive-functional. When categorizing verbal actions, it is crucial to determine whether an utterance is based solely on non-partner-related behavior or whether it constitutes a partner-directed language action (Searle, 1969). Action-oriented linguistics aims to describe the historical and social context of language use and the various social functions that language serves in specific speech situations. This approach emphasizes the communicative aspects of language rather than just the lexical ones, which allows for assigning a particular function to the lexicon in communication. Lexicology is a fundamental aspect of linguistic communication and serves as a means of intentional, partner-effective verbal interaction. As such, it acts as a repository for various communicative functions. The inventory of the lexicon is used for acts of reference, allowing speakers/writers to perform various speech acts like warning, commanding, explaining, judging, and evaluating. Performative verbs are lexical units that have their illocutions lexicalized in meaning. The selection of a performative verb indicates the illocutionary act performed by an utterance. By using verbs like "command," "instruct," and "prescribe," speakers/writers can perform exertive utterances (Austin, 1962) that readers/listeners can (mainly) accurately interpret. Additionally, selecting elements linked to a conventional positive or negative appraisal can function as a cue for a specific speech act or intensify its intended effect. When using utterances, speakers leverage their authority ("Be quiet please!", "Be quiet!", "Shut up!", "Shut the fuck up!", "Shut up!"). While the direct speech act can be seen as a polite request through the use of "please," the illocution of the elliptical utterance ("Shut up!") and that of the utterance ("Shut the fuck up!") can be interpreted as a command. The utterance ("Shut the fuck up!") intensifies the command of "Shut up!" by selecting "shut the fuck up" from the lexical potential of vulgar language, which signals an unyielding attitude on the speaker's part. This use of vulgar slang expresses the tone of command functionally and stylistically. The stylistically colored lexical choice provides additional information, whether as an intensification or specification of what is meant or as a subjective evaluation indicating a particular attitude of the speaker toward the hearer. This marking of lexical potential adds additional information specifying the illocution of an utterance. Distinctions in meaning can be conducted on the lexical level and the level of textual structure. Textual structure refers to how a text is organized, including elements such as sentences, paragraphs, headings, and subheadings (Crystal, 2001). The structure of a text

influences how it is perceived and understood by its audience. Discourse refers to how language is used in a particular context, while text refers to the written or spoken representation of that discourse. The two are closely related, as the discourse informs the structure and content of the text, and the text reflects the discourse it represents. How a text is structured can affect the reader's interpretation of its content, and different structures can convey different meanings and purposes. The structure also indicates how words and sentences are organized to form a coherent and meaningful whole. The cohesive factors of the textual structure are the linguistic ties that link the elements of a text together, making it a unified and integrated piece of discourse. These factors include lexical cohesion, reference, ellipsis, conjunction, and lexical sets. Therefore, pragmatics is a subfield of linguistics that studies how context influences language interpretation. It focuses on how speakers and listeners use language in context to convey meaning and how they can understand each other's intentions and meaning despite potential ambiguities in language. According to linguist David Crystal, pragmatics is concerned with the "unwritten rules" of language use, which are often context-specific and may vary from culture to culture. It examines how language is used in different social situations, such as how people use politeness strategies to convey respect or irony to convey humor or sarcasm. Pragmatics also examines how language is used to achieve specific goals, such as persuading, informing, or entertaining. It concerns how speakers and listeners negotiate and how they use language to convey and interpret social cues (Crystal, 2001).

3.2.1.2 Categorical Distinctions

We define units by membership in a particular category or class based on shared characteristics. A typical reference point for these units is any word or phrase that refers to a specific object, event, person, action, country, or idea. The grammatical structure or perspective used to refer to is of secondary importance compared to the unit's categorization. Categorial distinctions usually rely on taxonomies, aside from synonyms. These distinctions can also arise from a particular theoretical framework used for analysis. Early content analysts categorized symbols, typically single words, based on their denotations and associated values, attributes, and qualifications. Adjectives were considered necessary for proper categorization within this framework (Krippendorff, 2013).

3.2.2 Developing a Taxonomy

We refer to Allan (2002) and define a classification system as a methodical collection of abstract categories, concepts, or types used for organization and definition. These categories are often created by grouping objects based on shared characteristics or traits. Many classification systems are organized hierarchically, with varying levels of detail. The names of the categories form a controlled vocabulary. The process of assigning an object to a category within a classification system is called classification or class assignment. Taxonomies are constructed in order to classify things. Researchers design these models to capture individual cases and enable classification based on specific criteria. Taxonomies originate from the Greek words "taxis," meaning order, and "nomos," meaning law, which are standardized models used to categorize objects into classes or categories (also known as taxa). Scientific disciplines often use taxonomies to create a hierarchical classification system, including classes and subclasses. Taxonomies are vital to the development of scientific disciplines as they enable researchers to handle individual cases and make summary statements that can lead to an explanation of relationships. They help to clarify the distinctions between categories and enhance understanding of the research area (Allan, 2002).

3.3 Social Network Analysis

Social network analysis is an interdisciplinary field that encompasses several academic disciplines, including social psychology, sociology, statistics, and graph theory. Émile Durkheim and Ferdinand Tönnies initially suggested the concept of social networks in the late 1890s in their research on social groups. Tönnies argued that social groups consist of direct social connections that link individuals sharing common values and beliefs or impersonal, formal, and instrumental social ties. Durkheim offered a non-individualistic interpretation of social phenomena, stating that interacting individuals create a reality that cannot be explained solely based on individual actors' characteristics. Georg Simmel conducted a study that explored the characteristics of networks and how the size of a network affects interaction. He also focused on the nature of interaction in networks that are not tightly bound as opposed to organized groups. During the 1930s, social network theory gained significant popularity across multiple disciplines, including psychology, anthropology, and mathematics. Jacob L. Moreno conducted an in-depth analysis of social interaction in small groups. Early work

by Talcott Parsons in sociology paved the way for a relational approach to understanding the social structure, later further developed by Peter Blau's social exchange theory. In the upcoming years, there was a trend of combining different social network traditions and approaches. Harrison White and his students focused on social networks in various contexts. Charles Tilly explored the role of networks in politics and social movements. Mark Granovetter and Barry Wellman further developed and popularized social network analysis. Stanley Milgram contributed the concept of "six degrees of separation." The late 1990s saw social network analysis incorporate data from online and face-to-face networks, with contributions from scholars like Duncan J. Watts, Albert-László Barabási, and Peter Bearman, who applied new models and methods to study the emerging data in these areas. Social network analysis is an empirical social research method used to capture and analyze social relationships and networks. It promotes a particular view of social phenomena that emphasizes their relational nature. Connections and interdependencies between units (such as individuals or organizations) are the focus, rather than their attributes and characteristics. Hence, the social relationships and their structure become the unit of the analysis. Formal representations allow for graph-theoretical interpretations of social networks. A network is represented as a graph with a defined set of nodes representing the actors in a network and edges representing the relationship. It can be combined with sociometric and algebraic methods for more complex analyses. The network is then translated into a sociomatrix, a tabular listing of the nodes and their relationships.

A graph is an ordered pair $G = (V, E)$ comprising:

- V a set of nodes (also called vertices);
- $E \subseteq \{(u, v) \mid u, v \in V\}$ a set of edges (also called links or ties).

There are "undirected" and "directed" networks based on the symmetry or asymmetry of the relationship between pairs of nodes in a graph. When the relationship is symmetric, we refer to the edges as "undirected," which are considered unordered pairs of nodes. Conversely, when the relationship is asymmetric, the edges are referred to as "directed" and are considered to be ordered pairs of nodes. In cases where the relationship between nodes is assigned a strength value, the edges are assigned numeric weights, and we refer to the network as "weighted". Social network analysis involves using various metrics to characterize networks at the node level (degree and centrality) and the network level (density, diameter, and clustering coefficient). Advanced analysis techniques such as community

detection, diffusion dynamics, and link prediction can also be applied to social networks (Ghawi and Pfeffer, 2020). We refer to (Hennig et al., 2012) and summarize the various techniques for social networks analysis. One such technique is similarity, which refers to how people connect with others with similar traits, such as age, gender, education, values, and status. Another method is multiplexity, which measures the type of content present in a connection and is linked to the strength of the relationship, whether positive or negative. Mutuality, or reciprocity, assesses how much two actors reciprocate each other's interactions. Network closure measures the completeness of a set of relational triads, while propinquity refers to the tendency for individuals to form connections with those who are geographically close. Finally, transitivity, or the belief that friends of an individual are also friends with each other, is related to the need for cognitive closure, which is a trait or situation. In social networks, a bridge is a person who links two individuals or groups that do not have a direct relationship. It helps fill gaps in the structure and creates a direct link between different network parts. Centrality is a set of metrics used to measure the importance or influence of a node in a network. Common ways of measuring centrality include betweenness, closeness, eigenvector, alpha, and degree. Density is the proportion of direct connections in a network compared to the total possible number of connections. Distance is the minimum number of connections required to link two individuals. Structural holes are gaps in connections between two different networks. Tie strength is characterized by time, intensity, intimacy, reciprocity. It defines the strength of a connection. Strong connections are typically linked to similarity, proximity, and transitivity. Groups can be classified based on their connections, such as 'cliques' where all members are directly connected, 'social circles' where connections are less strict, or 'structurally cohesive blocks,' which have a precise definition. The clustering coefficient measures the likelihood of two contacts of a node being connected. A higher coefficient indicates a higher level of 'cliquishness'. Cohesion measures the strength of connections between members within a group, while structural Cohesion refers to the minimum number of members that must be removed to break the group's connections. Essential terms in social network analysis include density, centrality, in-degree, out-degree, and sociogram. The metric of density in social network analysis is calculated by dividing the number of connections an individual has by the total number of possible connections. Centrality is a measure of an individual's level of interaction within a network, with a higher number of connections indicating greater centrality.

In-degree centrality is a measure of the centrality of other individuals based on their connection to a specific individual, whereas out-degree centrality measures the frequency with which a particular individual interacts with others.

3.3.1 Change Detection

We refer to McCulloh and Carley (2008) and define social network change detection as monitoring networks to identify significant changes to their organizational structure. Changes can be detected by combining techniques from Social network analysis and statistical process control. To apply this approach, statistical process control charts are utilized to identify changes in measurable network factors. By monitoring network measures over time, a control chart can signal when significant changes occur (McCulloh and Carley, 2008). According to McCulloh, Social Network Change Detection offers a substantial improvement over previous methods of detecting changes by introducing a statistically sound probability space and powerful detection techniques that are uniformly effective. The literature provides various techniques for studying social networks over time (Goodreau, 2007; Snijders et al., 2007; McCulloh et al., 2007). Methods such as preferential attachment and fitness models have been employed as conceptual models to forecast the development of networks over time. Although it is unclear which approach most accurately represents the actual progression of networks, all techniques offer a way for analysts to comprehend the potential underlying statistical distribution of social network measures. McCulloh and Carley (2008) summarizes in his work that measures of average centrality, average betweenness, and density follow a normal distribution for networks with over 30 nodes. One can calculate these metrics for the entire network or each node separately. Network metrics such as betweenness and closeness centrality are often employed to gain insight into how information propagates in a social network because of the applications of real-world scenarios (McCauley and Moskalenko, 2008).

3.3.2 Statistical Process Control

In the field of quality engineering, Statistical Process Control (SPC) is a method used to monitor and identify changes in a process. SPC involves using control charts to measure the results of periodic product samples against a predefined control limit. Engineers use this method to detect any changes in the mean, determine the likely time of the change, and take action to prevent financial

loss. McCulloh and Carley (2008) discusses that control charts are designed to increase the sensitivity for detecting changes while reducing false alarms, which occur when no change occurs. A similar technique is applied to Social Network Change Detection, where control charts track measures such as density, closeness, and betweenness centrality over time. Significant changes in these measures are detected using statistical process control methods to identify any changes in the network structure (McCulloh and Carley, 2008).

3.4 Methodological Considerations

The chapter provided an overview of standard quantitative methods for analyzing and evaluating language and behavioral data. With these methods, properties can be extracted from text, emotions can be measured, experimental effects can be measured, changes in social networks can be captured, and content can be extracted from language and communication data. The quantitative methods enable descriptive capturing of text data on the content level, calculation of correlations (Strathern et al., 2020b), and higher statistical analyses with which we can develop models (Strathern et al., 2022b) that are applicable to many text data. The experimental methods and the linguistic metrics presented, in turn, enable the testing of effects. At the same time, the effectiveness of metrics from linguistics can be evaluated (Strathern et al., 2022a). Structural group and text properties can be measured through social network analysis and semantic network analysis (Strathern et al., 2020b, 2022b,a). In combination with quantitative methods, they deliver promising results. Quantitative content analysis, on the other hand, requires a more vital understanding of text, language, and culture and corresponds more closely to qualitative approaches. With the addition of theories and background literature, linguistic and social phenomena can be observed in depth (Strathern and Pfeffer, 2023; Wich et al., 2021). They are to be considered complementary to purely quantitative methods.

Chapter 4

Compilation of Social Media Data

There are challenges when working with social media data. The vast number of social media posts and comments are brief, informal exchanges containing limited information, making them difficult for Natural Language Processing tools to process. These texts often contain non-verbal contextual information, such as the user's profile, social network, and interactions with others. This rich context and the way it interacts make it challenging to automate the analysis of social media content. Traditional text mining methods struggle with this task because they need to consider the interactive dimension or the unique properties of social media data, which have spoken and written language characteristics. According to Chen et al. (2018) social data is generated by the interaction between humans and machines. It primarily comes from human language and human-machine interactions, as opposed to other instrument-measured data. Social data typically comprises metadata or structural data and content data. Metadata, such as user account information, time of post, and a serial number, are usually in list form and created by computer systems. Analyzing content data, however, is more challenging. It requires significant human and material resources to clean and organize. Human language is diverse and complex, and the connotations of text can be challenging to grasp. Additionally, there are various ways in which human language can be used, such as opinions, evaluations, and irony. Data processing for social media analysis has a specific procedure that includes collecting, cleaning, and visualizing data. A data analysis with computational methods has many challenges and limitations regarding interpretability and explainability (Radford and Joseph, 2020; Lipton, 2018). The notion of human behavior is multifaceted and often necessitates examining individual-level data to conclude the distribution of behaviors, attitudes, and attributes at a collective level.

4.1 Microblogging Data from Twitter

Social media text data comes as semi-structured data that presents a challenge for linguistic analysis. This is a crucial consideration for our methodological approach and is addressed in this chapter. The significant advantage is its interactive dimension that allows for analysis to discover changes in the network over time. The term "Twitter data" is not entirely accurate as it comprises both structured and unstructured data, making it more appropriate to describe it as semi-structured. The tweet text is an example of unstructured data. In contrast, structured data is information that can be organized into the predefined fields of a database table's columns and is usually stored in a table format. Structured data is often referred to as relational data because it has a unique identifier, or "key," that can be mapped to other tables. For example, the time a tweet was posted is structured data that can be easily mapped to a "time" column in a database table. Twitter data is structured in a way that allows for easy access and analysis. Each tweet is stored as a separate record, and each record includes a variety of fields or attributes that describe the tweet. These fields can include the tweet text, the user who posted the tweet, the date and time the tweet was posted, the location of the user, and any hashtags or other metadata associated with the tweet. Twitter data can be accessed through the Twitter API (Application Programming Interface) or third-party tools that allow for the collection and analysis of Twitter data. Twitter data can be analyzed in a variety of ways, including through the use of natural language processing techniques to extract meaning from the text of tweets and through the use of social network analysis techniques to study the relationships between users. One particular thing about Twitter's data structure is its focus on brevity. Twitter limits the length of tweets to 280 characters, encouraging users to be concise and communicate their thoughts and ideas in a condensed form. Another unique thing about Twitter's data structure is the use of hashtags and @mentions. Hashtags label tweets with specific keywords or themes and can be used to discover and follow conversations on specific topics. @mentions allow users to mention other users in their tweets, which can be used to draw attention to specific accounts or to engage in conversations with others. Twitter's data structure facilitates the rapid exchange of information and ideas and allows users to quickly discover and engage with content on a wide range of topics. Semi-structured data combines both structured and unstructured data, allowing for some degree of semantic tagging for data organization. The defining characteristic of semi-structured data is its classifica-

tion or tagging. Therefore, it is inaccurate to classify Twitter data as completely unstructured data; instead, it falls into the semi-structured data category. As data volumes expand and schema changes become more frequent, platforms are turning to graph-based storage to handle the complexity of unstructured data. Graphs allow for storing relationships between data points, which can be queried and organized using nodes and edges. In conclusion, Twitter data can be classified as semi-structured to unstructured data. The pitfalls and challenges of using Twitter data and its API have recently been discussed (Pfeffer et al., 2023b,a).

4.2 Chat Group Data from Reddit

Reddit data is also semi-structured data hierarchically, with posts and comments organized into "subreddits," which are dedicated forums or communities centered around a specific topic or theme. Each subreddit is organized into a series of threads: collections of posts and comments on a specific topic. Posts and comments are organized in a tree structure, with comments being nested under the post or comment to which they respond. Reddit data includes several fields or attributes for each post or comment, such as the text of the post or comment, the user who posted it, the date and time it was posted, and any metadata associated with the post or comment, such as upvotes or downvotes. Reddit data can be accessed through the API (Application Programming Interface) or third-party tools that allow for the collection and analysis of Reddit data. Reddit data can be analyzed in various ways, including natural language processing techniques to extract meaning from the text of posts and comments and network analysis techniques to study the relationships between users and communities on the platform. One thing that is special about Reddit's data structure is its hierarchical nature, with posts and comments organized into "subreddits" and threaded into discussions. This structure allows users to navigate and engage with content on specific topics easily and to discuss with others interested in the same topics. Another unique thing about Reddit's data structure is the option for users to upvote or comment on posts and comments. This system allows the community to collectively determine the relative importance or relevance of different pieces of content and can influence the visibility and engagement of specific posts or comments. Overall, Reddit's data structure facilitates discussion and community engagement and allows users to discover and participate in discussions on various topics quickly. During the relevant time period for this study, Reddit ex-

perienced a significant increase in monthly active users, rising from 46 million in 2012 to 430 million in 2019, as reported by Reddit Revenue and Usage Statistics (2023) - Business of Apps [3]. This growth underscores the increasing importance of Reddit, not just for its users, but also for research. The platform allows for lengthy, detailed comments, making the data more comprehensive than on other social media platforms. Additionally, Reddit's unique structure of niche-oriented communities makes it possible for researchers to analyze data in a more topic or community-specific manner, supporting studies such as language analysis, user modeling, sentiment analysis (Strathern et al., 2022a; Medvedev et al., 2019). However, due to Reddit's lack of content moderation until 2018, the platform's niche-oriented communities provided an ideal environment for radical and extreme content and groups of various kinds. Previous research has shown that movements such as the Alt-Right, Q-Anon, Incels, Men's Rights Activists, and different conspiracy theorists actively used Reddit to recruit new members and spread (mis)information (Horta Ribeiro et al., 2021). Communities on Reddit, known as "subreddits", which often contain explicit, violent, or hateful material and have been the subject of controversy, are referred to as controversial Reddit communities. These subreddits can receive significant media attention. The basis for our work is social media text and network data. The brevity of tweets and characteristics such as mutual mentions and indexing of topics make Twitter a particularly suitable source of data for analyzing firestorms (Strathern et al., 2020b, 2022b; Strathern and Pfeffer, 2023). The length and elaborate communication exchanges in Reddit's chat groups make this a more suitable platform for capturing the long-term effects of social processes in communities (Strathern et al., 2022a).

4.3 Ethical Considerations

Because the data obtained from Twitter and Reddit is of a delicate nature, it is crucial to acknowledge ethical issues in this investigation. The Twitter and Reddit discussions and comments in general are readily accessible to the public. To safeguard the confidentiality and privacy of the users, usernames should be omitted from the report, upholding the ethical principles of any study.

[3]https://www.businessofapps.com/data/reddit-statistics/

Chapter 5

Publications

5.1 Overview of Publications

This chapter presents the published articles clustered according to topics. The first article deals with the statistical analysis of large-scale web-based data (Strathern et al., 2021). The article discusses methodological requirements for the statistical evaluation of this new form of data. The second article presents a tool for anonymizing text data and presents the classification schema (Strathern et al., 2020a). The next two papers (Strathern et al., 2020b) and (Strathern et al., 2022b) are thematically linked to the discussion on negative word-of-mouth. These papers investigate and evaluate communication structures and behavioral changes in negative word-of-mouth, in Strathern et al. (2020b) we apply quantitative text analysis and extract features based on sentiment analysis. Based on these features we built a detection model. In addition, in Strathern et al. (2022b) we used the extracted features to built a prediction model. The next papers deal with polarization and radicalization and includes the analysis and experimental testing of behavioral changes applying metrics from linguistics Strathern et al. (2022a). Papers six (Wich et al., 2021) and seven (Strathern and Pfeffer, 2023) address the linguistic evaluation of hate speech, including considerations of radicalization and communicative aggression. The focus of these two works is on developing a taxonomy to capture hate in its various linguistic facets. We conclude the chapter with a conclusion in which we answer the overarching questions from the introduction with reference to the papers.

5.2 Methodological Approaches

5.2.1 Advanced Statistical Analysis of Large-Scale Web-based Data

Authors Wienke Strathern, Raji Ghawi, Jürgen Pfeffer

In Data in Economics and Finance for Decision Makers. Per Nymand-Andersen (Ed.), Risk Books, pp.43-72, London, 2021, ISBN: 978-1-78272-394-3, https://www.risk.net/data-science-in-economics-and-finance-for-decision-makers.

Publication Summary

Approach: In Strathern et al. (2021) we present work on advanced statistical analysis. The "Big Data" ecosystem contains millions of digital footprints left by individuals through their daily transactions. As more human interaction, communication, and culture is recorded digitally, text becomes a valuable input for economic research. Statistical and deep learning methods are applied to digital text to extract information on economic and social activity.

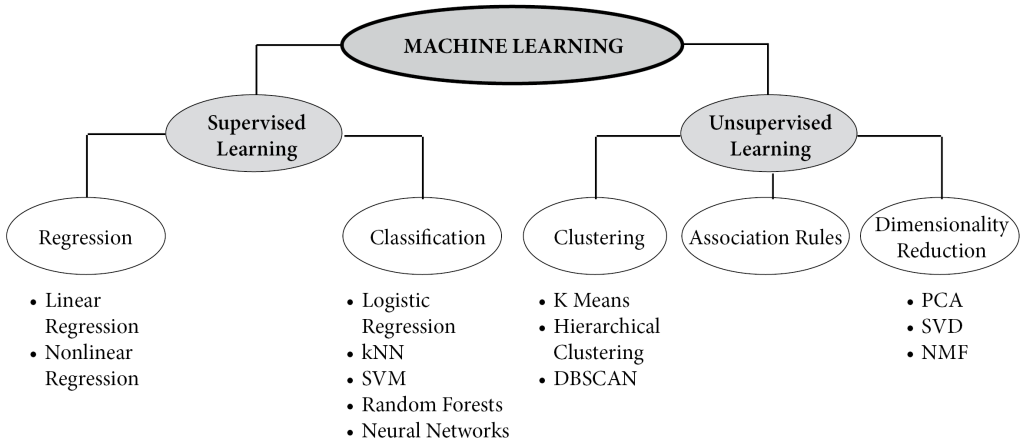


Figure 3: Advanced statistical methods as summarized in Strathern et al. (2021)

This book chapter aims to provide an overview of advanced statistical analysis methods, application areas, and potential pitfalls for decision-makers in economics and finance. The concept of "big data" is presented, highlighting the differences in data formats, and addressing the challenges that arise during analysis. **Methods:** Figure 3 provides a schematic overview of the advanced statistical methods. Decision-makers must know which methods to use for specific purposes, with particular attention paid to data quality. While machine learning methods can assist in decision-making, they cannot entirely replace human intuition. Decision-makers need to comprehend the underlying economics behind the data and signals to achieve desired investment results using large-scale web based data.

Author Contribution

Wienke Strathern headed the project, developed the conceptual framework, wrote the introduction, discussion and conclusion, did a literature review, revision and editing, and coordinated the team.

4

Advanced statistical analysis of large-scale Web-based data

Wienke Strathern, Raji Ghawi, Jürgen Pfeffer
Technical University of Munich

People leave millions of digital traces in the big data ecosystem. This ecosystem is a huge network with millions of daily personal transactions. And each of these transactions leaves traces that may be compiled into comprehensive information about individual and group behaviour (Lazer *et al* 2009, 2020). The capacity to collect huge amounts of data transforms the way people and organisations work and behave; hence, the market starts to react faster and increasingly anticipates traditional or other data sources. Data-driven computational economics capture changes in market, attitude and consumer behaviour over time and in real time. The quantitative techniques of machine learning have been applied to demonstrate a shift from a discretionary to a quantitative investment style (Kolanovic and Krishnamachari 2017). An increasing share of human interaction, communication and culture is recorded as digital text. Text is used as an input to economic research. Statistical methods and deep learning methods are applied to digital texts, as such data provides a rich repository of information about economic and social activity (Gentzkow *et al* 2019; Gentzkow and Shapiro 2010). More interesting for behavioural economics are the large-scale studies of social behaviour (Ruths and Pfeffer 2014). Research on big data analytics for economy and finance, especially quantitative finance, has been widely conducted (Ginsberg *et al* 2009; Engelberg and Parsons 2011; Goel *et al* 2010; Bańbura *et al* 2013; Cook *et al* 2011). A variety of studies have focused on social media data as a data source for finance and for decision makers. Bollen *et al* (2011) used Twitter data to predict

changes in stock market prices. Vermeer *et al* (2019) used machine learning and social media data from Facebook to better understand electronic word-of-mouth and its implications for brands. Ciulla *et al* (2012) used Twitter data to predict social events during elections to anticipate the voting outcome. Twitter data is used in financial market prediction (Mao *et al* 2011) and commonly used as a news source in mainstream media (Moon and Hadley 2014).

We refer to Laney (2001) and define big data by the following features.

- **Volume:** the size of collected and stored files, tables, numbers, etc.
- **Velocity:** the speed of transmitted data in real time or near real time.
- **Variety:** the number of different formats, ie, structured (structured query language (SQL) tables, comma-separated values (CSV) files), semi-structured (JavaScript object notation (JSON) or hypertext markup language (HTML)) or unstructured (social media post, video message).

According to Kolanovic and Krishnamachari (2017), we can differentiate big data sources as follows.

- **Data generated by individuals,** such as social media posts, product reviews and Internet search trends. Mostly recorded through textual mediums, such data is often unstructured and distributed across multiple platforms. We can further classify this data into data from social media, specialised sites such as business-reviewing websites (eg, e-commerce groups), Web searches and personalised data, data from personal inboxes, etc.
- **Data generated by business processes,** such as company exhaust data, commercial transactions, credit card data and order book data. This data refers to data produced or collected by corporations and public entities. An important subcategory is transaction records such as credit card data. Corporate data can be a byproduct or "exhaust" of corporate record-keeping such as banking records, supermarket scanner data and supply chain data. Data generated in this way is often highly structured (compared with individual data) and can act as a leading indicator for business metrics, which tend to be reported at a

significantly lower frequency. Business-processed data can, for example, arise from public agencies.

- **Data generated by sensors,** such as satellite images, foot and car traffic and ship location, is data collected mechanically through sensors embedded in various devices. The data generated is either structured or unstructured and is often much larger in size than either individual or process-generated data streams. An example would be satellite imaging used to monitor economic activities (construction, shipping, commodity production, etc). Geolocation data can be used to track foot traffic in retail stores (smartphone data, if allowed) or ships in ports. Other examples of sensors include cameras fixed at a location of interest and weather and pollution sensors. The practice of embedding microprocessors and networking technology into all personal and commercial electronic devices – the concept of the Internet of Things (IoT) – is the next step for sensor-generated data.

There have been three important trends that enabled big data analytics (Kolanovic and Krishnamachari 2017). The availability of different data sources and a possible application of quantitative strategies can be a huge informational advantage in complex systems. An exponential increase in the data available and an increase in computing power and data storage capacity at reduced cost (cloud computing) increases access to data. There have been increasingly fast developments in the advancement of machine learning methods to analyse complex data sets. One of the biggest advantages is the ability to collect large quantities of data and analyse it in real time. Simultaneously, there has been significant growth in the methodological advancements in pattern recognition and function approximation.

Machine learning methods are often extensions of well-known statistical methods; supervised learning methods attempt to establish a relationship between two data sets and use one data set to predict the other. The underlying concepts of machine learning methods are often as simple as regression models, improved to contain changing market regimes, data outliers and correlated variables. Unsupervised machine learning methods try to understand the underlying structure of data and identify the main patterns. Supervised machine learning methods try to find a rule that can be used to predict a variable (Kolanovic and Krishnamachari 2017). However, skills, infrastructure, market intuition and experiences in complex economic

and financial systems are required in order to handle and evaluate big data and gain insights about the economic drivers behind the data.

In this chapter we showcase the machine learning methods for analysing large-scale data and debate the strengths and weaknesses of these methods. First, we discuss in detail the machine learning methods used to work with big data. Then we discuss an application of these methods with representative data to illustrate advanced statistical analysis.

MACHINE LEARNING METHODS

New methods are needed to tackle the complexity and volume of new data sets. For instance, the automated analysis of unstructured data such as images and social media is not possible with standard analytical tools (eg, spreadsheets). Machine learning methods can be used to analyse big data, as well as to more efficiently analyse traditional data sets. Artificial intelligence (AI) is a broader scheme enabling machines to tackle complex problems in complex systems. In many cases, when a computing problem needs to be solved we often write a program that manually specifies a series of programming steps which need to be run to solve that particular problem. We can instruct a computer to perform certain operations based on a fixed set of rules. For instance, in finance, we can instruct a computer to sell an asset if the asset price drops by a certain amount (stop loss). This works well for a vast number of computing problems. However, not all problems lend themselves to being solved effectively by writing a handcrafted program or a set of rules. Image classification, speech recognition (converting human speech to text) and authorship identification (inferring the author of a document) are examples of tasks that cannot be accurately carried out by writing down a set of rules in a programming language.

Given how complex and delicate those problems are, writing by hand a set of program rules that could solve them would be a tremendous task. Even then, such a hand-crafted system would still likely be inflexible and not very robust at recognising different types of objects (images, speech or text). Moreover, if the system is required to be customised such that it could recognise new objects or other features that had not been encoded in the existing rules, we would have

to write a whole new set of rules, which would be a prohibitively difficult task.

Giving a machine a large number of complex rules for automating tasks is referred to as "symbolic AI". With this "symbolic AI", the machine will freeze the first time it encounters a situation that does not exactly match a set of pre-programmed rules. Machine learning, on the other hand, gives us technology that allows us to automatically learn these complex rules efficiently from labelled examples, called training data, in a way that is much more accurate and flexible than attempting to program all the rules by hand. The goal of machine learning is to enable computers to learn from their experience in certain tasks, and to improve their performance automatically as they gain more experience. This experience can take the form of data in a lot of different formats or situations, such as the labelled examples that are used to train the system's initial structure.

In machine learning, the computer is given an input (set of variables and data sets) and an output that is a consequence of the input variables. The machine then finds or "learns" a rule that links the input and output. The success of this learning task can be tested with respect to its ability to gain useful knowledge of the relationship between the variables and predict outcomes in as yet unseen situations. That is, since it is unlikely any future examples would match what was in the training set exactly, the primary goal of effective machine learning algorithms is to be able to generalise: to correctly predict or recognise new objects that were not seen during training.

Machine learning is a part of the broader fields of computer science and statistics. Statistical methods give machine learning ways to infer conclusions from data (learn from data) and also to estimate how reliable those conclusions are. Computer science methods, on the other hand, give machine learning algorithms the computing power (including effective large-scale computational architectures and algorithms for capturing, manipulating, indexing, combining and performing predictions with data) to solve problems.

Machine learning tasks can be categorised into two main types. The first type is known as "supervised learning", where the goal is to predict some output variable (a predefined label) associated with each input item. The second type deals with data that has no predefined labels, hence the name "unsupervised learning". Here the goal is to find structure in the data by finding some commonality in

its features (Kolanovic and Krishnamachari 2017; Domingos 2012). When we apply machine learning, using either a supervised or an unsupervised approach, a typical workflow consists of three basic components: representation, evaluation and optimisation.

- **Representation.** The first step in solving a problem with machine learning is to figure out how to represent the learning problem in terms of something the computer can handle. This serves two purposes.

    1. The representation of the data (eg, what features to use): we need to convert each input object, which we often call a sample, into a set of features that describe the object.
    2. The choice of the learning algorithm to apply: we need to pick a learning model, typically the type of classifier that you want the system to learn.

- **Evaluation.** The second step is to decide on an evaluation method that provides some type of quality or accuracy score for the predictions or the output of the machine learning algorithm. An evaluation function (or scoring function) is needed to assess and compare the effectiveness of different algorithms (models), and hence to distinguish good ones from bad ones. For example, a good classifier will have a high accuracy, making a high percentage of predictions matching the correct "true" label.

- **Optimisation.** The third step is to search all possible models for the optimal model that gives the best evaluation outcome for that particular problem, ie, the highest-scoring model. This involves an iterative process, where we make an initial guess about what some good features are for solving the problem, and which classifier might be appropriate. We then train the system using training data, produce an evaluation and see how well the classifier works. Then, based on that evaluation, we refine the model and repeat the process.

Typically, data instances are represented as vectors. The components of vectors correspond to the features of the data instances. When a feature is binary (Boolean) or numeric, its values can be used

directly in the corresponding vector component. Non-numeric features need some sort of transformation to be used in vector components. Ordinal features comprise a finite set of discrete values with a ranked ordering between values, such as size (small, medium, large, etc). Ordinal features can be transformed using an integer encoding, for example, "small" = 1, "medium" = 2 and "large" = 3. Categorical features comprise a finite set of discrete values with no relationship between values, such as colour (red, green, blue, etc). Categorical features can be transformed using a technique called one-hot encoding, for example, "red" = $(1, 0, 0)$, "green" = $(0, 1, 0)$ and "blue" = $(0, 0, 1)$.

A common way to represent text in machine learning is the "vector space" model, where each document is represented as a vector whose elements correspond to words in the whole document collection (vocabulary). The values in the vector can be binary (1 for the presence of the word and 0 for the absence of the word). Alternatively, it is common to use the within-document term frequency (TF), which is the number of occurrences of the given term in the given document. Moreover, TF is typically combined with the inverse document frequency (IDF), which is a measure of how common or rare a word is across all documents. The TF–IDF scheme is the most popular scheme for text representation. Using this representation, the similarity between two text objects (sentences, paragraphs or documents) can be assessed using the dot product of the vectors representing them or, more commonly, using cosine similarity (Huang 2008).

When data instances are properly represented as vectors, they are ready to be used in machine learning algorithms. Many machine learning algorithms (in particular, clustering algorithms and the k-nearest neighbours (k-NN) classification algorithm) need some measure of distance (or similarity) between data instances.

Let two data instances be represented by two n-dimensional vectors, $\boldsymbol{A}$ and $\boldsymbol{B}$, with $a_1, a_2, \ldots, a_n$ the components of vector $\boldsymbol{A}$ that represent the values of the features (raw or transformed) of data instance A (such as the occurrence of words in text A), and $b_1, b_2, \ldots, b_n$ the components of vector $\boldsymbol{B}$. Distance measures that are commonly used in machine learning algorithms include the following.

- The Euclidean distance, which represents the shortest distance between two points

$$D(A,B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

- The Manhattan distance, which is the sum of the absolute differences between points across all the dimensions

$$D(A,B) = \sum_{i=1}^{n}|a_i - b_i|$$

- The Minkowski distance, which is a generalised form of the Euclidean and Manhattan distances; a Minkowski distance of order p between two points is defined as

$$D(A,B) = \left(\sum_{i=1}^{n}|a_i - b_i|^p\right)^{1/p}$$

  If $p = 1$, the Minkowski distance reduces to the Manhattan distance. If $p = 2$, the Minkowski distance reduces to the Euclidean distance.

- Cosine similarity, which is a measure of similarity between two vectors defined to equal the cosine of the angle between them

$$\cos(A,B) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$$

  The resulting similarity ranges from -1 to $+1$, where -1 means exactly the opposite, and $+1$ means exactly the same. The cosine distance is the complement of the cosine similarity, ie, $D_{\mathrm{C}}(A,B) = 1 - S_{\mathrm{C}}(A,B)$, where D_{C} is the cosine distance and S_{C} is the cosine similarity.

SUPERVISED LEARNING

The first type of machine learning methods is known as supervised learning. The goal is to predict some output variable that is associated with each input item. The output variable could be a category (with a finite number of possibilities), such as a spam or not-spam email, a fraudulent or not-fraudulent prediction for a credit card

transaction or the topic of a document (eg, sport, politics or the economy). In this case, we call this a classification problem within supervised learning, and the function that we learn is called the classifier. Conversely, if the output variable we want to predict is not a category, but a real-valued number such as the price of a house, then we call this a regression problem, and we are learning something called a regression function.

Regression

Regression is one of the most widely used machine learning tools. It allows us to make predictions from data by learning the relationship between features of the data and some observed, continuous-valued response. Regression is used in a very large number of applications, ranging from predicting stock prices to understanding gene regulatory networks. Regression aims to estimate the relationships between a dependent variable (outcome variable, or target) and a group of independent variables (predictors, or features). The most common type of regression is linear regression, where the relationship is typically in the form of a line (or a linear combination) that best approximates all the individual data points. On the other hand, in non-linear regression, observational data is modelled by a function that is a non-linear combination of the model parameters.

Linear regression

A linear model expresses the target output value in terms of a sum of weighted input variables that predict the target value given an input data instance.

Let $(\boldsymbol{x}, y)$ be a data instance, where $\boldsymbol{x} = (x_0, x_1, \ldots, x_n)$ is a vector of features representing the input data instance and y is the target output value. The predicted output will be of the form $\hat{y} = \hat{w}_0 x_0 + \hat{w}_1 x_1 + \cdots + \hat{w}_n x_n + \hat{b}$, where $\hat{\boldsymbol{w}} = (\hat{w}_0, \hat{w}_1, \ldots, \hat{w}_n)$ is a vector of feature weights (model coefficients) and $\hat{b}$ is a constant bias term (intercept). The goal of the linear regression algorithm is to estimate the model parameters $\hat{\boldsymbol{w}}$ and $\hat{b}$.

A common method to estimate the model parameters is the ordinary least squares technique. The aim of this technique is to minimise the difference (the mean squared error) between the predicted value and the actual value of the target variable. Formally, the objective is to minimise the sum of $(y - \hat{y})^2$ over all the data instances in a data set.

There are several extensions to the ordinary least squares technique, such as the least absolute shrinkage and selection operator (Lasso) (Tibshirani 1996) and ridge regression (Hoerl and Kennard 2000), which aim to control the model complexity.

Non-linear regression

In non-linear regression, the relationship between the feature vector $\boldsymbol{x}$ of a data instance and the target output y takes the form of an arbitrary function, $y = f(\boldsymbol{x}, \boldsymbol{\beta})$. The function f is non-linear in the components of the vector of parameters $\boldsymbol{\beta}$. Examples of non-linear functions include exponential functions, logarithmic functions, trigonometric functions and power functions.

Other regression methods that are non-linear include polynomial regression and k-NN regression. Polynomial regression is a form of regression in which the relationship between the input $\boldsymbol{x}$ and the output y is modelled as an nth-degree polynomial in $\boldsymbol{x}$. k-nearest neighbours (k-NN) is a nonparametric method used for classification and regression. In k-NN regression, the output is the property value for the object. This value is the average of the values of the k nearest neighbours (Altman 1992).

Classification

The goal of the classification methods in the supervised learning group is to classify observations into distinct categories, ie, the target value is a discrete class value. Furthermore, classification can be binary or multi-class. In binary classification, the target value can be 0 (negative class) or 1 (positive class), eg, email classification as spam or not-spam. On the other hand, in multi-class classification, the target value is one of a set of discrete values, eg, labelling the topic of a document based on its text.

Logistic regression

Logistic regression is a classification algorithm that produces the output as a binary decision, eg, "spam" or "not-spam". It is the simplest adaptation of linear regression to a specific case when the output variable is binary (0 or 1). Logistic regression is derived via a simple change to ordinary linear regression. We first form a linear combination of the input variables (as in conventional regression) and then apply a function that maps this number to a value between 0 and 1. The mapping function is called the logistic function.

k-nearest neighbours

The *k*-NN algorithm is a non-parametric method, proposed by Thomas Cover, used for classification and regression (Cover and Hart 1967). In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether *k*-NN is used for classification or regression. In *k*-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbours, with the object being assigned to the class that is most common among its *k* nearest neighbours (*k* is a positive integer, typically small).

Support-vector machines

Support-vector machines (SVMs) are one of the most popular classification algorithms. Their popularity stems from their ease of use and calibration. A support-vector machine constructs a hyperplane or set of hyperplanes in a high-dimensional space. The goal of an SVM is to separate the data hyperplane into non-overlapping parts. Intuitively, a good separation is achieved by the hyperplane that has the greatest distance to the nearest training-data point of any class (the so-called functional margin), since in general the larger the margin, the lower the generalisation error of the classifier (Hastie *et al* 2009).

Random forests

A random forest is a meta classifier that fits a number of decision-tree classifiers. A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences. A random forest classifier works by constructing a multitude of decision trees at training time on various subsamples of the data set. It uses averaging to improve the predictive accuracy and to control over-fitting. Thus, a random forest outputs the class that is the mode of the classes (the class that appears most often) of the individual trees (Ho 1995).

Neural networks

Neural networks are complex models that try to mimic the way the human brain develops classification rules. A neural network consists of many different layers of neurons, with each layer receiving inputs from previous layers and passing outputs to further layers. A neural network is composed of artificial neurons (conceptually

derived from biological neurons). Each artificial neuron has inputs and produces a single output that can be sent to multiple other neurons. The inputs can be the feature values of a sample of data, or they can be the outputs of other neurons. The outputs of the final output neurons of the neural net accomplish the classification task.

Evaluation

Evaluation metrics for regression

As regression tasks seek to predict a continuous-valued response, the output is some numeric value. Evaluating the performance of a regression algorithm is hence based on assessing how the predicted values deviate from the actual values of the target variable. Various metrics are typically used to evaluate the results of the prediction.

- Mean squared error (MSE) is the average of the squared difference between the target value and the value predicted by the regression model

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Root mean squared error (RMSE) is the square root of the averaged squared difference between the target value and the value predicted by the model

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- Mean absolute error (MAE) is the average of the absolute difference between the target value and the value predicted by the model

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- The R^2 or coefficient of determination is the proportion of the variance in the dependent variable that is predictable from the independent variables

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Table 4.1 Confusion matrix.

	Actual	
Predicted	**Positive (1)**	**Negative (0)**
Positive (1)	True positives (TP)	False positives (FP)
Negative (0)	False negatives (FN)	True negatives (TN)

is the mean of the observed data.

In the best case, the predicted values would exactly match the observed values; in this case, the MSE, RMSE and MAE results are 0 and R^2 result is 1.

Evaluation metrics for classification

Classification tasks seek to predict a discrete class value for a target variable. Evaluating the performance of a classification algorithm is hence based on assessing the extent to which the predicted classes match the actual ones over all the instances in the data set.

Classification evaluation metrics are mainly based on a confusion matrix, which consists of two dimensions (actual and predicted) and sets of classes in both dimensions. In this matrix, the columns represent actual classifications, and the rows represent predicted ones.

Table 4.1 shows how the confusion matrix looks for a binary classification, where there are two classes, labelled positive and negative. Several terms are associated with the confusion matrix. True positives (TP) are the cases where the actual and predicted classes are both positive. True negatives (TN) are the cases where the actual and predicted classes are both negative. False positives (FP) are the cases where the actual class is negative while the predicted one is positive. Conversely, false negatives (FN) are the cases where the actual class is positive while the predicted one is negative. True positives and true negatives are the cases that are correctly classified, whereas false positives and false negatives are those that are predicted incorrectly by the model. Using these terms, the evaluation metrics for classification are defined as follows:

- Accuracy is the fraction of the total number of cases that are correctly classified

$$A = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

- Precision is the fraction of all predicted positive cases that are correctly classified as positive

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Recall is the fraction of all actual positive cases that are correctly classified as positive

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- The F_1-score is the harmonic mean of recall and precision

$$F = \frac{2PR}{P + R}$$

For multi-class classification, the precision, recall and F_1-measure are calculated for each class. To combine the per-class scores into a single number, three methods are typically used.

- **Micro averaging:** first the values of TP, FN, TN and FP are summed over all instances, and then the performance measures are calculated using the accumulated values.
- **Macro averaging:** a simple arithmetic mean of per-class scores.
- **Weighted averaging:** similar to macro averaging, but the contribution of each class is weighted by the number of samples from that class.

UNSUPERVISED LEARNING

We have seen that supervised machine learning algorithms and techniques aim to develop models where the data has (previously known) labels, ie, the data has some target variables with specific values that are used to train the models (Bousquet *et al* 2004). However, when dealing with real-world problems, most of the time the data will not come with predefined labels. Therefore, there is a need to develop machine learning models that can classify data autonomously by finding commonality in the features. The main goal of unsupervised learning is to study the intrinsic structure of the data. The major applications of unsupervised learning include

- segmenting data sets by some shared attributes,
- detecting anomalies that do not fit into any group, and
- simplifying data sets by aggregating variables with similar attributes.

Clustering

The objective of clustering analysis is to find different groups within the data elements. To do this, clustering algorithms find a structure in the data so that elements of the same cluster (or group) are more similar to each other. Clustering algorithms have a wide range of applications, and are quite useful to solve real-world problems such as anomaly detection, recommending systems, document grouping or finding customers with common interests based on their purchases. Some of the most common clustering algorithms are the *K*-means, hierarchical clustering (agglomerative or divisive) and density-based spatial clustering of applications with noise (DBSCAN).

K-means

K-means clustering is a data mining technique used to group objects or data sets into clusters based on their similarities. The similarity is the total distance from the values in each cluster to the centroid, where each centroid has an average cluster value. The shorter the distance, the greater the similarity, and vice versa.

K-means clustering algorithm works as follows:

1. determine the number of clusters *K*;
2. choose *K* random points from the data as centroids;
3. set all the data points to the closest cluster centroid;
4. recalculate the centroid of newly formed clusters;
5. repeat until convergence, ie, the data points stop changing clusters.

Hierarchical clustering

Hierarchical clustering methods seek to build a hierarchy of clusters, using either an agglomerative strategy or a divisive strategy (Rokach and Maimon 2005). Agglomerative clustering is a bottom-up approach, where each observation starts in its own cluster, and

pairs of clusters are merged as we move up the hierarchy. Divisive clustering is a top-down approach, where all observations start in one cluster, and splits are performed recursively as we move down the hierarchy. The results of hierarchical clustering are usually presented in a dendrogram.

DBSCAN

Density-based spatial clustering of applications with noise is a density-based clustering that is able to find arbitrarily shaped clusters and clusters with noise, ie, outliers (Ester *et al* 1996). Given a set of points in some space, this algorithm groups together points that are closely packed together (points with many nearby neighbours), marking as outliers points that lie alone in low-density regions (whose nearest neighbours are farther away).

Association rule mining

Association rule mining (Agrawal *et al* 1993; Agrawal and Srikant 1994; Larose and Larose 2014) is used for discovering interesting relationships between variables in a large database. Association rules were first introduced for discovering regularities between products in large-scale transaction data recorded by point-of-sale systems in supermarkets (Agrawal *et al* 1993). Such rules can be used in supermarket basket analysis as the basis for decisions about marketing activities such as promotional pricing or product placements. Association rules can also be used in many application areas, including Web usage mining, intrusion detection and bioinformatics.

An association rule has the form $A \rightarrow B$, where A and B are disjoint sets of items (called the antecedent and the consequent of the rule, respectively). For example, $\{\text{milk}, \text{eggs}\} \rightarrow \{\text{bread}\}$ is an association rule that says that when milk and eggs are purchased, bread is likely to be purchased as well. Mining algorithms of association rules are based on various measures of significance and interest, such as support, confidence and lift (Geng and Hamilton 2006). Algorithms apply some constraints on such significance measures in order to select interesting rules from the set of all possible ones. The best-known constraints are minimum thresholds on support and confidence.

Generally, the association rule mining problem can be decomposed into two sub-problems. First, find all combinations of items

that have a certain statistical significance (frequent itemset mining). Second, given a significant itemset, generate all rules that have a certain strength.

Dimensionality reduction

Dimensionality reduction is the process of reducing the number of random variables (features, predictors) under consideration by obtaining a set of principal variables. Dimensionality reduction techniques are used for several reasons, including

- simplification of models to make them easier to interpret by researchers and users (James *et al* 2014),
- shorter training times,
- avoiding the curse of dimensionality (Bellman 1957), and
- enhanced generalisation by reducing overfitting (reduction of variance (James *et al* 2014)).

Data analysis such as regression or classification can be done in the reduced space more accurately than in the original space (Sulayes 2017).

Approaches to dimensionality reduction can be divided into feature selection (returns a subset of the features) and feature extraction (creates new features from functions of the original features). Feature selection is the process of selecting a subset of relevant features for use in model construction (Bousquet *et al* 2004; Blum and Langley 1997).

Feature extraction (also known as feature project and feature reduction), on the other hand, aims at transforming the data from a high-dimensional space to a space of fewer dimensions. Feature extraction starts from a set of initial features (measured data) and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalisation steps. The data transformation may be linear, as in principal component analysis (PCA), but many non-linear dimensionality reduction techniques also exist, such as Sammon's mapping (Sammon 1969), curvilinear component analysis (Demartines and Herault 1997) and kernel PCA (Schölkopf *et al* 1998).

The aim of PCA (Abdi and Williams 2010), also known as the Karhunen–Loeve transformation, is to perform a linear mapping

of the data to a lower-dimensional space in such a way that maximises the variance of the data in the low-dimensional representation. In other words, PCA reshapes the data along the directions of maximal variance. Simply speaking, PCA transforms data linearly into new properties that are not correlated with each other. Singular value decomposition is another factorisation method that transforms a matrix into special matrices that are easy to manipulate and to analyse.

Non-negative matrix factorisation (NMF) is a group of algorithms that factorise (decompose) a matrix into two matrices, with the property that all three matrices have no negative elements. This non-negativity makes the resulting matrices easier to inspect. NMF has many applications in astronomy, text mining and spectral data analysis (Berry *et al* 2007).

Evaluation

Evaluation metrics for clustering

The methods for evaluating the performance of a clustering algorithm are classified as either

- extrinsic, requiring ground truth labels, or
- intrinsic, not requiring ground truth labels.

Extrinsic measures are the most commonly used in clustering problems, and are based on comparisons between the output of the clustering algorithm and a gold standard usually built using human assessors. Extrinsic evaluation is based on determining the distance between both clustering solutions: the system output and the gold standard. Evaluation metrics can be grouped into four families (Amigó *et al* 2009; Meila and Heckerman 2001; Meila 2005), based on counting pairs, set matching, entropy and edit distance. Metrics that are based on set matching share the feature of assuming a one-to-one mapping between clusters and categories, and they rely on the precision and recall concepts inherited from information retrieval (Zaki and Meira 2014).

- Purity (Zhao and Karypis 2001; Manning *et al* 2008) is a measure that quantifies the extent to which a cluster contains entities from only one partition, ie, it measures how "pure" each cluster is.

- The precision, recall and F_1-measure metrics typically used for classification evaluation can also be used to evaluate the performance of clustering algorithms.
- Normalised mutual information is a measure of the mutual dependence between the system clustering and the ground truth based on the shared object membership, with a scaling factor corresponding to the number of objects in the respective clusters.

In intrinsic evaluation, the aim is to identify sets of clusters that are compact, with a small variance between members of the cluster, and well separated (where the means of different clusters are sufficiently far apart) compared with the within cluster variance. Intrinsic measures include the following.

- The Davies–Bouldin index (DBI) is a metric for evaluating clustering algorithms, where the validation of the clustering is based on quantities and features inherent to the data set, such as the scatter of points within the cluster, and the separation between different clusters (Davies and Bouldin 1979). Thus, it captures the intuition that clusters which are well-spaced from each other and are themselves very dense are likely to be "good". As the DBI shrinks, the clustering is considered to become "better".
- The Dunn index captures the same idea as the DB index, as it improves when clusters are dense and far apart from each other. But the Dunn index increases as performance improves (Dunn 1974). However, while the DBI considers the dispersion and separation of all clusters, the Dunn index only considers the worst cases in the clustering: the clusters that are closest together and the single most dispersed cluster.
- Silhouette is a method of validation of consistency within clusters (Rousseeuw 1987). The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared with other clusters (separation). A high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters. The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance.

In addition to extrinsic and intrinsic evaluation metrics, there are relative evaluation metrics that are used to compare two clusterings,

such as the Rand index and adjusted Rand index. The Rand index is a measure of the similarity between two data clusterings, which is similar to the accuracy metric of classification evaluation.

Evaluation metrics for association rules

Typically, the evaluation of association rules mining is not in terms of the performance of the mining algorithm, but rather in terms of the quality (interestingness) of the discovered rules.

Various measures are commonly used to assess the significance and interest of association rules (Geng and Hamilton 2006). For a given association rule $A \rightarrow B$, the interest measures include the following.

- "Support" is an indication of how frequently the rule occurs in the database, and defined as the proportion of transactions in which the itemsets A and B appear together

$$\text{supp}(A \rightarrow B) = P(A \cup B)$$

- "Confidence" is defined as the proportion of the transactions containing A that also contain B

$$\text{conf}(A \rightarrow B) = P(B \mid A) = \frac{\text{supp}(A \cup B)}{\text{supp}(A)}$$

- "Lift" is the ratio of the observed support to that expected if A and B were independent

$$\text{lift}(A \rightarrow B) = \frac{P(B \mid A)}{P(B)}$$

- "Leverage" is a symmetric measure expressing the difference between the actual probability of $A \cup B$ occurring in a transaction and the probability when A and B are statistically independent

$$\text{leverage}(A \rightarrow B) = \text{supp}(A \cup B) - \text{supp}(A)\,\text{supp}(B)$$

CASE STUDY

On July 18, 2012, the McDonalds fast-food chain started a social media campaign using the hashtag "#McDstories" to emphasise their product quality. Within hours, thousands of people turned to Twitter and used this hashtag to share their negative stories about McDonalds (Lubin 2012). A seemingly arbitrarily occurring outrage

towards people, companies, media campaigns or politicians is called an online firestorm (Pfeffer *et al* 2014). These online firestorms have the potential to seriously affect a company's reputation or stock market value. Consequently, early detection of these negative word-of-mouth events is of high significance. We will use some of the above-mentioned approaches to exemplify how machine learning methods can be used on large-scale Web-based data to better understand the dynamics in the data.

Data

To study the dynamics of this Twitter firestorm, we use historical data from the 10% sample application programming interface (API) data of Twitter: a random 10% of all tweets over a three week period around the time of the incident. While Twitter's data samples are viewed critically in academia (Pfeffer *et al* 2018), they are widely used as a data source by social media teams of companies, business consultants and government entities for real-time analysis of public opinion (Hong and Nadler 2011; Younus *et al* 2011; Cody *et al* 2016). For the purpose of this case study, we extracted 110,898 tweets including the term "mcdonalds" (case insensitive).

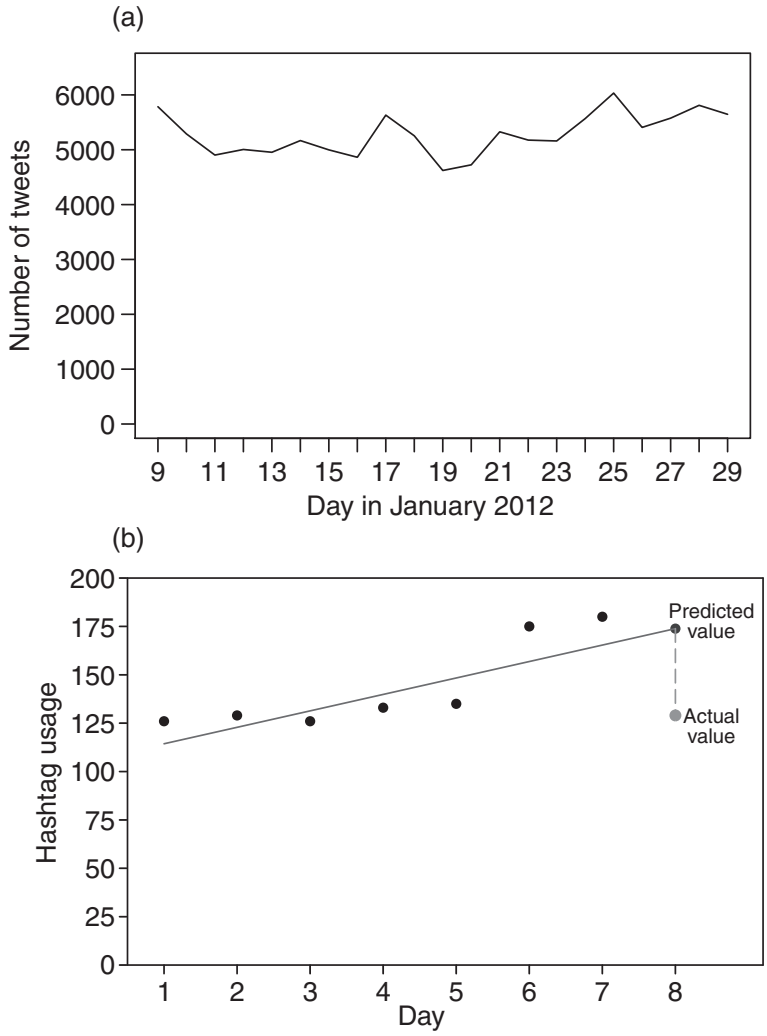
Change detection

In a first step we will identify whether a firestorm is going on. Figure 4.1(a) shows the time series plot of the overall number of tweets per day from our data source. This figure does not show suspicious changes caused by the firestorm (starting on January 18). In order to detect a change in the data, we need to apply methods that investigate the content of the tweets and that can be employed for real-time analysis. For the purpose of this case study we analysed data on a daily basis. Adapting the approaches to an hourly analysis or, in the case of more data, to a more granular temporal level is easily possible.

In order to systematically detect suspicious changes caused by the firestorm, we observe the usage of Twitter entities, such as hashtags, on a daily basis. The goal is to examine whether there is any deviation between the expected usage and the actual usage of an entity. When a significant deviation is observed, this is an indication that something unusual is going on which will need further inspection.

First, we use a supervised machine learning approach, namely linear regression, in order to predict the usage of a hashtag on a

Figure 4.1 Tweets per day and an example for regression.

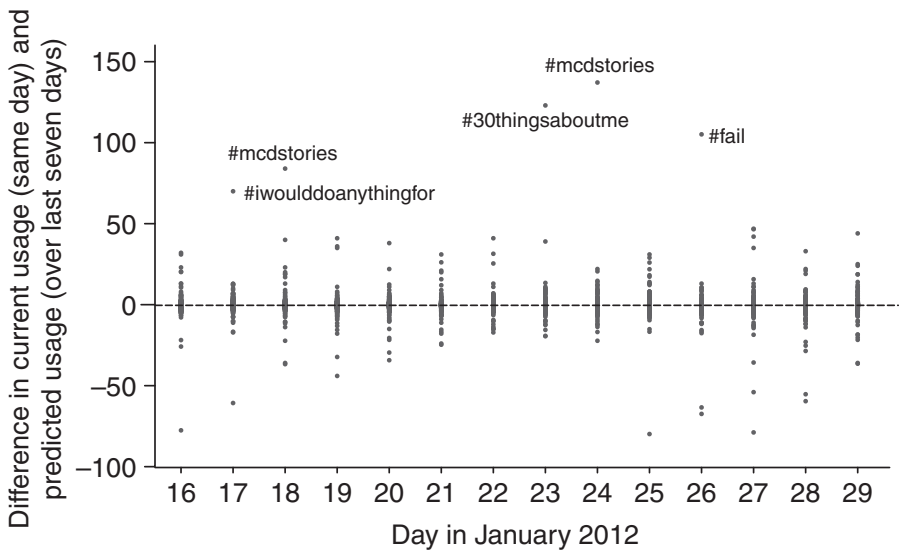


Part (a) shows the number of tweets per day in January 2012 that include the term "mcdonalds" in the 10% sample data source from Twitter. In part (b), using linear regression, we can predict the usage of a hashtag on a given day by extrapolating the usage over the last seven days. The dashed grey line shows the deviation.

given day d by extrapolating its usage over the previous seven days: $[d-7, d-1]$. As shown in Figure 4.1(b), we compare the predicated value of usage with the actual value to obtain the deviation. In this case, we can see a general upwards trend that is captured by the seven-day regression model. The actual data point on day eight deviates negatively from this trend.

We then repeated the operation of predicting the next data point with linear regression models for each hashtag that was used at least five times on every day (starting from January 16, since we need seven days of previous usage). The results are shown in Figure 4.2.

Figure 4.2 Deviation of actual hashtag usage from expected usage, on daily basis.



The hashtag "#mcdstories" appears on January 18 and January 24.

We observe that most hashtags form a cluster centred around zero, where the deviation, either positive or negative, is not significant. However, several outliers to that cluster can be observed on different days, which indicates significant positive or negative deviations. While a negative deviation of a hashtag indicates its decay, a significant positive deviation indicates a rise in new trending hashtags. In particular, we observe that our hashtag of interest, "#mcdstories", appears way above expectation on January 18, and then reappears again on January 24, followed by "#fail", which was used to describe McDonalds' response to the online firestorm.

Understanding what is going on

In order to understand what is going on, we extracted a subset of the tweets that contain the hashtag #mcdstories and similar terms, such as #mcdonaldstories. On this subset, for each tweet (text) as a document, we applied a preprocessing pipeline, typical for text analysis methods, including tokenisation, lower-case conversion and stop-words removal. Hence, each document became a list of tokens (terms). Then, we applied the unsupervised machine learning approach described above, namely association rules mining, using

Table 4.2 Association rules.

Rule	Support	Confidence	Lift
{“backfires”} → {“#mcdstories”}	0.142	1.0	1.021
{“horror”} → {“#mcdstories”}	0.107	1.0	1.021

the Apriori algorithm (Agrawal and Srikant 1994). A few of the interesting association rules among the many that we can identify in this data set are shown in Table 4.2.

Many tweets associated the hashtag “#mcdstories” with keywords such as “horror” and “backfires”. Examples of such tweets are the following:

- “McDonalds’ Twitter promotion fail: Users hijack #McDstories hashtag to share fast food horror stories”;
- “The @McDonalds social media campaign backfired. Now people are using #McDstories to share McDonalds horror stories”.

Text classification task

In the previous step, we identified possible story lines about tweets related to specific topics. Generalising this idea to identify newly emerging topics with negative stories about the brand brings us to topic modelling and text classification methods. Using latent Dirichlet allocation, the standard topic modelling approach, we could identify topics in the tweets about McDonald’s. Here, we focus on a classical machine learning challenge, namely classification. The underlying idea is to have a set of tweets precoded as being positive or negative towards the brand or unrelated to the brand. With these codes, a machine learning model is trained and applied to newly incoming tweets in order to automatically classify them.

For training and testing purposes, we use only tweets related to the #McDStories hashtag; we establish a ground truth by the manual labelling of tweets as good, bad or unrelated. The allocation to those classes is as follows: bad (10%), good (52%), unrelated (38%). Given the tweet text as a document, after preprocessing (tokenisation, etc) and computing of TF–IDF scores, each data entry is a vector of TF–IDF scores.

Then, we split the annotated tweets into two subsets: training data (80%) to train the classifier, and test data (the remaining 20%) to

Table 4.3 Classification results.

	Accuracy	Precision	Recall	F_1-score
Random forest	0.85	0.77	0.85	0.80
Support-vector machine	0.53	0.28	0.53	0.37
Logistic regression	**0.89**	0.90	0.89	**0.86**

Notes: Boldface denotes the highest values (see text).

evaluate the accuracy of the classifier. For comparison, we use three classification approaches: random forests, SVMs and logistic regression. Each of those classifiers is trained using the training data set and evaluated using the test data set.

The classification results are shown in Table 4.3, where the evaluation metrics used are accuracy, recall, precision and F_1-score (see p. 55).

We find that the random forest classifier has a reasonable accuracy (with 85% of instances being correctly classified). It also has a good enough recall and precision; hence the F_1-score is good (0.80). In contrast, the SVM classifier has a moderate accuracy (53% correctly classified instances). This classifier has a moderate recall but a low precision, hence the F_1-score is low too (0.37). In fact, in this case, SVM classifier performed badly in identifying the bad and unrelated classes. Hence, the precision and recall for those classes were very low, which is why the overall performance of this classifier was not good. Finally, the logistic regression outperforms the other classifiers, with an accuracy of 89% and an F_1-score of 0.86.

DISCUSSION

In the previous section, we showcased and discussed the machine learning methods used to analyse data characteristics from millions of tweets. The use of large-scale Web data can be a very good complement to gain insight into real-time dynamics, ie, knowledge and information for anticipating and controlling processes. The combination of traditional data sets (data from information service providers, financial reports, institutions, etc) and big data can be an information advantage here.

While these methods are relatively easy to use, some of them are algorithmically very complex and almost impossible to comprehend in detail for researchers from most fields. This leads to the

biggest issue related to computational methods: researchers deploying methods without considering their limitations or preconditions for the data.

Humans and machines: possible pitfalls of big data and machine learning

As Kolanovic and Krishnamachari (2017) discussed, the methods provided cannot entirely replace human intuition. Machine learning models can, if not properly guided, overfit or uncover spurious relationships and patterns. Data scientists who lack subject matter expertise may not achieve the desired investment results. When using big data, it is still necessary to understand the economics behind the data and signals. The role of humans and machines is twofold: machines have the ability to rapidly collect and analyse news feeds and tweets, scrape websites and trade on these continuously, but they are unlikely to be able to compete with strong macroanalysis and the refined intuition of human investors (Kolanovic and Krishnamachari 2017). Regarding the validity of data sets, biases and inaccuracies not only occur at the source of the data, but also are introduced during processing. The rigour with which these issues are addressed by different researchers is known to vary widely. In practice, a variety of dangers regarding social media data have been identified and studied (Pfeffer *et al* 2018; Olteanu *et al* 2019). As Lazer *et al* (2014) demonstrate, research on whether search data or social media "can predict *x*" has become commonplace and is often presented in sharp contrast with traditional methods and hypotheses. Although these studies have shown the value of such data, it is far from supplanting more traditional methods or theories. Perspectives on these challenges address the scientific infrastructure supporting data sharing, data management, informatics and statistical methodology. Research ethics and policy are discussed in the literature, and suggestions on how to tackle these challenges need to be discussed as well (Lazer *et al* 2020; King 2011; Vespignani 2009).

Because of these challenges, we recommend decision makers are sensitive to the following issues when applying machine learning methods, especially when using social media big data as alternative data sets (Ruths and Pfeffer 2014; Olteanu *et al* 2019).

- In designing your method, consider carefully what representation to use for your data, which algorithm to use, how to optimise it and how to evaluate its performance.
- Test the validity of both internal and external data.
- Data might lack quality due to sparsity, noise or bias effects.
- Population biases may affect the representativeness of a data sample.
- Data acquisition involves a query specifying a set of criteria for selecting, ranking and returning the data being requested, but different APIs may support different types of queries.
- Data filtering entails the removal of irrelevant portions of the data; sometimes this cannot be done during data acquisition due to the limited expressiveness of an API or query language.
- Biases introduced by data processing operations such as cleaning, enrichment and aggregation are likely to compromise the internal validity.

Therefore, we encourage decision makers to combine data science expertise with high levels of competence in the given field (eg, economics, behavioural economics, statistics, methodology). Applied properly, machine learning methods have the ability to complement established techniques and, when applied to very large data sets, have the potential to capture complex facts and dynamics.

REFERENCES

Abdi, H., and L. J. Williams, 2010, "Principal Component Analysis", *WIREs Computational Statistics* 2(4), pp. 433–59.

Agrawal, R., T. Imieliński, and A. Swami, 1993, "Mining Association Rules between Sets of Items in Large Databases", in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pp. 207–16 (New York, NY: Association for Computing Machinery).

Agrawal, R., and R. Srikant, 1994, "Fast Algorithms for Mining Association Rules in Large Databases", in *Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487–99 (San Francisco, CA: Morgan Kaufmann).

Altman, N. S., 1992, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression", *American Statistician* 46(3), pp. 175–85.

Amigó, E., J. Gonzalo, J. Artiles and F. Verdejo, 2009, "A Comparison of Extrinsic Clustering Evaluation Metrics Based on Formal Constraints", *Information Retrieval* 12(4), pp. 461–86.

Bańbura, M., D. Giannone, M. Modugno and L. Reichlin, 2013, "Now-Casting and the Real-Time Data Flow", in *Handbook of Economic Forecasting*, pp. 195–237 (Elsevier).

Bellman, R., 1957, *Dynamic Programming*, Rand Corporation Research Study (Princeton University Press).

Berry, M. W., M. Browne, A. N. Langville, V. P. Pauca and R. J. Plemmons, 2007, "Algorithms and Applications for Approximate Nonnegative Matrix Factorization", *Computational Statistics and Data Analysis* 52(1), pp. 155–73.

Blum, A. L., and P. Langley, 1997, "Selection of Relevant Features and Examples in Machine Learning", *Artificial Intelligence* 97(1), pp. 245–71.

Bollen, J., H. Mao and X.-J. Zeng, 2011, "Twitter Mood Predicts the Stock Market", *Journal of Computational Science* 2(1), pp. 1–8.

Bousquet, O., U. von Luxburg and G. Ratsch, 2004, *Advanced Lectures on Machine Learning: ML Summer Schools 2003* (Berlin: Springer).

Ciulla, F., D. Mocanu, A. Baronchelli, B. Gonçalves, N. Perra and A. Vespignani, 2012, "Beating the News Using Social Media: The Case Study of American Idol", *EPJ Data Science* 1, pp. 1–11.

Cody, E. M., A. J. Reagan, P. S. Dodds and C. M. Danforth, 2016, "Public Opinion Polling with Twitter", e-print, arXiv:1608.02024 [physics.soc-ph].

Cook, S., C. Conrad, A. L. Fowlkes and M. H. Mohebbi, 2011, "Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic", PLoS ONE 6(8), e23610.

Cover, T. M., and P. E. Hart, 1967, "Nearest Neighbor Pattern Classification" *IEEE Transactions on Information Theory* 13, pp. 21–7.

Davies, D. L., and D. W. Bouldin, 1979, "A Cluster Separation Measure", *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1(2), pp. 224–7.

Demartines, P., and J. Herault, 1997, "Curvilinear Component Analysis: A Self-Organizing Neural Network for Nonlinear Mapping of Data Sets", *Transactions on Neural Networks* 8(1), pp. 148–54.

Domingos, P., 2012, "A Few Useful Things to Know about Machine Learning", *Communications of the ACM* 55(10), pp. 78–87.

Dunn, J. C., 1974, "Well-Separated Clusters and Optimal Fuzzy Partitions", *Journal of Cybernetics* 4(1), pp. 95–104.

Engelberg, J. E., and C. A. Parsons, 2011, "The Causal Impact of Media in Financial Markets", *Journal of Finance* 66(1), pp. 67–97.

Ester, M., H.-P. Kriegel, J. Sander and X. Xu, 1996, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–31 (Palo Alto, CA: AAAI Press).

Geng, L., and H. J. Hamilton, 2006, "Interestingness Measures for Data Mining: A Survey", *ACM Computing Surveys* 38(3), 9-es.

Gentzkow, M., B. Kelly and M. Taddy, 2019, "Text as Data", *Journal of Economic Literature* 57, pp. 535–74.

Gentzkow, M., and J. M. Shapiro, 2010, What Drives Media Slant? Evidence from US Daily Newspapers", *Econometrica* 78(1), pp. 35–71.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski and L. Brilliant, 2009, "Detecting Influenza Epidemics Using Search Engine Query Data", *Nature* 457(7232), pp. 1012–14.

Goel, S., J. M. Hofman, S. Lahaie, D. M. Pennock and D. J. Watts, 2010, Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences* 107(41), pp. 17486–90.

Hastie, T., R. Tibshirani and J. H. Friedman, 2009, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer Series in Statistics (Berlin: Springer).

Ho, T. K., 1995, "Random Decision Forests", in *Proceedings of the Third International Conference on Document Analysis and Recognition*, Volume 1, pp. 278–82 (Hoboken, NJ: IEEE Press).

Hoerl, A. E., and R. W. Kennard, 2000, "Ridge Regression: Biased Estimation for Nonorthogonal Problems", *Technometrics* 42(1), pp. 80–86.

Hong, S., and D. Nadler, 2011, "Does the Early Bird Move the Polls? The Use of the Social Media Tool 'Twitter' by US Politicians and Its Impact on Public Opinion", in *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, pp. 182—6 (New York, NY: Association for Computing Machinery).

Huang, A., 2008, "Similarity Measures for Text Document Clustering", in *Proceedings of the Sixth New Zealand Computer Science Research Student Conference*, pp. 49–56 (Wellington: New Zealand Computer Society).

James, G., D. Witten, T. Hastie and Tibshirani, R., 2014, *An Introduction to Statistical Learning* (Berlin: Springer).

King, G., 2011, "Ensuring the Data-Rich Future of the Social Sciences", *Science* 331(6018), pp. 719–21.

Kolanovic, M., and R. T. Krishnamachari, 2017, "Big Data and AI Strategies, Machine Learning and Alternative Data Approach to Investing", Technical Report, JP Morgan.

Laney, D., 2001, "3D Data Management: Controlling Data Volume, Velocity, and Variety", Blog Post, February 21, Application Delivery Strategies.

Larose, D. T., and C. D. Larose, 2014, *Discovering Knowledge in Data: An Introduction to Data Mining*, Second Edition (Hoboken, NJ: John Wiley & Sons).

Lazer, D., R. Kennedy, G. King and A. Vespignani, 2014, "The Parable of Google Flu: Traps in Big Data Analysis", *Science* 343(6176), pp. 1203–5.

Lazer, D., A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy and M. V. Alstyne, 2009, "Computational Social Science", *Science* 323(5915), pp. 721–3.

Lazer, D. M. J., A. Pentland, D. J. Watts, S. Aral, S. Athey, N. Contractor, D. Freelon, S. Gonzalez-Bailon, G. King, H. Margetts, A. Nelson, M. J. Salganik, M. Strohmaier, A. Vespignani and C. Wagner, 2020, "Computational Social Science: Obstacles and Opportunities", *Science* 369(6507), pp. 1060–62.

Lubin, G., 2012, "McDonald's Twitter Campaign Goes Horribly Wrong #McDStories", Blog Post, January 24, Business Insider.

Manning, C. D., P. Raghavan and H. Schütze, 2008, *Introduction to Information Retrieval* (Cambridge University Press).

Mao, H., S. Counts and J. Bollen, 2011, "Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data", e-print, arXiv:1112.1051 [physics, q-fin].

Meila, M., 2005, "Comparing Clusterings", in *Proceedings of the 22nd International Conference on Machine Learning*, pp. 577–84 (New York, NY: Association for Computing Machinery).

Meila, M., and D. Heckerman, 2001, "An Experimental Comparison of Model-Based Clustering Methods", *Machine Learning* 42, pp. 9–29.

Moon, S. J., and P. Hadley, 2014, "Routinizing a New Technology in the Newsroom: Twitter as a News Source in Mainstream Media", *Journal of Broadcasting and Electronic Media* 58(2), pp. 289–305.

Olteanu, A., C. Castillo, F. Diaz and, E. Kıcıman, 2019, "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries", *Frontiers in Big Data* 2, Article 13.

Pfeffer, J., K. Mayer and F. Morstatter, 2018, "Tampering with Twitter's Sample API", *EPJ Data Science* 7, Article 50, pp. 1–21.

Pfeffer, J., T. Zorbach and K. M.Carley, 2014, "Understanding Online Firestorms: Negative Word-Of-Mouth Dynamics in Social Media Networks. *Journal of Marketing Communications* 20(1), pp. 117–28.

Rokach, L., and O. Maimon, 2005, "Clustering Methods", in O. Maimon and L. Rokach (eds), *Data Mining and Knowledge Discovery Handbook*, pp. 321–352 (Boston, MA: Springer).

Rousseeuw, P. J., 1987, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis", *Journal of Computational and Applied Mathematics* 20, pp. 53–65.

Ruths, D., and J. Pfeffer, 2014, "Social Media for Large Studies of Behavior", *Science* 346(6213), pp. 1063–64.

Sammon, J. W., 1969, "A Nonlinear Mapping for Data Structure Analysis", *IEEE Transactions on Computers* 18(5), pp. 401–9.

Schölkopf, B., A. Smola and K.-R. Müller, 1998, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem", *Neural Computation* 10(5), pp. 1299–1319.

Sulayes, A. R., 2017, "Reducing Vector Space Dimensionality in Automatic Classification for Authorship Attribution", *Revista Ingeniería Electrónica, Automática y Comunicaciones* 38(3), pp. 26–35.

Tibshirani, R., 1996, "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society (Series B)* 58, pp. 267–88.

Vermeer, S. A. M., T. Araujo, S. F. Bernritter and G. van Noort, 2019, "Seeing the Wood for the Trees: How Machine Learning Can Help Firms in Identifying Relevant Electronic Word-of-Mouth in Social Media", *International Journal of Research in Marketing* 36(3), pp. 492–508.

Vespignani, A., 2009, "Predicting the Behavior of Techno-Social Systems", *Science* 325, pp. 425–8.

Younus, A., M. A. Qureshi, F. F. Asar, M. Azam, M. Saeed and N. Touheed, 2011, "What Do the Average Twitterers Say: A Twitter Model for Public Opinion Analysis in the Face of Major Political Events", in *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining*, pp. 618–23 (Hoboken, NJ: IEEE Press).

Zaki, M. J., and W. Meira, Jr, 2014, *Data Mining and Analysis: Fundamental Concepts and Algorithms* (Cambridge University Press).

Zhao, Y., and G. Karypis, 2001, "Criterion Functions for Document Clustering: Experiments and Analysis", Technical Report 01-40, Army HPC Research Center, Minneapolis, MN.

5.2.2 QualiAnon – The Qualiservice Tool for Anonymizing Text Data

Authors Wienke Strathern, Moritz Issig, Kati Mozygemba, Jürgen Pfeffer

In Technical Report, TUM-I2087, Technical University of Munich, Department of Informatics, 2020, https://mediatum.ub.tum.de/1537509?show_id=1575928

Abstract

The anonymization of qualitative interview data is of high importance. For secondary use of data, anonymized data is essential. However, anonymization procedures are complex and time-consuming. Due to issues with automated processes and a lack of control that did not allow researchers to use earlier tool versions outside the RDC, we decided to provide a tool that keeps researchers in control of their data. Automated decisions give all-in-one solutions, but studying qualitative interview data depends on the needs of every single researcher. We provide a tool that enables researchers to make individual decisions with the information needed on the level required. This report proposes a solution to anonymize qualitative interview data to create its coding schemes and individual abstraction levels. We built a tool that assists in working with textual interview data. By using the tool, processes can be optimized, and important information can be obtained at the same time.

Publication Summary

Approach: In Strathern et al. (2020a) we present the development of a tool that allows for anonymizing interview data and a manual on how to apply the tool. Here, we present the technical report without the manual. In collaboration with Qualiservice, we identified the essential information from interview data for the tool's development. **Methods:** We then created a classification scheme that enables to maintain the desired information while safeguarding sensitive participant information. The critical factors are the type of anonymization and the level of abstraction used.

name	color	labelsub	subcategories	labelA	examples
Person	Red	Gender	Male / Female / Diverse	Role	Partent / Sister / Brother / Friend / Neighbor / Colleague / Contract / Student
Location	Light Green	Type	Country / Fed. Country / Region / City / Street	Size	Big / Medium / Small
Institutions	Light Blue	Type	Companies / Schools / Theater / Association / Party / Prison / Orphanage	Activity	Further Training / Hobby / Sport / Education
Profession	Light Grey	Sector	Trade / Services / Industry / Craft / Science / research / Administration	Work	Assistants and Trainees / Specialist Activities / Complex Specialist Activities / Highly Complex Activities
Personal Circumstances	Pink	Type	Accident / Death / Alcoholism / Nursing / Awards / Disease	Description	
Time	Yellow	Type	Date / Time Span	Time	Date 2009-2019/ Time Span 0-3 Years / 4-6 Years / 7-10 Years
Education	Purple	Type	University / University of Applied Sciences / High School / Secondary School / Company	Description	Academic Degree / Vocational Training / School Degree / No Degree
Other	Dark Grey	Type	Religion / Politics / Sexuality	Description	

Figure 4: Structure of Categories as developed in Strathern et al. (2020a)

To anonymize research data, we can use pseudonyms, aggregation, or replace sensitive information with relevant information to the social sciences. Figure 4 provides an overview of the tool's core element, outlining the nine default categories for anonymization, including Person, Location, Institution, Profession, Personal Circumstances, Time, Education, and Other. Standardized lists can also be used to replace places, diseases, and professions to provide the researcher with standardized information. The degree of anonymization can be adjusted by using pseudonyms, aggregating information by applying classes, or replacing text by adding descriptions or attributions. The tool was developed using Java.

Author Contribution

Wienke Strathern headed the project, developed the conceptual framework, wrote the introduction and conclusion, built the classification schema, did a literature review and worked out the theoretical background, revision and editing, and coordinated the team.

QualiAnon - The Qualiservice tool for anonymizing text data

Wienke Strathern, Moritz Issig, Kati Mozygemba, Jürgen Pfeffer

September 2020

Abstract

The anonymization of qualitative interview data is of high importance. For the purpose of secondary use of data, anonymized data is essential. However, anonymization procedures are complex and time consuming. This is why the Research Data Center Qualiservice at the University of Bremen aimed at providing a tool to support the anonymization of textbased research data. In cooperation with the RDC Qualiservice teh working group of Jürgen Pfeffer at the Bavarian School of Public Policy at the Technical University of Munich developed the basis for the Qualiservice Anonymization Tool - QualiAnon by starting to technically implement the Qualiservice Anonymization Concept [4][6]. Due to issues with automated processes and a lack of control that did not allow to use earlier tool versions outside the RDC, we decided to provide a tool that keeps researcher in control of their data. Automated decisions give all-in-one solutions, but studying qualitative interview data depends on the needs of every single researcher. We provide a tool that enables researcher to make individual decisions with the information needed, on the level required. In this report, we propose a solution to anonymize qualitative interview data with the purpose to create own coding schemes and individual abstraction levels. We built a tool that assists in working with textual interview data. By using the tool, processes can be optimized and important information can be obtained at the same time.

The basic elements of the Qualiservice Anonymization Tool has been further developed by the world data archive PANGAEA, cooperation partner of the Qualiservice consortium. Qualiservice currently validates the tool in different use cases. The release of the tool is scheduled for spring 2021.

The work on QualiAnon was funded by the German Research Foundation between 2018 and 2021 (project HO 2120/9-1 QualiService: Implementation of a nationwide archive and data service center for qualitative social science interview data; Head of project: Prof. Dr. Betina Hollstein)

Index Terms: anonymization, qualitative social science interview data, text analysis

Contents

Chapter 1

Technical Report

1.1 Motivation

The database of the Research data Center Qualiservice is qualitative social science data. These are currently available mostly as transcripts. In order to make this data accessible for secondary research, the personal and person-related data must be anonymized. The process of anonymizing qualitative data is costly and complex. In order to facilitate the work of primary researchers, the Qualiservice anonymisation tool - QualiAnon - was developed to implement Qualiservice Anonymization Concept [4] [6] and to allow for a a feasible anonymization process. Therefore, the original documents are available to secondary users in an anonymized version. The anonymization process refers to flexible rules that meet the respective requirements of the secondary researcher. One objective is to control and implement the application of the rules to the original documents and the construction of the documents to be used for further research. Furthermore, the tool should facilitate the anonymization of qualitative research data (mainly transcripts) by providing protective measures for the anonymisation of qualitative interview data - so that this data can be used for secondary purposes. In accordance with the Qualiservice Anonymization Concept the data should remain researchable (information relevant to the social sciences, such as the size of a site or the general concept of a disease) or be reopened for specific research questions, that is "flexible anonymization" [4]. [1]

1.1.1 Objectives

One of the objective is to develop a tool that intuitively assists in text editing processes without making automated decisions. Researchers are in control to determine the objects and the degree of anonymization. Information can be extracted and changed by their meaning for the researcher. Following the Qualiservice Anonymization Concept the framework should allow for an appropriate level of anonymity whilst trying to maintain maximum meaningful information in the research data.

[1]The work on QualiAnon was funded by the German Research Foundation between 2018 and 2021 (project HO 2120/9-1 QualiService: Implementation of a nationwide archive and data service center for qualitative social science interview data; Head of project: Prof. Dr. Betina Hollstein)

1.1.2 Contributions

- We build a tool with an easy to use interface that assists in the anonymization process of interview data.
- We provide a framework with which entities from text can be extracted.

1.1.3 Outline

This work is structured as follows. In a first step information about the secondary use of interview data is provided. For the purpose of secondary use anonymization characteristics will be defined following the Qualiservice Anonymization Concept. To approach different levels of abstraction during anonymization we refer methodologically to information extraction and provide our framework. This is followed by the code documentation. The technical report ends with some closing remarks. Chapter 2 contains the tool manual in which all steps containing the basic version are shown in detail with screenshots and example text. We have selected three different interview types to illustrate the functions of the tool step by step.

1.2 Background

1.2.1 Secondary Use of Interview Data

The concept of qualitative longitudinal data archives to conduct qualitative secondary analysis has been broadley discussed [8] [4]. Referring to the Qualiservice Anonymization Concepts as well as to the conceptual elaboration from a project of the University of Leeds [1] it is important to reach an appropriate level of anonymity, whilst trying to maintain maximum meaningful information in the research data. Information should not be crudely removed or blanked-out, but rather pseudonyms, replacement terms or vaguer descriptors should be used. Some data that combine many difficult features: geographically specific references, sensitive and potentially harmful content, longitudinal detail that increase disclosiveness, will be difficult or impossible to anonymize in a manner that both protects the quality of the data and the confidentiality of participants. Other strategies will be necessary for such data, for example, the anonymization of a small subset of data for illustrative purposes and might be highly valuable for methodological insights. It can be summarized that the objective for all data is to achieve a reasonable level of anonymization which is then combined with other strategies, namely consent agreements and access controls, in order to maintain confidentiality.

1.2.2 Anonymization of Interview Data

1.2.2.1 Objects of Anonymization

Objects of anonymization are personal and person-related features, e.g., personal names, place names, street names, federal states, institutions and organizations (e.g. companies, schools, institutes), professions, titles and educational qualifications, age, times/calendar dates, pictures and voices. Furthermore, indirect, but specific contextual information. Characteristics of the subjects as well as those of third parties mentioned in the interviews (also the personal rights of the interviewees, transcribers, etc. must be taken

into account). Sensitive information are information on ethnic origin, political opinion, religious or philosophical beliefs, trade union membership, health or sex life.

1.2.2.2 Degree of Anonymization

The Qualiservice Anonymization Concept works with different levels of abstraction. The degree of anonymization refers to the level of abstraction. The term **Pseudonym** refers to renaming people, objects etc. A third way of making research data anonymous is **aggregation**. That means to coarsen or aggregate information by creating classes or categories. This includes, for example, replacing the concrete age with age classes, replacing the concrete employer with the industry or company size class, and replacing a specific girl's name with student. Another way to make the information anonymous is to **replace** it with information that includes the meaning for and relationship with the researcher. Following the Qualiservice Anonymization Concept, people can be described by the meaning the information to be replaced has for the interviewer, for example, girlfriend, mother, teacher. Place names can be paraphrased based on the meaning of the place to the subject, such as birthplace, place of residence, place of work. On ecentral aspect of the Qualiservice Anonymization Concept is to replace a sensible information with information relevant to the social sciences. So you could replace a country's name and add information such as country with high youth unemployment or welfare state.

1.3 Framework

1.3.1 Information extraction

To approach different levels of abstraction regarding interview texts, we ask the following questions: Which categories of information, parts of a text are needed to understand content, intention, associations, relations that can be replaced, thus anonymized. Therefore, we distinguish the level of abstraction by the degree of information that can easily still be obtained. To obtain the required information coding schemes with categories can be applied. Building a taxonomy of quantitative text analysis techniques is based on two types, instrumental schemes categorize tokens in a text corpus according to theoretical or conceptually-driven frameworks (sentiment analysis based on affect control theory or narrative analyses based on story grammars). Representational schemes, on the other hand, are data-driven schemes (open schemes) that categorize tokens according to simple semantic relations such as synonymy, meronymy, hypernomy, or hyponymy [5]. To build coding schemes for text, we ask the five W's of journalism - the who, what, where, when, why, and how of things. This kind of content analysis requires entity extraction and ontologically categorized entities [7][2][3]. The applying coding scheme refers to central replacement categories of the Qualiservice Anonymization Concept.

1.3.2 Implementation

For the anonymization of objects in interview data we built a **category** scheme as mentioned in the framework. This is the core element of this tool. These nine categories are intended as default categories for anonymization: Person, location, Institution, Profession, Personal Circumstances, Time, Education, Other. Standardised lists, developed by Qualiservice [6], for the replacements of places, diseases and professions can be applied.

Thus, the researcher has the possibility to use standardized information. The **degree of anonymization** refers to the labels in the category scheme: pseudonym, aggregation of information by applying classes or replacing text by adding descriptions or attributions. A category refers to the overall entity. A pseudonym replaces names and makes it easier to read and follow interviews where many different friends or siblings are mentioned. Aggregating information is more convenient for obtaining just the information. Whereas paraphrasing enables to reflect the own analysis. Furthermore, categories can be edited in a flexible way. Categories can be re-used and are saved as XML-files. They can be edited and be part of the publication. Information in the interviews can be replaced by these categories, changed or paraphrased, depending on the researchers interest for secondary use of data. Further core elements of this basic tool are the option to design individual categories and labels. The replacement scheme is visible at any time. Identical text passages can be marked with different labels. And different text passages can be marked with the same labels. The replacements are stored individually at the corresponding text passages and not globally. Encoded saving to new text can be applied. Exported categories and text can be part of publications and can be shared with a research data center and within the community. The table for replacements includes the originals, the category, pseudonym, replacement and the number of occurrence and can be exported for a first analysis. Codes of different interviews can be compared.

1.4 Code Documentation

1.4.1 Used technologies and libraries

1.4.1.1 Maven

Maven is used to manage the various libraries of the program. Maven makes it easy to add different libraries via the pom.xml file. Maven also builds the final jar file.

1.4.1.2 Libraries

- **JavaFX** is used to create the graphical user interface (GUI). JavaFX uses among other things fxml.files, which serve as the basic framework for the GUI. Each window has its own fxml.file, which communicates with the rest of the program via a controller object. SceneBuilder was used for visual editing of the fxml.files. SceneBuilder was used for visual editing of the fxml.files. Furthermore, JavaFX allows the use of CSS stylesheets to change the interface design. CSS was hardly used in the program.

- **RichTextFX** extends the TextAreas of JavaFX and makes it possible to highlight text sections in such a TextArea.

1.4.2 Class-Diagram

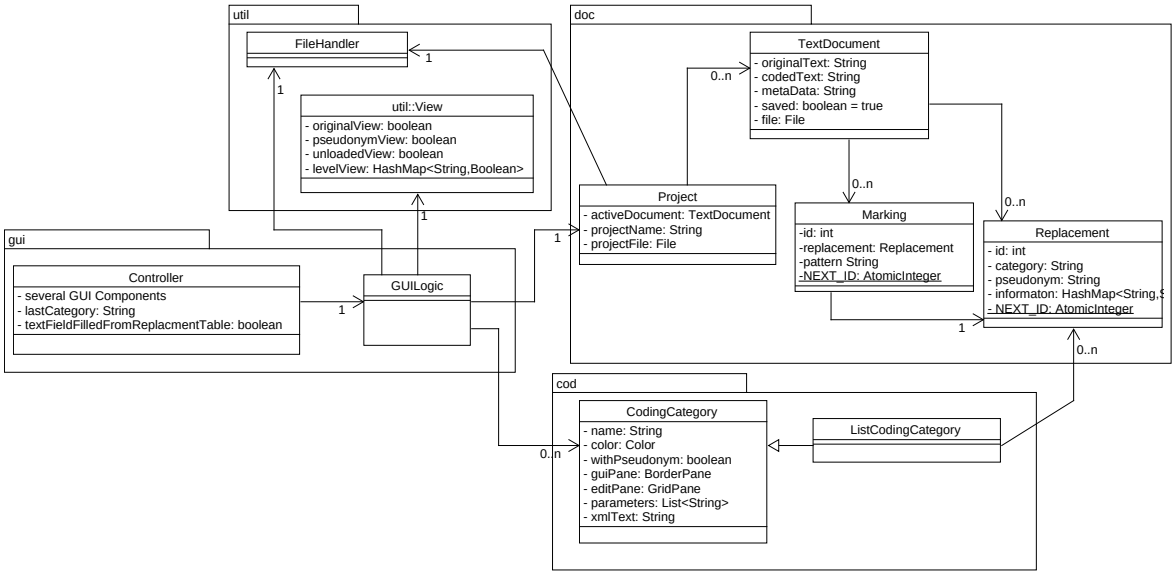


Figure 1.1: Class-Diagram reduced to Data-Objects and their Enitities

The program is divided into 4 packages as shown in Figure 1.1:

1. **gui:** In this package all classes are united, which take care of the interaction with and presentation on the GUI. This includes all controllers as well as GUILogic, which is a logic component that prepares the data for the GUI or prepares the input from The GUI for storage.

2. **util:** Different auxiliary classes are combined in this package. Thus the FileHandler takes care of the interaction with the operating system file system and a view object describes the view selected by the user.

3. **doc:** All classes that represent the structure of the opened documents are combined in this package. So an opened project has several documents, which in turn contain markings. Each of these markings will be replaced by a referenced replacement.

4. **cod:** To allow loading Projects independent of the encoding, the classes that contain the encoding categories are located separately from the doc-package in the cod-package. While CodingCategory represents a user-created category, a ListCodingCategory is created from a csv-list and has predefined replacements.

1.4.3 Text storage

To allow different views of the document text, the text is stored in two ways. First, the unchanged original text of the document is saved. On the other hand the coded text, in which the places to be replaced are replaced with the ID of the corresponding marker. Together with the lists of markings and replacements, a third text, called ExportText, can be created. This text is created dynamically depending on the selected view. Figure 1.2 shows an example of the differences in the different texts.

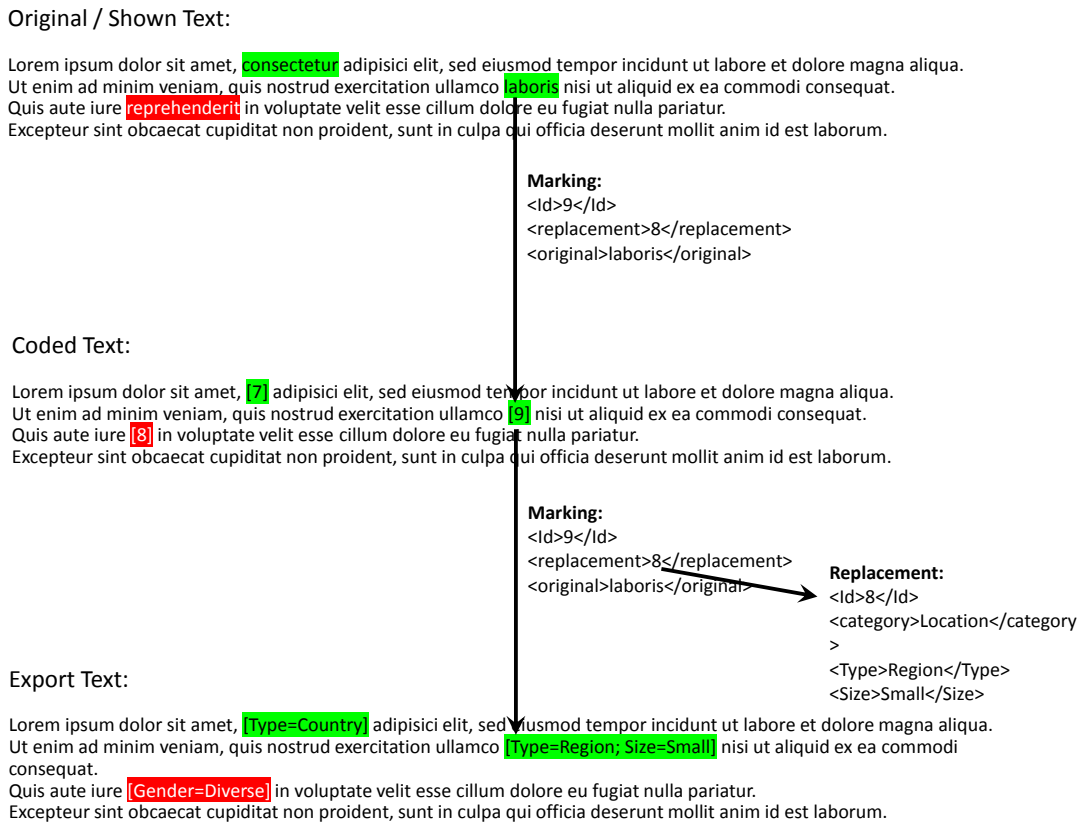


Figure 1.2: Text-Example for Original-, Coded- and Export-Text

1.4.4 GUI-Controller

This chapter lists which fxml files, which controllers are used in the code and which functions the windows have.

- **mainGUI.fxml/Controller.java**
  - Main window
  - The replacement table and file list are filled dynamically
  - Dynamically created pane for the categories
- **editMeta.fxml/ MetaController.java**
  - Allows the user to change the meta information of a document
  - No dynamic GUI components
- **exportFile.fxml/ ExportController.java**
  - Allows the user to export the currently selected document
  - Choice of different levels for export
  - The different levels are added dynamically
- **loadCategoryList.fxml/ ListController.java**

- Allows to select the desired labels for the new list category
- Divided into 4 pages
  1. Selection of the identifier
  2. Selection of a pseudonym (optional)
  3. Selection of labels
  4. Level assignment for the labels
- The selection options are created dynamically the frame is fixed in the fxml file
- Every time you turn the pages, the interface changes

- **singleListMarker.fxml/ SingleListMarkController.java**
  - Allows to select a replacement from a loaded list
  - Statically Built from the fxml file
  - However, the table is filled dynamically during the search

- **editCategories.fxml/ EditCategoriesController.java**
  - HelferObjekt: CodingCategoryGUIBuilder.java
  - Except for the frame and the "+" tab, everything is created dynamically

1.5 Conclusion

In cooperation with the RDC Qualiservice we built a tool that assists in working with interview data. This tool was the basis of the new Qualiservice Anonymization Tool which will be released in 2021. The conceptual framework of the Qualiservice Anonymization Tool allows to reach an appropriate level of anonymity, whilst trying to maintain maximum meaningful information in the research data. Studying qualitative interview data depends on the needs of every single researcher. Deciding individually which information to anonymize was the main scope of this project. For further information regarding the Qualiservice Anonymization Tool "QualiAnon" please contact the RDC Qualiservice via www.qualiservice.org. The basic elements of the tool presented here has been further developed by the world data archive PANGAEA as part of the Qualiservice consortium. Qualiservice currently validates the tool in different use cases. The release of the tool is scheduled for 2021.

Bibliography

[1] `https://timescapes-archive.leeds.ac.uk/`.

[2] Kathleen M. Carley. Coding choices for textual analysis: A comparison of content analysis and map analysis. *Sociological Methodology*, 23:75–126, 1993.

[3] Kathleen M. Carley. Extracting team mental models through textual analysis. *Journal of Organizational Behavior*, 18:533–558, 1997.

[4] Susanne Kretzer. Arbeitspapier zur konzeptentwicklung der anonymisierungs-/pseudonymisierung in qualiservice. *Arbeitspapier.*, 2013.

[5] Michael K. Martin, Juergen Pfeffer, and Kathleen M. Carley. Network text analysis of conceptual overlap in interviews, newspaper articles and keywords. *Social Network Analysis and Mining*, 3, 2013.

[6] Kati Mozygemba and et al. (o.J.). Handreichung zur anonymisierung qualitativer textgebundener forschungsmaterialien. *Qualiservice Working Paper*, Forthcoming.

[7] Jürgen Pfeffer and Kathleen M. Carley. Rapid modeling and analyzing networks extracted from pre-structured news articles. *Computational and Mathematical Organization Theory*, 18(3):280–299, 2012.

[8] Irena Medjedovic und Andreas Witzel. Wiederverwendung qualitativer daten: Archivierung und sekundärnutzung qualitativer interviewtranskripte. *VS Verlag für Sozialwissenschaften*, 2010.

5.3 Negative Word-of-Mouth

5.3.1 Against the Others! Detecting Moral Outrage in Social Media Networks

Authors Wienke Strathern, Mirco Schönfeld, Raji Ghawi, Jürgen Pfeffer

In 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, The Hague (Netherlands), pp. 322-326,
ISBN: 978-1-7281-1056-1,
https://doi.org/10.1109/ASONAM49781.2020.9381415.

Abstract

Online firestorms on Twitter are seemingly arbitrarily occurring outrages towards people, companies, media campaigns, and politicians. Moral outrage can create an excessive collective aggressiveness against one single argument, one single word, or one action of a person resulting in hateful speech. With a collective "against the others" the negative dynamics often start. Using data from Twitter, we explored the starting points of several firestorm outbreaks. As a social media platform with hundreds of millions of users interacting in real-time on topics and events all over the world, Twitter serves as a social sensor for online discussions and is known for quick and often emotional disputes. The main question we pose in this article is whether we can detect the outbreak of a firestorm. Given 21 online firestorms on Twitter, the key questions regarding anomaly detection are 1) How can we detect changing points? 2) How can we distinguish the features that indicate a moral outrage? In this paper, we examine these challenges by developing a method to detect the point of change by systematically focusing on linguistic cues in tweets. We are able to detect outbreaks of firestorms early and precisely only by identifying linguistic cues. The results of our work can help

detect negative dynamics and may have the potential for individuals, companies, and governments to mitigate hate in social media networks.

Publication Summary

Questions: In Strathern et al. (2020b), we aim to map the structure and dynamics of communication conflicts by examining their linguistic properties. Twitter is known for its quick and emotional disputes, which often escalate into hate speech, threats, and collective aggression. These online firestorms typically begin with a collective "against the others" mentality, targeting individuals and companies alike. The motivations for such attacks are varied, but revenge seems to be a common driving force, and triggers can be arbitrary and emotion-based. The consequences of these firestorms can be severe, causing intimidation, reputational damage, and deepening polarization. Our investigation centers on identifying the tipping point in times of unrest, tracking the evolution of moral outrage, and analyzing users' differing commenting patterns. We also explore how networks, sentiments, and linguistics shift during a firestorm. Our primary research question is whether we can detect the onset of a firestorm and pinpoint the features that signify moral outrage. To address this, we employ a combination of network analysis and text statistics to examine the occurrence of sentiments and linguistic cues over time. However, detecting the signal that indicates a firestorm anomaly poses a significant challenge.

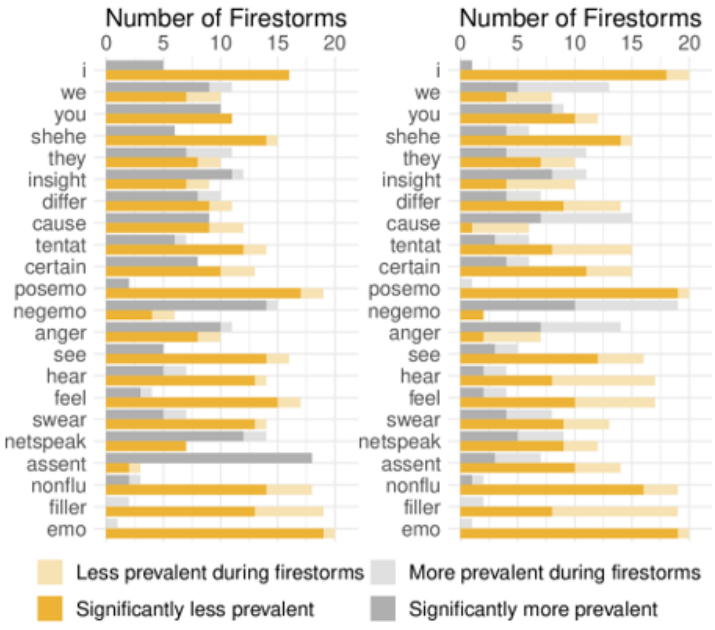


Figure 5: Example change point detection conducted on a linguistic category according to Strathern et al. (2020b)

Data: We used 21 firestorms, which includes almost 8200 tweets from ap-

proximately 6600 users in the first week of each event. To enhance our dataset, we include all tweets from the users that participated in the firestorms from the week before and the week after the starting day of the firestorm, giving us data from 15 days in total, with the start of the firestorm in the middle.

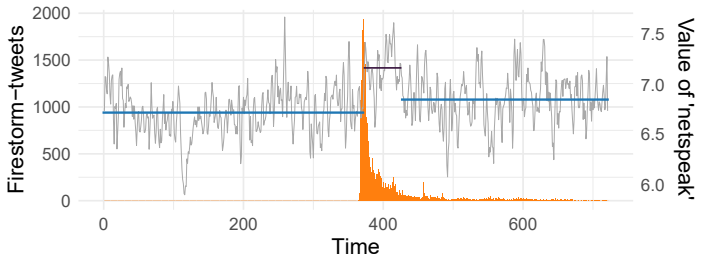


Figure 6: Change point detection model according to Strathern et al. (2020b)

Methods and Analysis: From a linguistic perspective, moral outrages are fascinating to examine. The language people use can reveal a great deal about their behavior and attitude. We hypothesize that moral outrage contains strong emotional expressions, and we aim to investigate the types of words used. To address this question, we first analyzed the linguistic characteristics of firestorm tweets. We employed the Linguistic Inquiry Word Count classification scheme, which is a common method for analyzing large text corpora. This tool includes an extensive dictionary, and each text is matched with about 90 different word categories based on psychological and linguistic research. We focused on sentiments and inspirational words to investigate whether users exhibit forms of aggressiveness during firestorms. **Results and Interpretation:** Our study's main finding is that during firestorms, users tend to use fewer first-person pronouns (such as "I") (cf. Figure 5), and instead, they mention the person or group being targeted in the attack more frequently. This shift in perspective indicates a change in focus from the individual user to the attack on others. We refer to this phenomenon as "Against the others!" based on our observations. To predict the onset of a firestorm, we utilized change point detection and the linguistic features we extracted. In Figure 6, we demonstrate how we used this approach to identify a firestorm at an early stage.

Author Contribution

Wienke Strathern headed the project, developed the research question, elaborated the study design and the methods, conducted and wrote the literature

review, selected the dataset, conducted the linguistic analysis of the tweets using the dictionary, wrote the introduction, wrote the conclusion and discussion, overall manuscript writing, revision, and editing, coordinated the team. Wienke Strathern was responsible for the overall manuscript.

Against the Others! Detecting Moral Outrage in Social Media Networks

Wienke Strathern*, Mirco Schoenfeld†, Raji Ghawi*, and Juergen Pfeffer*

* Bavarian School of Public Policy
Technical University of Munich, Munich, Germany
{wienke.strathern, raji.ghawi, juergen.pfeffer}@tum.de

† University of Bayreuth
Bayreuth, Germany
mirco.schoenfeld@uni-bayreuth.de

***Abstract*—Online firestorms on Twitter are seemingly arbitrarily occurring outrages towards people, companies, media campaigns and politicians. Moral outrage can create an excessive collective aggressiveness against one single argument, one single word, or one action of a person resulting in hateful speech. With a collective "against the others" the negative dynamics often start. Using data from Twitter, we explored the starting points of several firestorm outbreaks. As a social media platform with hundreds of millions of users interacting in real-time on topics and events all over the world, Twitter serves as a social sensor for online discussions and is known for quick and often emotional disputes. The main question we pose in this article is whether we can detect the outbreak of a firestorm. Given 21 online firestorms on Twitter, the key questions regarding the anomaly detection are: 1) How can we detect changing points? 2) How can we distinguish the features that indicate a moral outrage? In this paper we examine these challenges developing a method to detect the point of change systematically spotting on linguistic cues of tweets. We are able to detect outbreaks of firestorms early and precisely only by applying linguistic cues. The results of our work can help detect negative dynamics and may have the potential for individuals, companies, and governments to mitigate hate in social media networks.**

***Index Terms*—Firestorms, Twitter, Change Detection**

I. Introduction

Twitter is a social media platform with millions of users exchanging ideas about daily topics [1]. Its influence on societal processes is widely discussed. Acting as social sensors for real-time discussions, users provide information about ongoing discussions for live events and media topic strategies. User interactions have provided real-time information about the success and not-success of media campaigns and public relation events. One phenomena are online firestorms [2]–[8]. Negative online dynamics can be very dangerous in real life and can do harm to people. A single statement or media outlet can trigger a collective brawl that seems to escalate uncontrollably until a certain point of exhaustion.

Research questions. Analyzing real-time communication data on Twitter can help to understand the emergence of online moral outrages, negative dynamics and collective action [9]. Hence, the key research questions of our study are, "what are the major features that indicate the outbreak of a firestorm?" and "how can we detect relevant occurrences by exploring firestorm data?"

Methods. In order to address our research question of the relationship between lexical features and firestorm participation, we use the extracted characteristics of firestorms to detect an outbreak at an early stage. Our approach provides a method from network analysis and text statistics by examining the dynamics of linguistic cues over time.

On this account, we assume that detecting change based on sentiment analysis plus the usage of pronouns is more significant in how people connect with each other to form an outrage. Combining the automated processes that is done by the LIWCTool [10] and looking for explicit lexical features, could help to answer the above posted questions. Function words are psychologically and linguistically interesting and have been studied broadly [11]. Pronouns refer to a referent, hence, tell to whom somebody is speaking [12]. In this way, we might figure out if actors in social media networks stop talking about themselves and start talking collectively against somebody emotionally and with the words they use.

Contributions. The goal is to detect sentimental and lexical changes as a signal of an underlying change in a social network. In summary, our contributions are:

- Model: We propose a novel change detection model that accounts for linguistic cues and is able to detect the outbreak of a firestorm closely and quickly.
- Algorithm: We are able to detect firestorms on streaming Twitter data by only monitoring a couple of lexical features.

II. Related Work

Online firestorms are similar to rumors to some extent, e.g. they often rely on hearsay and uncertainty, online firestorms pose new challenges due to the speed and potential global reach of social media dynamics [2]. Why do people join online firestorms? Based on the concept of moral panics the authors argue that participation behavior is driven by a moral compass and a desire for social recognition [7]. Social norm theory refers to understanding online aggression in a social-political online setting, challenging the popular assumption that online anonymity is one of the principle factors that promotes aggression [4].

With respect to firestorms on social media, the analysis of dynamics and their early detection often involves research

from the field of sentiment analysis, network analysis as well as change point detection.

Sentiment analysis. Sentiment analysis was applied to analyze the emotional shape of moral discussions in social networks [13]. It has been argued that moral-emotional language increased diffusion more strongly. Highlighting the importance of emotion in the social transmission of moral ideas, the authors demonstrate the utility of social network methods for studying morality. A different approach is to measure emotional contagion in social media and networks by evaluating the emotional valence of content the users are exposed to before posting their own tweets [14]. Modeling collective sentiment on Twitter gave helpful insights about the mathematical approach to sentiment dynamics [15]. Arguing that rational and emotional styles of communication have strong influence on conversational dynamics, sentiments were the basis to measure the frequency of cognitive and emotional language on Facebook [16]. Extracting the patterns of word choice in an online social platform reflecting on pronouns is one way to characterize how a community forms in response to adverse events such as a terrorist attack [17].

Network analysis. Social media dynamics can be described with models and methods of social networks [18]. Approaches mainly evaluating network dynamics are, for example, proposed by Snijders et al. Here, network dynamics were modeled as network panel data [19]. This study demonstrated ways in which network structure reacts to users posting and sharing content. While examining the complete dynamics of the Twitter information network, the authors showed where users post and reshare information while creating and destroying connections. Dynamics of network structure can be characterized by steady rates of change, interrupted by sudden bursts [20]. Dynamics of online firestorms where analyzed applying an agent-based computer simulation (ABS) [21]—information diffusion and opinion adoption are triggered by negative conflict messages. In other works, techniques from social network analysis were combined with those from statistical process control in order to detect when significant change occurs in longitudinal network data [22].

Change point detection. The best known approaches for change point detection include Binary Segmentation [23], [24], Segment Neighborhood [25], and Optimal Partitioning [26], all of which suffer from certain drawbacks when considering monitoring streaming data: Binary Segmentation is quite efficient in terms of computational complexity, i.e. $\mathcal{O}(n \log n)$, but it cannot guarantee to find the global minimum. Segment Neighborhood approaches suffer from computational complexity which might degenerate to $\mathcal{O}(n^3)$. A more recent approach was proposed by Killick et al. and it is based on the Optimal Partitioning in that it yields a guaranteed identification of the exact minimum while retaining a computational complexity that is linear in the number of samples n [27]. Their approach is called the Pruned Exact Linear Time (PELT) method and is based on a work by Jackson et al. [26]. Most importantly, their method has a linear computational complexity which renders it especially useful for applications on streaming data.

Mixed approaches. More recent approaches analyze online firestorms by analyzing both content and structural information. A text-mining study on online firestorms evaluates negative eWOM that demonstrates distinct impacts of high- and low-arousal emotions, structural tie strength, and linguistic style match (between sender and brand community) on firestorm potential [28]. Online Firestorms were studied to develop optimized forms of counteraction, which engage individuals to act as supporters and initiate the spread of positive word-of-mouth, helping to constrain the firestorm as much as possible [5]. By monitoring both linguistic and psychological features of anomaly in the mention networks of online firestorms, we also combine analysis of content with the focus on structural information. To be able to detect online firestorms quickly, we also employ a method of change point detection on time series of the extracted features.

III. Data

We used the same set of 21 firestorms as in [3], whose data source is an archive of the Twitter decahose, a random 10% sample of all tweets. Mention and re-tweet networks based on these samples can be considered as *random edge sampled* networks [29] since sampling and network construction is based on Tweets that constitutes the links in the network. The set of tweets of each firestorm covers the first week of the event including on average 8199.29 tweets from 6641.76 users.

We augmented this dataset by including all decahose tweets from the users that participated in the firestorms from the 7 days before and the 7 days after the starting day of the firestorm, i.e. 15 days overall with the start of the firestorm in the middle. The fraction of firestorm-related tweets is between 2% and 8% of the tweets of each event—it is important to realize at this point that even for users engaging in online firestorms, this activity is a minor part of their overall activity on the platform.

A. Mention and Retweet Networks

To get insight on the evolution of each event, we opt to split time into units of *half hours*. This allows us to perform analysis at fine granularity. The result of this splitting is a series of about 720 time slices (since the studied time-span of an event is 15 days, this period corresponds to 720 half hours). At each time point we construct *mention networks*, and *retweet networks* taking into account all the tweets during the last 12 hours. This way we obtain a moving window of tweets: with a window size of 24 slices at steps of half hours. The *mention network* of each moving window contains an edge ($user_1$, $user_2$) if a tweet (among tweets under consideration) posted by $user_1$ contains a mention to $user_2$. The *retweet network* of each moving window contains an edge ($user_1$, $user_2$) if a tweet (among tweets under consideration) posted by $user_1$ is a retweet of another (original) tweet posted by $user_2$.

IV. Analysis

A. Analysis of Mention Networks

For each event, the mention networks constructed at the different time points are directed, unweighted networks.

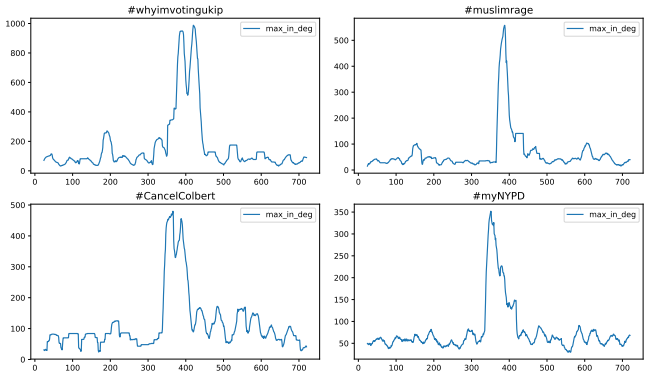


Fig. 1. Maximum in-degree in mention networks.

We performed several types of social network analysis and extracted a set of metrics, including: number of nodes N, and edges E, and density, average out-degree (which equals avg. in-degree), maximum out-degree, and maximum in-degree absolute and relative size of the largest connected component, as well as ratio of mention tweets to all tweets, mention per tweet ratio: E / nr. tweets, mention per 'mention' user ratio: E / N, tweet per 'mention' user ratio: nr. tweets / N.

Each of the aforementioned features leads to a time-series when taken over the entire time-span of the event. We find the maximum in-degree feature is one of the best features to detect this change. Figure 1 shows the time-series of maximum in-degree for the events with the largest number of tweets. The ability of this feature to detect a firestorm can be interpreted by considering that, generally speaking, a firestorm occurs when one user is being mentioned unusually high. However, the change of focus to a particular user can be the result of different (including positive) events.

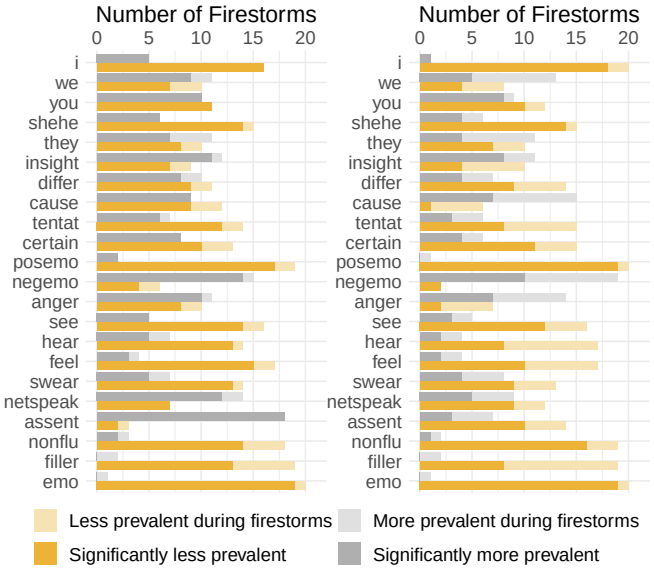
A more rigorous analysis of the change in behavior of such features is necessary in order to devise a formal method/algorithm of change detection as we will see in the next section.

B. Change of language

The first step was to uncover the linguistic peculiarities of firestorms. We classified all tweets using the LIWC classification scheme [30] and compared between firestorm tweets and non-firestorm tweets. The comparisons refer to the following categories: personal pronouns, affective processes, cognitive processes, perceptual processes, and informal language.

These categories each contain several subcategories that can be subsumed under the category names. The category of personal pronouns, for example, contains several subcategories referring to personal pronouns in numerous forms. One of these subcategories 'I', for example, includes—besides the pronoun 'I'—'me', 'mine', 'my', and special netspeak forms such as 'idk' (which means "I don't know").

Netspeak is a written and oral language, an internet-chat, which has developed mainly from the technical circumstances: the keyboard and the screen. The combination of technology



(a) All Tweets (b) Mention Networks

Fig. 2. Comparison of linguistic features between firestorm tweets and non-firestorm tweets

and language makes it possible to write the way you speak [31]. For each individual subcategory, we obtain the mean value of the respective LIWC values for the firestorm tweets and the non-firestorm tweets.

In Figure 2 the comparisons between firestorm tweets and non-firestorm tweets are shown with regard to the individual subcategories. The firestorm-tweets were compared with tweets from the week immediately before the firestorm.

Figure 2a refers to all tweets, while for Figure 2b only tweets from the mention network were considered. In both cases every subcategory was examined separately for all 21 firestorms. The grey bars represent the number of firestorms in which terms from the respective category occurred more frequently during the firestorms. The orange bars visualize the number of firestorms in which the same words occurred less frequently during the firestorms. The sum of the orange and grey bars is therefore always 21.

The light areas of the bars indicate that the mean values were different from each other. The strongly colored areas of the bars indicate that these differences were significant in terms of t-tests with $p < 0.01$. For category 'I' in Figure 2a, this means that in 5 Firestorms people used words of this category significantly more often, while in 16 Firestorms these words were used significantly less. Words of the same category were used less in 20 Firestorms considering the mention networks alone as depicted in Figure 2b. In 19 of these Firestorms, the differences were also significant.

In addition to the category 'I', the categories 'posemo' and 'assent' should also be highlighted. Words representing positive emotions like 'love', 'nice', 'sweet'—the 'posemo' category—are used clearly (and often significantly) less in almost all firestorms: positive emotions were less present in 19

out of 21 firestorms. In 17 out of 19 firestorms the differences were significant. This effect even increases when looking at mention networks. For the third remarkable category 'assent', which contains words like 'agree', 'OK', 'yes', this effect is reversed for all tweets—words in this category are used significantly more often during almost all firestorms (18 out of 21). When looking at the mention networks, however, this feature lacks accuracy. The differences are significant only in 13 firestorms.

Finally, we constructed our own category 'emo' by calculating the difference between positive and negative sentiments in tweets. Thus, weights of this category can be negative and should describe the overall sentiment of a tweet. There are 19 firestorms in which the 'emo' values were significantly lower during a firestorm. At the same time, there was only one firestorm with higher values of 'emo' but these differences were not significant. Checking if the differences remain visible decomposing the mention networks into components comparing tweets inside the largest component to tweets outside that component, we see no effect. There are only a few firestorms, in which the use of 'we' is significantly larger inside the largest component of the mention network.

C. Change point detection

The goal is to identify a firestorm at an early stage with the help of linguistic features. For the detection of change points, we use an efficient method that is suitable for being applied to streaming data and that was proposed by Killick et al. [27].

We constructed individual time series of the linguistic features. For this purpose, we first split the timeline of each of the firestorm data sets into buckets of half hours and assign tweets to buckets based on their timestamp. When constructing a time series of linguistic characteristics, we describe a bucket of tweets by the mean value of the corresponding LIWC values. To detect the change points in streaming data, we simulate the arrival of new tweets every half hour. Being able to decide at any time t whether a change has occurred, we use historical data from the past 24 hours. The time series applied for change point detection thus consists of 49 values for the interval $[t-48:t]$ each representing the mean of the LIWC values of the respective tweets. By doing so, we create separate time series for each of the subcategories mentioned above—see Figure 3. We do not apply any smoothing to the time series.

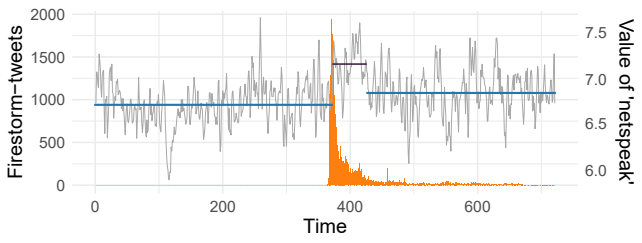


Fig. 3. Example change point detection conducted on a linguistic category. The number of Firestorm tweets is depicted in the background.

Decisive for successful change point detection is the choice of the penalty parameter. We select this parameter using elbow criterion. Therefore, we iterate penalty parameters from 2 to 10 and obtain the number of identified change points. From this data we determine the optimal penalty parameter as the configuration with the maximum absolute second derivative using the approximation of the second derivative of a point x_i as $x_{i+1} + x_{i-1} - 2x_i$. Choosing a higher penalty value results in fewer change points detected and vice versa.

We were able to detect the start of the firestorms in -0.55 ∓ 2.34 hours. We have defined the start time as the first interval of half an hour at which the hashtag or @user mention of the firestorm was the most frequent hashtag or @user mention in the data set. Due to the focused data collection process, the set contains little hashtags or @user mentions that were used frequently. Hence, this definition of a starting point of firestorms is quite sensitive, i.e. a low number of tweets suffices to boost the relevant hashtag or @user mention.

Also, it takes two intervals of half an hour until the beginning of the firestorm is noticed, i.e. the minimum deviation from the start time was measured. Hence, we are able to detect a change in the linguistic behaviour of the users quickly.

In a next step, we explored change near the peak of a firestorm. Determining this peak was done in two ways. We were able to approximate the peak of network dynamics with an average of $+1.19 \mp 2.51$ hours meaning that the change point closest to the peak is on average shortly after this peak. The second, natural definition of the peak is the interval of half an hour in which most firestorm-related tweets were recorded. We were able to approximate this peak of tweet accumulations with an average of $+0.14 \mp 1.30$ hours.

With respect to the identified differences in language use that were discussed in relation to Figure 2, we further evaluated how many change points were identified on the timelines of the linguistic categories. Figure 4 depicts how often a change point could be detected from the timeline of an individual characteristic. The top three categories were 'netspeak', 'I', and 'posemo' which corresponds to the insights from Figure 2—these were the categories with the most significant differences just after our own category 'emo'.

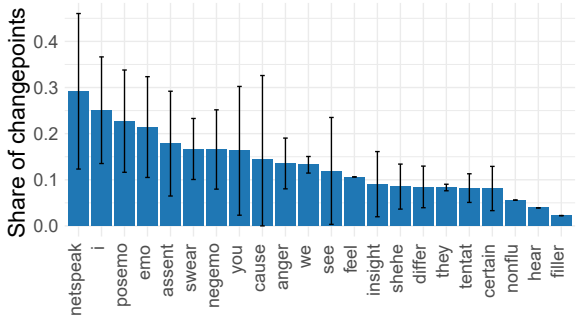


Fig. 4. How often were predictors relevant for the detection of a relevant changepoint

V. Discussion & Conclusion

From a network perspective, a firestorm occurs when one user is being mentioned unusually high—focusing on a Twitter handle or a hashtag. The maximum in-degree in mention networks is significantly deviating from comparable time periods. By evaluating lexical cues from the Tweet comments, we evaluated collective behavior manifesting in individual choices of words.

During firestorms, users talk significantly less about themselves compared to non-firestorm periods. Simultaneously, the positivity in firestorms tweets vanishes and negativity rises. The extracted lexical features were applicable to streaming data. Using lexical features to monitor change in behavior has the advantage of constant memory requirements.

By applying a straightforward change point detection, we were able to detect the starting point of the firestorms closely and quickly. We further provide insight into which linguistic categories proved to be useful for monitoring change.

According to our posed questions, combining sentiment analysis and text statistics to explore firestorm data can reveal how people connect with each other to form an outrage. The usage of vocabulary changes at a certain point when every single user stops commenting with the I-perspective and starts commenting on others. As mentioned, pronouns refer to a referent. If the 'I' diminishes, the focus changes significantly. All of a sudden people stop talking collectively about themselves positively and collectively more negatively—against the others!

Our model picks up these features and is able to detect the starting point of outrages giving insights into collective changing behavior. Further research questions regarding spreading of rumours and moral outrages might be: What causes evolving collective emotionality? Why does a community or society may at times come together and simultaneously communicate the same thought and participate in the same action? A better knowledge of individual motivations and collective action can help to better understand and detect online firestorms.

VI. Acknowledgements

The author(s) gratefully acknowledge the financial support from the Technical University of Munich - Institute for Ethics in Artificial Intelligence (IEAI). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the IEAI or its partners.

References

[1] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in *IEEE/HICSS*, 2010, pp. 1–10.

[2] J. Pfeffer, T. Zorbach, and K. M. Carley, "Understanding online firestorms: Negative word-of-mouth dynamics in social media networks," *Journal of Marketing Communications*, vol. 20/1–2, pp. 117–128, 2014.

[3] H. Lamba, M. M. Malik, and J. Pfeffer, "A Tempest in a Teacup? Analyzing Firestorms on Twitter," in *2015 IEEE/ACM ASONAM*, New York, NY, USA, 2015, p. 17–24.

[4] K. Rost, L. Stahel, and B. S. Frey, "Digital Social Norm Enforcement: Online Firestorms in Social Media," *PLOS ONE*, vol. 11, no. 6, p. e0155923, 2016.

[5] A. Mochalova and A. Nanopoulos, "Restricting the spread of firestorms in social networks," *ECIS 2014 Proceedings*, 2014.

[6] B. Drasch, J. Huber, S. Panz, and F. Probst, "Detecting Online Firestorms in Social Media," *ICIS*, 2015.

[7] M. Johnen, M. Jungblut, and M. Ziegele, "The digital outcry: What incites participation behavior in an online firestorm?" *New Media & Society*, vol. 20, no. 9, pp. 3140–3160, 2018.

[8] L. Stich, G. Golla, and A. Nanopoulos, "Modelling the spread of negative word-of-mouth in online social networks," *Journal of Decision Systems*, vol. 23, no. 2, pp. 203–221, 2014.

[9] M. J. Crockett, "Moral outrage in the digital age," *Nature Human Behaviour*, vol. 1, pp. 769–771, 2017.

[10] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of Language and Social Psychology*, 2009.

[11] P. Dekker, "Pronouns in a pragmatic semantics," *Journal of Pragmatics*, vol. 34, no. 7, pp. 815–827, 2002.

[12] J. W. Pennebaker, *The secret life of pronouns: What our words say about us*, ser. The secret life of pronouns: What our words say about us. Bloomsbury Press/Bloomsbury Publishing, 2011, pages: xii, 352.

[13] W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, and J. J. V. Bavel, "Emotion shapes the diffusion of moralized content in social networks," *PNAS*, vol. 114, no. 28, pp. 7313–7318, 2017.

[14] E. Ferrara and Z. Yang, "Measuring emotional contagion in social media," *PLOS ONE*, vol. 10, no. 11, 2015.

[15] N. Charlton, C. Singleton, and D. V. Greetham, "In the mood: the dynamics of collective sentiments on Twitter," *Royal Society Open Science*, vol. 3, no. 6, 2016.

[16] C. A. Bail, T. W. Brown, and M. Mann, "Channeling Hearts and Minds: Advocacy Organizations, Cognitive-Emotional Currents, and Public Conversation," *American Sociological Review*, vol. 82, no. 6, pp. 1188–1213, 2017.

[17] S. Shaikh, L. B. Feldman, E. Barach, and Y. Marzouki, "Tweet Sentiment Analysis with Pronoun Choice Reveals Online Community Dynamics in Response to Crisis Events," in *Advances in Cross-Cultural Decision Making*. Springer International Publishing, 2017, pp. 345–356.

[18] M. Hennig, U. Brandes, J. Pfeffer, and I. Mergel, *Studying Social Networks. A Guide to Empirical Research*. Campus Verlag, 2012.

[19] T. A. Snijders, J. Koskinen, and M. Schweinberger, "Maximum likelihood estimation for social network dynamics," *The annals of applied statistics*, vol. 4, no. 2, pp. 567–588, 2010.

[20] S. A. Myers and J. Leskovec, "The bursty dynamics of the Twitter information network," in *WWW*, Seoul, Korea, 2014, pp. 913–924.

[21] F. Hauser, J. Hautz, K. Hutter, and J. Füller, "Firestorms: Modeling conflict diffusion and management strategies in online communities," *Journal of Strategic Information Systems*, vol. 26/4, pp. 285–321, 2017.

[22] I. McCulloh and K. M. Carley, "Detecting change in longitudinal social networks," *Journal of Social Structure*, vol. 12/3, pp. 1–37, 2011.

[23] A. J. Scott and M. Knott, "A cluster analysis method for grouping means in the analysis of variance," *Biometrics*, vol. 30, pp. 507–512, 1974.

[24] A. Sen and M. S. Srivastava, "On tests for detecting change in mean," *The Annals of Statistics*, vol. 3, no. 1, pp. 98–108, 1975.

[25] I. E. Auger and C. E. Lawrence, "Algorithms for the optimal identification of segment neighborhoods," *Bulletin of Mathematical Biology*, vol. 51, no. 1, pp. 39–54, Jan 1989.

[26] B. Jackson, J. D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and Tun Tao Tsai, "An algorithm for optimal partitioning of data on an interval," *IEEE Signal Processing Letters*, vol. 12, no. 2, pp. 105–108, 2005.

[27] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012.

[28] D. Herhausen, S. Ludwig, D. Grewal, J. Wulf, and M. Schoegel, "Detecting, Preventing, and Mitigating Online Firestorms in Brand Communities," *Journal of Marketing*, vol. 83, no. 3, pp. 1–21, 2019.

[29] C. Wagner, P. Singer, F. Karimi, J. Pfeffer, and M. Strohmaier, "Sampling from social networks with attributes," in *Proceedings of the WWW Conference*, 2017, pp. 1181–1190.

[30] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The Development and Psychometric Properties of LIWC2015," The University of Texas at Austin, Tech. Rep., 2015.

[31] D. Crystal, "Language and the internet," *IEEE Transactions on Professional Communication*, pp. 142–144, 2002.

5.3.2 Identifying Lexical Change in Negative-Word-of-Mouth on Social Media

This peer-reviewed paper is relevant for examination.

Authors Wienke Strathern, Raji Ghawi, Mirco Schönfeld, Jürgen Pfeffer

In Social Network Analysis and Mining. 12, 59 (2022), Cham: Springer International Publishing, https://doi.org/10.1007/s13278-022-00881-0.

Abstract

Negative word-of-mouth is a strong consumer and user response to dissatisfaction. These moral outrages enrage out of a collective aggressiveness against a single argument, word, or action of a person or a company. In this work, we examine the vocabulary change to explore the outbreak of online firestorms on Twitter. The sudden change in an emotional state can be captured in language. It reveals how people connect to form an outrage. We find that when users turn their outrage against somebody, the occurrence of self-referencing pronouns like 'I', 'me', ... reduces significantly. Using data from Twitter, we derive such linguistic features and features based on retweet and mention networks to use them as indicators for negative word-of-mouth dynamics in social media networks. Based on these features, we build three classification models that can predict the outbreak of a firestorm with high accuracy.

[4]https://creativecommons.org/licenses/by/4.0/

Publication Summary

Questions: In Strathern et al. (2022b) were are interested in testing machine learning methods to predict the start of a firestorm. Typically, people tend to talk about themselves a lot on social media. However, during a firestorm, the usage of pronouns such as I' and me' disappears, and the use of hashtags and mentions directed towards the target of the firestorm increases significantly. The focus shifts from the individual user to the attack on the other person. To investigate the relationship between linguistic changes and the emergence of online firestorms, we apply techniques from social network analysis, text statistics, and machine learning. Therefore, identifying changes in speech enables us to provide predictive analysis and early warnings of potential issues. In this study, we examine the lexical and network characteristics of tweets in different time periods. We primarily focus on textual data from tweets, as well as mention and retweet networks. **Data:** We use 21 firestorms, which includes almost 8200 tweets from approximately 6600 users in the first week of each event. To enhance our dataset, we include all tweets from the users that participated in the firestorms from the week before and the week after the starting day of the firestorm.

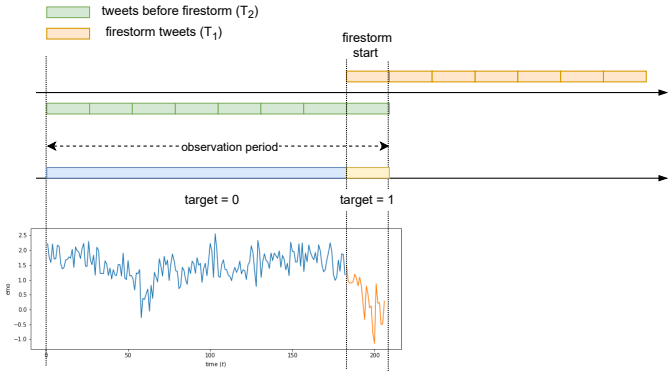


Figure 7: Timeline of a firestorm according to Strathern et al. (2022b)

Methods and Analysis: The aim of our study is to predict the onset of a firestorm using the aforementioned linguistic and network features. This task involves a binary classification approach where we determine whether a particular time point falls within the start period of a firestorm, based on various tweet features. To analyze the development of each firestorm, we divide the dataset into approximately 360-time slices, as depicted in Figure 7. This approach allows us to investigate the temporal evolution of each firestorm and identify when significant changes occur. **Results and Interpretation:** The results indicate that all

Table 4: Accuracy of prediction models according to Strathern et al. (2022b)

	basic	linguistic	mention	retweet
askbg	0.926	0.916	0.958	0.953
askjpm	0.953	0.953	0.995	0.995
cancelcolbert	0.948	0.953	0.990	0.984
celebboutique	0.915	0.945	0.937	0.963
david_cameron	1.000	0.995	0.995	0.995
fafsa	0.932	0.943	0.989	0.989
gaelgarciab	0.906	0.885	0.956	0.956
klm	0.891	0.902	0.907	0.907
mcdstories	0.943	0.938	0.923	0.961
muslimrage	0.956	0.990	0.980	0.969
mynypd	0.958	0.984	1.000	0.995
notintendedto..	0.990	0.984	0.989	0.989
qantas	0.922	0.922	0.939	0.956
qantasluxury	0.932	0.943	0.972	0.972
spaghettios	0.944	0.964	0.989	0.989
suey_park	0.943	0.948	0.990	0.995
theonion	0.974	0.974	0.989	0.989
ukinusa	0.870	0.875	0.995	0.995
voguearticles	0.951	0.967	0.971	0.977
whyimvotingukip	0.943	0.969	0.956	0.950
avg.	0.940	0.948	0.971	0.974

prediction models can accurately forecast the start of a firestorm. The network models show slightly higher precision compared to the basic model, while the linguistic model is slightly less accurate.

Author Contribution

Wienke Strathern headed the project, developed the research question, elaborated the study design and the methods, conducted and wrote the literature review, selected the dataset, conducted the linguistic analysis of the tweets, wrote the introduction, wrote the conclusion and discussion, wrote the manuscript, revisions and editing, coordinated the team. Wienke Strathern was responsible for the overall manuscript.

Social Network Analysis and Mining (2022) 12:59
https://doi.org/10.1007/s13278-022-00881-0

ORIGINAL ARTICLE

Identifying lexical change in negative word-of-mouth on social media

Wienke Strathern[1] · Raji Ghawi[1] · Mirco Schönfeld[2] · Jürgen Pfeffer[1]

Received: 25 March 2021 / Revised: 13 April 2022 / Accepted: 15 April 2022

Abstract
Negative word-of-mouth is a strong consumer and user response to dissatisfaction. Moral outrages can create an excessive collective aggressiveness against one single argument, one single word, or one action of a person resulting in hateful speech. In this work, we examine the change of vocabulary to explore the outbreak of online firestorms on Twitter. The sudden change of an emotional state can be captured in language. It reveals how people connect with each other to form outrage. We find that when users turn their outrage against somebody, the occurrence of self-referencing pronouns like 'I' and 'me' reduces significantly. Using data from Twitter, we derive such linguistic features together with features based on retweets and mention networks to use them as indicators for negative word-of-mouth dynamics in social media networks. Based on these features, we build three classification models that can predict the outbreak of a firestorm with high accuracy.

1 Introduction

As social media platforms with hundreds of millions of users interacting in real time on topics and events all over the world, social media networks are social sensors for online discussions and are known for quick and often emotional disputes (Chadwick 2017). Online firestorms can be defined as the sudden discharge of large quantities of messages containing negative word of mouth and complaint behavior against a person, company or group in social media networks (Pfeffer et al. 2014). The negative dynamics often start with a collective "against the others" (Strathern et al. 2020).

In social media, negative opinions about products or companies are formed by and propagated via thousands or millions of people within hours. Furthermore, massive negative online dynamics are not only limited to the business domain, but they also affect organizations and individuals in politics. Even though online firestorms are a new phenomenon, their dynamics are similar to the way in which rumors are circulated. In 1947, Gordon Allport and Leo Postman defined a rumor as a "proposition for belief, passed along from person to person, usually by word of mouth, without secure standards of evidence being presented" (Allport and Postman 1947).

When people are active on social media, they act in a socio-technical system that is mediated and driven by algorithms. The goal of social media platforms is to keep users engaged and to maximize their time spent on the platform. Highly engaged users who spend a lot of time on platforms are the core of a social media business model that is based on selling more and better targeted ads. But the question is always which content will be interesting for a particular user? To answer this, recommendation systems are developed to increase the chance that a user will click on a suggested link and read its content. These recommendation algorithms incorporate socio-demographic information, but also data of a user's previous activity (Leskovec et al. 2014; Anderson 2006).

Furthermore, behavioral data of alters (friends) of a user are also used to suggest new content (Appel et al. 2020). Social scientists have studied the driving forces of social relationships for decades, i.e., why do people connect with each other. Homophily and transitivity are the most important factors for network formation. Homophily means that your friends are similar to yourself (McPherson et al. 2001). They like similar things and are interested in similar topics.

✉ Wienke Strathern
wienke.strathern@tum.de

Raji Ghawi
raji.ghawi@tum.de

Mirco Schönfeld
mirco.schoenfeld@uni-bayreuth.de

Jürgen Pfeffer
juergen.pfeffer@tum.de

[1] School of Social Science and Technology, Technical University of Munich, Munich, Germany

[2] University of Bayreuth, Bayreuth, Germany

Published online: 07 June 2022

Transitivity describes the fact that a person's friends are often connected among each other (Heider 1946; Cartwright and Harary 1956). Combining these two aspects results in the fact that most people are embedded in personal networks with people that are similar to themselves and who are to a high degree connected among each other.

The above-described forces of how humans create networks combined with recommendation systems have problematic implications. Recommendation systems filter the content that is presented on social media and suggest new "friends" to us. As a result, filter bubbles (Pariser 2011) are formed around individuals on social media, i.e., they are connected to like-minded people and familiar content. The lack of diversity in access to people and content can easily lead to polarization (Dandekar et al. 2013). If we now add another key characteristic of social media, abbreviated communication with little space for elaborate exchange, a perfect breeding ground for online firestorms emerges. Consider a couple of people disliking a statement or action of a politician, celebrity or any private individual and these people voicing their dislike aggressively on social media. Their online peers, who most likely have similar views (see above), will easily and quickly agree by sharing or retweeting the discontent. Within hours, these negative dynamics can reach tens of thousands of users (Newman et al. 2006). A major problem, however, is to capture first signals of online outrage at an early stage. Knowing about these signals would help to intervene in a proper way to avoid escalations and negative dynamics.

In previous work, Strathern et al. (2020) tackled the question of anomaly detection in a network by exploring major features that indicate the outbreak of a firestorm; hence, the goal was to early detect change and extract linguistic features. Detection of outrage (e.g., hate speech) is based on identification of predefined keywords, while the context in which certain topics and words are being used has to be almost disregarded. To name just one extreme example, hate groups have managed to escape keyword-based machine detection through clever combinations of words, misspellings, satire and coded language (Udupa 2020). The focus of the analysis of Strathern et al. was on more complex lexical characteristics, which they applied as a basis for automated detection.

Our research question is the following: On Twitter, there is constant fluctuation of content and tweets and the question arises if, in these fluctuations, we can detect early that a negative event starts solely based on linguistic features. We assume that the start of a firestorm is a process, and because of a sudden change of emotions it can be early detected in sentiments and lexical items. With this work, we aim at answering the following question: Once we identify the linguistic changes as indicators of a firestorm, can we also predict a firestorm? In an abstract view

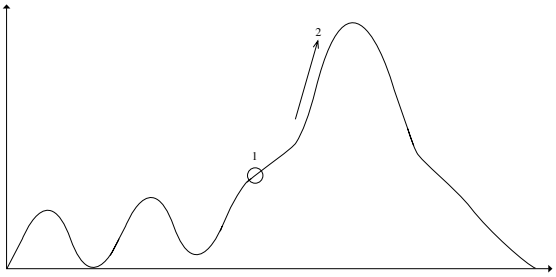


Fig. 1 Early detection of linguistic indicators (1) and prediction of firestorm (2)

on a firestorm as depicted in Fig. 1, the indicators show at time point 1), whereas the firestorm takes place starting during the phase marked by 2) in the figure. Hence, in this paper, we build upon and extend the work presented by Strathern et al. (2020).

Our choice of methods to answer our research question regarding the prediction of the beginning of online firestorms is based on text statistics and social network analysis for longitudinal network data. We assume that anomalies in behavior can be detected by statistical analysis applied to processes over time. Hence, in this work, we extract lexical and network-based properties, measure their occurrence for different tweet periods and use these features to predict the outbreak of a firestorm. For the scope of this work, we are mainly interested in textual data from tweets and in mention and retweet networks. We use quantitative linguistics to study lexical properties. For our linguistic analysis, we apply the Linguistic Inquiry Word Count Tool by Pennebaker et al. (2015). To contrast this linguistic perspective, we also investigate mention and retweet networks. Mentions and hashtags represent speech acts in linguistic pragmatics and are interesting in that they represent behavioral properties in addition to the lexical properties (Scott 2015). For predictive analysis, we define models based on linguistic features as well as models based on features derived from mention and retweet networks and compare them with each other.

Our contributions are:

- Extracting linguistic and sentimental features from textual data as indicators of firestorms.
- Defining a prediction model that accounts for linguistic features.

The remainder of the paper is organized as follows: Sect. 2 highlights important related works. In Sect. 3, we introduce the dataset used for this analysis together with a few descriptive statistics. What follows in Sects. 4 and 5 is a description of the linguistic and network-based features that our

prediction is based upon. The prediction task is described in detail in Sect. 6. Section 7 concludes the paper.

2 Related work

While online firestorms are similar to rumors to some extent, e.g. they often rely on hearsay and uncertainty, online firestorms pose new challenges due to the speed and potential global reach of social media dynamics (Pfeffer et al. 2014). With respect to firestorms on social media, the analysis of social dynamics, their early detection and prediction often involves research from the field of sentiment analysis, network analysis as well as change detection. There is work asking why do people join online firestorms (Delgado-Ballester et al. 2021). Based on the concept of moral panics, the authors argue that participation behavior is driven by a moral compass and a desire for social recognition (Johnen et al. 2018). Social norm theory refers to understanding online aggression in a social–political online setting, challenging the popular assumption that online anonymity is one of the principle factors that promote aggression (Rost et al. 2016).

2.1 Sentiment analysis

Approaches to the analysis of firestorms focusing on the mood of the users and their expressed sentiments unveil, for example, that in the context of online firestorms, non-anonymous individuals are more aggressive compared to anonymous individuals (Rost et al. 2016). Online firestorms are used as a topic of news coverage by journalists and explore journalists' contribution to attempts of online scandalization. By covering the outcry, journalists elevate it onto a mainstream communication platform and support the process of scandalization. Based on a typology of online firestorms, the authors have found that the majority of cases address events of perceived discrimination and moral misconduct aiming at societal change (Stich et al. 2014). Online firestorms on social media have been studied to design an Online Firestorm Detector that includes an algorithm inspired by epidemiological surveillance systems using real-world data from a firestorm (Drasch et al. 2015).

Sentiment analysis was applied to analyze the emotional shape of moral discussions in social networks (Brady et al. 2017). It has been argued that moral–emotional language increased diffusion more strongly. Highlighting the importance of emotion in the social transmission of moral ideas, the authors demonstrate the utility of social network methods for studying morality. A different approach is to measure emotional contagion in social media and networks by evaluating the emotional valence of content the users are exposed to before posting their own tweets (Ferrara and Yang 2015). Modeling collective sentiment on Twitter gave helpful insights about the mathematical approach to sentiment dynamics (Charlton et al. 2016).

Arguing that rational and emotional styles of communication have strong influence on conversational dynamics, sentiments were the basis to measure the frequency of cognitive and emotional language on Facebook. Bail et al. (2017).

Instead, the analysis of linguistic patterns was used to understand affective arousal and linguist output (Sharp and Hargrove 2004). Extracting the patterns of word choice in an online social platform reflecting on pronouns is one way to characterize how a community forms in response to adverse events such as a terrorist attack (Shaikh et al. 2017). Synchronized verbal behavior can reveal important information about social dynamics. The effectiveness of using language to predict change in social psychological factors of interest can be demonstrated nicely (Gonzales et al. 2010). In Lamba et al. (2015), the authors detected and described 21 online firestorms discussing their impact on the network. To advance knowledge about firestorms and the spread of rumors, we use the extracted data as a starting point to follow up on the research findings.

2.2 Network analysis

Social media dynamics can be described with models and methods of social networks (Wasserman and Faust 1994; Newman 2010; Hennig et al. 2012). Approaches mainly evaluating network dynamics are, for example, proposed by Snijders et al. Here, network dynamics were modeled as network panel data (Snijders et al. 2010). The assumption is that the observed data are discrete observations of a continuous-time Markov process on the space of all directed graphs on a given node set, in which changes in tie variables are independent conditional on the current graph. The model for tie changes is parametric and designed for applications to social network analysis, where the network dynamics can be interpreted as being generated by choices made by the social actors represented by the nodes of the graph. This study demonstrated ways in which network structure reacts to users posting and sharing content. While examining the complete dynamics of the Twitter information network, the authors showed where users post and reshare information while creating and destroying connections. Dynamics of network structure can be characterized by steady rates of change, interrupted by sudden bursts (Myers et al. 2012). Network dynamics were modeled as a class of statistical models for longitudinal network data (Snijders 2001). Dynamics of online firestorms were analyzed using an agent-based computer simulation (ABS) (Hauser et al. 2017)—information diffusion and opinion adoption are triggered by negative conflict messages.

2.3 Classification in machine learning

In order to efficiently analyze big data, machine learning methods are used, with the goal of learning from experience in certain tasks. In particular, in *supervised learning*, the goal is to predict some output variable that is associated with each input item. This task is called *classification* when the output variable is a category. Many standard classification algorithms have been developed over the last decades, such as logistic regression, random forests, k nearest neighbors, support vector machines and many more (Friedman et al. 2001; James et al. 2014).

Machine learning methods have been used widely for studying users' behavior on social media (Ruths and Pfeffer 2014), predicting the behavior of techno-social systems (Vespignani 2009) and predicting consumer behavior with Web search (Goel et al. 2010). Moreover, such methods are also used in identifying relevant electronic word of mouth in social media (Vermeer et al. 2019; Strathern et al. 2021).

2.4 Mixed approaches

More recent approaches analyze online firestorms by analyzing both content and structural information. A text-mining study on online firestorms evaluates negative eWOM that demonstrates distinct impacts of high- and low-arousal emotions, structural tie strength, and linguistic style match (between sender and brand community) on firestorm potential (Herhausen et al. 2019). Online Firestorms were studied to develop optimized forms of counteraction, which engage individuals to act as supporters and initiate the spread of positive word of mouth, helping to constrain the firestorm as much as possible (Mochalova and Nanopoulos 2014). By monitoring psychological and linguistic features in the tweets and network features, we combine methods from text analysis, social network analysis and change detection to early detect and predict the start of a firestorm.

3 Data

To address our research question, we examined 20 different firestorms. Some are directed against individuals and a single statement; some are against companies, campaigns and marketing actions. They have all received widespread public attention in social media as well as mainstream media. As shown in Table 1, there are hashtags and also @mentions that name the target.

3.1 Dataset

We used the same set of firestorms as in Lamba et al. (2015), whose data source is an archive of the Twitter decahose, a random 10% sample of all tweets. This is a scaled up version of Twitter's Sample API, which gives a stream of a random 1% sample of all tweets.

Table 1 Firestorm events sorted by number of tweets

Firestorm hashtag/mention	Tweets	Users	First day
#whyimvotingukip	39,969	32,382	2014-05-21
#muslimrage	15,721	11,952	2012-09-17
#CancelColbert	13,277	10,353	2014-03-28
#myNYPD	12,762	10,362	2014-04-23
@TheOnion	9959	8803	2013-02-25
@KLM	8716	8050	2014-06-29
#qantas	8649	5405	2011-10-29
@David_Cameron	7096	6447	2014-03-06
suey_park	6919	3854	2014-03-28
@celebboutique	6679	6189	2012-07-20
@GaelGarciaB	6646	6234	2014-06-29
#NotIntendedtobeaFactualStat.	6261	4389	2011-04-13
#AskJPM	4321	3418	2013-11-14
@SpaghettiOs	2890	2704	2013-12-07
#McDStories	2374	1993	2012-01-24
#AskBG	2221	1933	2013-10-17
#QantasLuxury	2098	1658	2011-11-22
#VogueArticles	1894	1819	2014-09-14
@fafsa	1828	1693	2014-06-25
@UKinUSA	142	140	2014-08-27

Mention and retweet networks based on these samples can be considered as *random edge sampled* networks (Wagner et al. 2017) since sampling and network construction is based on Tweets that constitute the links in the network. As found by Morstatter et al. (2013), the Sample API (unlike the Streaming API) indeed gives an accurate representation of the relative frequencies of hashtags over time. We assume that the decahose has this property as well, with the significant benefit that it gives us more statistical power to estimate the true size of smaller events.

The dataset consists of 20 firestorms with the highest volume of tweets as identified in Lamba et al. (2015). Table 1 shows those events along with the number of tweets, number of users, and the date of the first day of the event. The set of tweets of each firestorm covers the first week of the event. We also augmented this dataset via including additional tweets, of the same group of users, during the same week of the event (7 days) and the week before (8 days), such that the volume of tweets is balanced between the 2 weeks (about 50% each). The fraction of firestorm-related tweets is between 2 and 8% of the tweets of each event (Table 1)—it is important to realize at this point that even for users engaging in online firestorms,

this activity is a minor part of their overall activity on the platform.

Thus, for each of the 20 firestorms, we have three types of tweets: (1) tweets related to the firestorm, (2) tweets posted 1 week before the firestorm and (3) tweets posted during the firestorm (same week) but not related to it. Let us denote these three sets of tweets T_1, T_2 and T_3, respectively.

For each event, we also extracted tweets metadata including timestamp, hashtags, mentions and retweet information (user and tweet ID).[1]

4 Linguistic features

Negative word-of-mouth sometimes contains strong emotional expressions and even highly aggressive words against a person or a company. Hence, the start of a firestorm might be indicated by a sudden change of vocabulary and emotions. Do people become emotionally thrilled and can we find changes in tweets? Can we capture a change of perspective in the text against a target? Emotionality is reflected in words, the first analysis is based on the smallest structural unit in language: words (Bybee and Hopper 2001).

4.1 Extraction of features

To extract linguistic features and sentiment scores we use the Linguistic Inquiry Word Count classification scheme, short LIWCTool (Pennebaker et al. 2015). In this way, first textual differences and similarities can be quantified by simple word frequency distribution (Baayen 1993). Furthermore, to understand emotions in tweets we use the sentiment analysis provided by the LIWCTool. Essentially, sentiment analysis is the automatic determination of the valence or polarity of a text part, i.e., the classification of whether a text part has a positive, negative or neutral valence. Basically, automatic methods of sentiment analysis work either lexicon based or on the basis of machine learning. Lexicon-based methods use extensive lexicons in which individual words are assigned positive or negative numerical values to determine the valence of a text section (usually at the sentence level) of a text part (mostly on sentence level) (Tausczik and Pennebaker 2009).

LIWC contains a dictionary with about 90 output variables, so each tweet is matched with about 90 different categories. The classification scheme is based on psychological and linguistic research. Particularly, we were interested in sentiments to see if users show ways of aggressiveness during firestorms compared to non-firestorm periods. Furthermore, we would like to know which lexical items differ in different phases. We extracted 90 lexical features for each tweet of each of the 20 firestorms. We used variables that give standard linguistic dimensions (percentage of words in the text that are pronouns, articles, auxiliary verbs) and informal language markers (percentage of words that refer to the category assents, fillers, swear words, netspeak). To discover sentiments, we also used the variables affective processes, cognitive processes, perceptual processes. The categories provide a sentiment score of positivity and negativity to every single tweet. We also considered the category posemo and negemo to see if a tweet is considered positive or negative. We also constructed our own category 'emo' by calculating the difference between positive and negative sentiments in tweets. Thus, weights of this category can be negative and should describe the overall sentiment of a tweet.

These categories each contain several subcategories that can be subsumed under the category names. The category of personal pronouns, for example, contains several subcategories referring to personal pronouns in numerous forms. One of these subcategories 'I,' for example, includes—besides the pronoun 'I'—'me,' 'mine,' 'my,' and special netspeak forms such as 'idk' (which means "I don't know").

Netspeak is a written and oral language, an internet chat, which has developed mainly from the technical circumstances: the keyboard and the screen. The combination of technology and language makes it possible to write the way you speak (Crystal 2002).

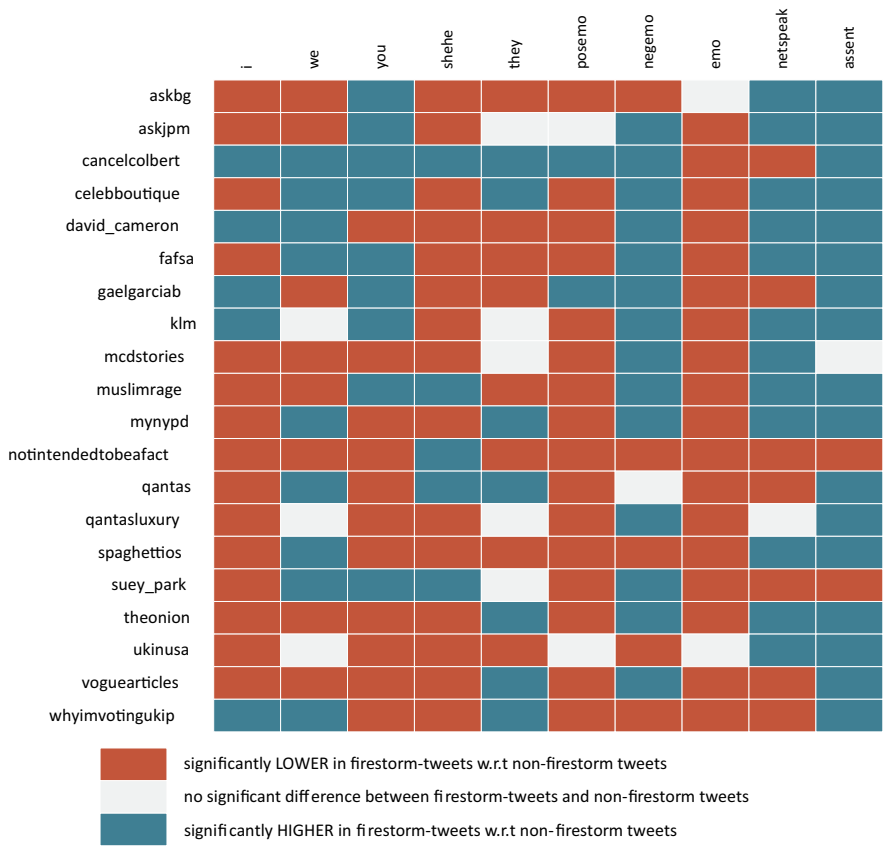
Finally, for each individual subcategory, we obtain the mean value of the respective LIWC values for the firestorm tweets and the non-firestorm tweets. Comparing these values gives first insights about lexical differences and similarities.

4.2 Comparing firestorm and non-firestorm tweets

In order to explore how the linguistic and sentiment features of tweets change during firestorms, we perform comparisons between firestorm tweets and non-firestorm tweets with regard to the individual LIWC subcategories. The firestorm tweets (T_1) were compared with tweets from the same user accounts from the week immediately before the firestorm (T_2) and the same week of the firestorm (T_3). We used t-tests to compare the mean value of the respective LIWC values for the firestorm tweets and the non-firestorm tweets, where the level of statistical significance of those tests is expressed using p-values (we used $p < 0.01$).

Figure 2a depicts the comparisons between firestorm tweets and non-firestorm tweets with regard to the individual subcategories. Every subcategory was examined separately for all 20 firestorms.

[1] Comparing with (Lamba et al. 2015), we have excluded 'AskThicke' firestorm, because it has a gap of 24 h between T_2 and T_1; hence, we added 'suey_park' firestorm instead.



(a) Results of comparison

	I	we	you	shehe	they	posemo	negemo	emo	netspeak	assent
lower	15	8	11	15	8	16	5	18	7	2
same	0	3	0	0	5	2	1	2	1	1
higher	5	9	9	5	7	2	14	0	12	17

(b) Number of firestorms

Fig. 2 Comparison between firestorm-related tweets ($T1$) and non-firestorm tweets ($T2$ and $T3$) w.r.t various linguistic features using T-tests with p value < 0.01

The blue (turquoise) cells represent the firestorms in which terms from the respective category occurred more frequently during the firestorms. The red (brick) cells represent the firestorms in which the same words occurred less frequently during the firestorms. The light gray cells represent the firestorms in which there is no significant difference between firestorm tweets and non-firestorm tweets.

The results of comparison are aggregated in the table in Fig. 2b, which shows, for each feature, the number of firestorms according to the three cases of comparison: lower, higher and same (no significant difference).

Results For category 'I' this means that in five firestorms people used words of this category significantly more often, while in 15 firestorms these words were used significantly less. Similar results are observed for category 'she/he' In addition to the category 'I', the categories 'posemo' and 'negemo' should also be highlighted. Words representing positive emotions like 'love,' 'nice,' 'sweet'—the 'posemo' category—are used significantly less in almost all firestorms: positive emotions were less present in 16 out of 20 firestorms. For the category 'negemo,' which contains words representing negative emotions, this effect is reversed for all tweets—words in this category are used significantly more often during most of the firestorms (14 out of 20). There are 18 firestorms in which the 'emo' values were significantly lower during a firestorm. At the same time, there

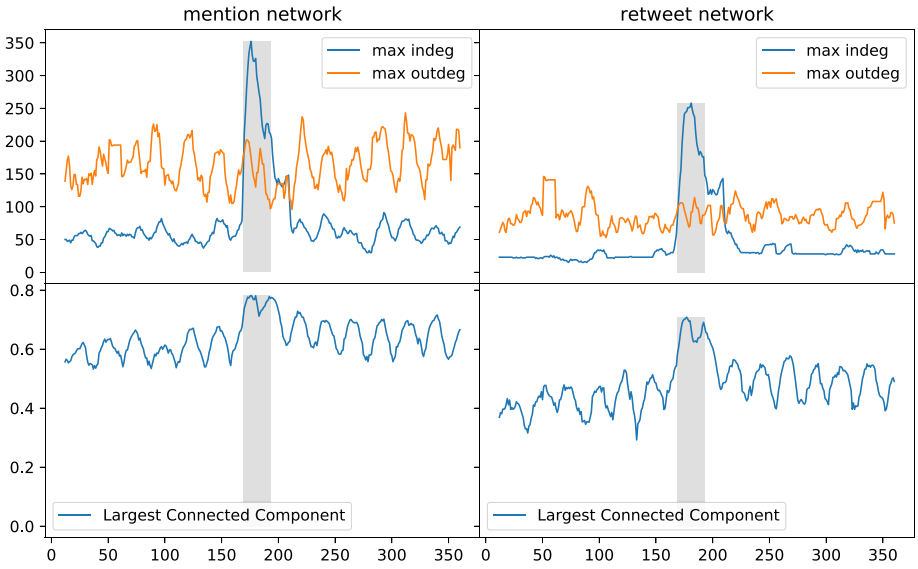


Fig. 3 Evolution of network features over time (#myNYPD firestorm). Highlighted area indicates the start of the firestorm (first 24 h)

were only two firestorms where the differences in the values of 'emo' were not significant. Another remarkable category is 'assent,' which contains words like 'agree,' 'OK,' 'yes.' In this category, the effect is also reversed—words in this category are used significantly more often during almost all firestorms (17 out of 20). *Interpretation.* We can state that during firestorms, the I vanishes and users talk significantly less about themselves compared to non-firestorm periods. Simultaneously, the positivity in firestorms tweets vanishes and negativity rises.

5 Mention and retweet networks

Besides linguistic features and sentiments expressed in tweets, online firestorms have also impact on the structure of user's social networks, such as mention and retweet networks.

To get insight on the evolution of each firestorm over time, we first split the time-line of each of the firestorm datasets into buckets of *one hour* and assign tweets to buckets based on their timestamp. The result of this splitting is a series of about 360 time slices (since the studied time-span of an event is 15 days). This allows us to perform analysis at fine granularity.

First, at each time slice, we extract several *basic features* of the corresponding hourly buckets of tweets, including:

- Number of tweets N_t
- Number of mention tweets N_{mt}
- Number of mentions N_m
- Ratio of mention tweets to all tweets N_{mt}/N_t.
- Mention per tweet ratio: N_m/N_t.

Moreover, at each time point we construct *mention networks*, and *retweet networks* taking into account all the tweets during the last 12 h. This way, we obtain a moving window of tweets: with a window size of 12 slices at steps of 1 h. The *mention network* of each moving window contains an edge ($user_1$, $user_2$) if a tweet (among tweets under consideration) posted by $user_1$ contains a mention to $user_2$. The *retweet network* of each moving window contains an edge ($user_1$, $user_2$) if a tweet (among tweets under consideration) posted by $user_1$ is a retweet of another (original) tweet posted by $user_2$.

For each event, the mention networks constructed at different time points are directed, unweighted networks. We performed several types of social network analysis and extracted a set of metrics, including:

- Number of nodes N and edges E,
- Average out-degree (which equals avg. in-degree).
- Maximum out-degree and maximum in-degree.
- Relative size of the largest connected component.

Each of the aforementioned features leads to a time-series when taken over the entire time-span of the event. For example, Fig. 3 depicts some of those time-series for the features of the mention and retweet networks of #myNYPD firestorm, showing how those features evolve over time. While network metrics are affected by sampled datasets, we still believe that these metrics are meaningful since the sampling process was consistent over all firestorms.

Results One can clearly observe the oscillating behavior of those features. This oscillation is due to the alternation of tweeting activity between daytime and night. More interesting observation is the manifest change of behavior that occurs near the middle of the time span, which evidently

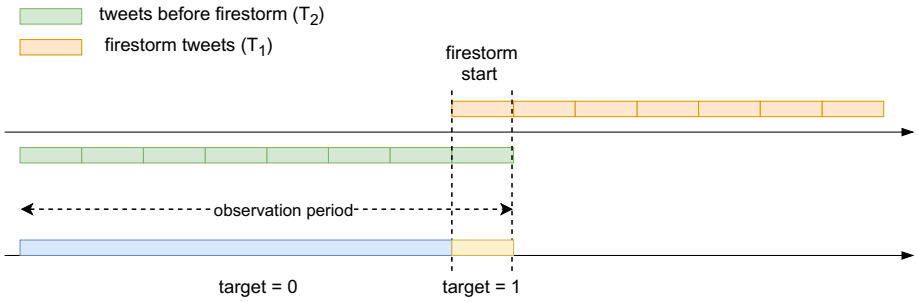


Fig. 4 Timeline of a firestorm

signals the beginning of the firestorm event. This apparent change can be observed in most of the features for the event at hand. However, not all the features are useful to detect the trigger of the firestorm in all events. In particular, we find the maximum in-degree feature is one of the best features to detect this change. This feature can clearly detect the start of the firestorm (in all events). The maximum in-degree in mention networks means the highest number of mentions received by a particular user.

Interpretation Thus, the ability of this feature to detect a firestorm can be interpreted by considering that, generally speaking, a firestorm occurs when one user is being mentioned unusually high. This result is intuitive since Tweets related to a certain Firestorm normally mention the victim's Twitter account.

Monitoring this feature in real-time would be certainly handy at detecting firestorms as early as possible, by signaling abnormal changes (increase) in this feature. However, the change of focus to a particular user can be the result of different (including positive) events.

From a network perspective, an online firestorm occurs when one user is mentioned unusually high, focusing on a Twitter handle or a hashtag. The maximum in-degree in @ mention networks is significantly deviating from comparable time periods.

6 Predicting the start of a firestorm

In the previous section we identified slight changes in lexical and sentimental cues as indicators of a firestorm. From a network perspective, we identified the maximum in-degree to be a very good indicator for a firestorm to occur. Based on these findings we want to test and compare our extracted features for a classification task in order to build models for predicting the start of a firestorm.

6.1 Prediction models (predictor variables)

As mentioned earlier, we split the time-line of each firestorm into buckets of *one hour* and assign tweets to buckets based on their timestamp.

Thus, for each time slice, the corresponding bucket of tweets is described by several features. Mainly, we distinguish between different types of features; each type of them defines a prediction model:

- *Baseline model* includes the basic features, such as number of tweets N_t, number of mentions N_m, etc. (see Sect. 5).
- *Mention-network model* includes network features, such as, number of nodes and edges, density, reciprocity, average and max in-degree and out-degree, etc., extracted from *mention* networks.
- *Retweet-network model* includes the same set of network features extracted from *retweet* networks.
- *Linguistic model* extends the basic model by including linguistic features, i.e., the mean values of extracted LIWC features (over the hourly bucket of tweets). In particular, we are interested in the following features: pronouns: namely: 'i,' 'we,' 'you,' 'shehe,' and 'they'; emotions: 'posemo,' 'negemo' and 'emo'; and 'netspeak' and 'assent.'

By doing so, we create separate time series for each of the features mentioned above.

6.2 Target variable

As shown in Fig. 4, the time span of the two sets of tweets T_2 and T_1 is 8 and 7 days, respectively, with an overlap of 1 day between the two periods. We consider the first day of the firestorm as its start. Hence, we create a *target* variable whose value is 0 for the time points t occurring entirely before the firestorm (the first 7 days of T_2) and 1 for the time points t occurring during the first day of the firestorm. The rest of the firestorm days are omitted. Hence, we obtain about $7 \times 24 = 168$ time points[2] where $target = 0$ (negative instances), as well as 24 points where $target = 1$ (positive instances).

[2] This number slightly varies from one firestorm to another.

Table 2 Pearson correlation of basic features, network features and linguistic features with the *target* variable (#myNYPD firestorm)

Basic features		Network features			Linguistic features	
			Mention	Retweet		
N_t	0.70	N	0.76	0.83	i	−0.16
N_{mn}	0.71	E	0.80	0.86	we	0.12
N_m	0.67	density	−0.61	−0.68	you	−0.15
N_{mt}/N_t	0.40	recip.	0.34	−0.38	she/he	−0.06
N_m/N_t	0.21	lwcc	0.82	0.85	they	0.35
		avg d_{in}	0.82	0.87	posemo	0.01
		max d_{in}	0.96	0.96	negemo	0.27
		max d_{out}	−0.00	0.11	emo	−0.23
					netspeak	0.56
					assent	0.19

Our objective is thus to predict the value of this *target* variable using the aforementioned sets of predictors. Hence, the prediction turns into a binary classification task, where we want to classify whether a time point t belongs to the period of firestorm start (target = 1) or not (belongs to the period before the firestorm: target = 0), using different types of features of the tweets. This classification task needs to be performed for each firestorm separately and independently from other firestorms.

6.3 Comparing features between before and the start of the firestorm

Before we dive deeper into the details of the classification task, it is interesting at this point to look at how different predictor features correlate with our target variable (which indicates the firestorm start). This would help us get insight on the ability of those features to *predict* that target variable. For this purpose, we calculate the Pearson correlation of each feature with the target variable (its numeric value 0 or 1). Table 2 shows the correlation values for the case of #myNYPD firestorm.

We can observe that basic features—in particular, number of tweets N_t, number of mention tweets N_{mt} and number of mentions N_m—have a relatively strong positive correlation with the target variable.

This effect of strong positive correlation can be also observed for most of network features, such as number of nodes N and edges E, relative size of largest (weakly) connected component *lwcc*, avg. and max. in-degree. In contrast, density has a strong negative correlation, which means that this feature is lower at the start of the firestorm compared to before the firestorm. On the other hand, reciprocity has rather a weak correlation with the target variable; this correlation is positive for mention networks (+0.34) and negative for retweet networks (-0.38). Finally, max d_{out}, the maximum out-degree, has no correlation at all.

Regarding linguistic features, most of those features have weak correlation (positive or negative), or no correlation with the target variable. The highest correlations are for 'netspeak' (0.56) and 'they' (0.35).

6.4 Design of the classification task

6.4.1 Split into training and test sets

As in any supervised machine learning task, data instances need to be split into training and test subsets: the first is used to train the classifier while the other is used to test it, i.e., to evaluate its performance. Typically, such splitting of the dataset is performed in a *random* fashion, with, for example, 75% of instances for training and the remaining 25% for testing. Moreover, in order to make a more reliable evaluation, a *cross validation* approach is typically used, such as the k-folds method. In k-folds cross-validation, the dataset is split into k consecutive folds, and each fold is then used once as a validation while the $k-1$ remaining folds form the training set. This method generally results in a less biased model compared to other methods, because it ensures that every observation from the original dataset has the chance of appearing in the training and test set.

However, in our firestorm dataset(s), positive and negative classes are highly *unbalanced*, with a 1:7 ratio, i.e., for each positive instance there are 7 negative instances. To tackle this unbalanced issue, we use *stratified k-folds*, which is a variation in k-folds cross-validation that returns *stratified* folds, that is, the folds are made by preserving the percentage of samples for each class.

In this study, we opt to use $k = 4$, and the dataset is split hence into 4 stratified folds. Thus, when the dataset contains 24 positive samples, and 168 negative ones, then each fold will contain $24/4 = 6$ positive samples, and about $168/4 = 42$ negative ones. The training is also performed 4 times, each time one of the folds is used as a test set while the remaining 3 folds are used as a training set.

This means that, each time, the 24 positive instances will be distributed such that 6 instances will be in the test set and 18 instances in the training set. This approach avoids the undesired situations where the training is performed with very few or with too many positive instances. The overall evaluation score is calculated as the average over the 4 training times.

6.4.2 Feature scaling

In our case, different features have their values on very different scales. For instance, regarding network features, the number of nodes N and edges E are usually $> 10^3$, while density is $< 10^{-3}$ and reciprocity is $< 10^{-2}$. Thus, in order to improve the prediction accuracy, we need to avoid some low-scale features being overwhelmed by other high-scale ones; therefore, we use feature scaling in order to put the features roughly on the same scale.

We use the *standard scaling* approach, where each feature is standardized by centering and scaling to unit variance. The standard score of a sample x is calculated as: $z = (x - \mu(x))/\sigma(x)$ where μ is the mean of the samples, and σ is the standard deviation of the samples.

Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. Mean and standard deviation are then stored to be used on later data using transform. Standardization of a dataset is a common requirement for many machine learning algorithms, as they might behave badly if the individual features do not roughly look like standard normally distributed data (e.g., Gaussian with 0 mean and unit variance).

6.4.3 Algorithm

As a classification algorithm, we used the logistic regression algorithm. Logistic regression is a well-known and widely used classification algorithm which extends linear regression. Instead of fitting a straight line or hyperplane, the logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1. The logistic function is defined as: $\sigma(x) = 1/(1 + \exp(-x))$

6.4.4 Evaluation

As an evaluation measure, we used Accuracy, which is simply the fraction of correctly classified instances (to all instances). For each firestorm, the prediction accuracy is calculated as the average of the accuracy over the 4 folds.

6.5 Results

We applied the logistic regression algorithm to each firestorm using different prediction models: basic model, mention network model and linguistic model. Table 3 shows the overall accuracy for each firestorm, with respect to each prediction model. We can see that the prediction accuracy is pretty high in general where the accuracy is within the range of 87% to 100%.

Table 3 Accuracy of prediction models

	Basic	Linguistic	Mention	Retweet
askbg	0.926	0.916	0.958	0.953
askjpm	0.953	0.953	0.995	0.995
cancelcolbert	0.948	0.953	0.990	0.984
celebboutique	0.915	0.945	0.937	0.963
david_cameron	1.000	0.995	0.995	0.995
fafsa	0.932	0.943	0.989	0.989
gaelgarciab	0.906	0.885	0.956	0.956
klm	0.891	0.902	0.907	0.907
mcdstories	0.943	0.938	0.923	0.961
muslimrage	0.956	0.990	0.980	0.969
mynypd	0.958	0.984	1.000	0.995
notintendedto.	0.990	0.984	0.989	0.989
qantas	0.922	0.922	0.939	0.956
qantasluxury	0.932	0.943	0.972	0.972
spaghettios	0.944	0.964	0.989	0.989
suey_park	0.943	0.948	0.990	0.995
theonion	0.974	0.974	0.989	0.989
ukinusa	0.870	0.875	0.995	0.995
voguearticles	0.951	0.967	0.971	0.977
whyimvotingukip	0.943	0.969	0.956	0.950
avg.	0.940	0.948	0.971	0.974

For the basic model, the accuracy ranges between 87% (for 'ukinusa') and 100% (for 'david_cameron'), with an average of 94%. For the linguistic model, the accuracy ranges between about 87% (e.g., 'ukinusa') and 99.5% (@David_Cameron), with an average of 95%.

Finally, the two network models, mention and retweet, show very similar results in general. The accuracy ranges between about 90% (klm) and 100% (myNYPD), with an average of 97%. Overall we can see that all the prediction models are able to predict the start of the firestorm with very high accuracy.

Interpretation Network models are slightly more accurate than the linguistic model, which is in turn slightly more accurate than the basic model. It is logical that in times of firestorms there are a lot of mentions, hashtags and retweets, i.e., explicit network properties. Even more important and interesting is the result that we can measure early changes already in the language and that these properties are much more important for the early detection of changes. The fact that we make a comparison here should illustrate how well our model works alongside other more explicit models.

7 Conclusion

Our goal was to predict the outbreak of a firestorm using linguistic and network-based features. Therefore, we examined the vocabulary of tweets from a diverse set of firestorms and compared it to non-firestorm tweets posted by the same users. Additionally, we measured features describing the mention and retweet networks also comparing firestorm with non-firestorm tweets. We used the features in a logistic regression model to predict the outbreak of firestorms. The identified linguistic and sentimental changes were good indicators for the outbreak of a firestorm.

Observing linguistic features, we found that during firestorms users talk significantly less about themselves compared to non-firestorm periods which manifested in significantly fewer occurrences of self-referencing pronouns like 'I,' 'me' and the like. Simultaneously, the positivity in firestorm tweets vanishes and negativity rises. Especially the change in the use of personal pronouns served as a good indicator for the outbreak of online firestorms. This change of subject to a different object of discussion could be observed in an increased mentioning of a user or a hashtag who/that was the target of a firestorm, hence the perspective changes. Users start pointing at others. This expressed itself in a maximum in-degree in mention networks that significantly deviated from comparable time periods giving evidence for the pragmatic action from a network perspective. However, we are aware of the fact that we have only measured cases in which the in-degree change happens in the context of something negative.

Our models were able to predict the outbreak of a firestorm accurately. We were able to classify the outbreak of a firestorm with high accuracy (above 87%) in all scenarios. It showed, however, that classification models using features derived from the mention and retweet networks performed slightly better than models based on linguistic features.

Overall, verbal interaction is a social process and linguistic phenomena are analyzable both within the context of language itself and in the broader context of social behavior (Gumperz 1968). From a linguistic perspective, the results give an idea of how people interact with one another. For this purpose, it was important to understand both the network and the speech acts. Changes in the linguistic and sentimental characteristics of the tweets thus proved to be early indicators of change in the parts of social media networks studied. Besides the fact that users changed their perspective, we could also observe that positivity in words vanished and negativity increased.

Future work could consider clustering firestorms according to their dynamics, i.e., can firestorms be differentiated in the way users ally against a target? This is of interest insofar as we know that negative PR can also mean profit for a company and that this is seen as less bad. Another pathway worth following would be to leverage contextualized word embeddings (Peters et al. 2018) to identify especially harmful words that demand early attention. Generally, the question of what motivates people to ally against a target is of great scientific and social interest.

Our results give insights about how negative word-of-mouth dynamics on social media evolve and how people speak when forming an outrage collectively. Our work contributed to the task of predicting outbreaks of firestorms. Knowing where a firestorm is likely to occur can help, for example, platform moderators to know where an intervention in a calming manner will be required. Ultimately, this can save individuals from being harassed and insulted in online social networks.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The lists of Tweet IDs of the analyzed data can be shared upon request.

Declarations

Conflict of interests The author(s) gratefully acknowledge the financial support from the Technical University of Munich—Institute for Ethics in Artificial Intelligence (IEAI). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the IEAI or its partners.

References

Allport G, Postman L (1947) The psychology of rumor. J Clin Psychol 3(4):402

Anderson C (2006) The long tail: why the future of business is selling less of more. Hyperion Books, New York

Appel G, Grewal L, Hadi R, Stephen AT (2020) The future of social media in marketing. J Acad Mark Sci 48(1):79–95

Auger IE, Lawrence CE (1989) Algorithms for the optimal identification of segment neighborhoods. Bull Math Biol 51(1):39–54

Baayen H (1993) Statistical models for frequency distributions: a linguistic evaluation. Comput Humanit 26:347–363

Bail CA, Brown TW, Mann M (2017) Channeling hearts and minds: advocacy organizations, cognitive-emotional currents, and public conversation. Am Sociol Rev 82(6):1188–1213
Brady WJ, Wills JA, Jost JT, Tucker JA, Bavel JJV (2017) Emotion shapes the diffusion of moralized content in social networks. PNAS 114(28):7313–7318
Bybee J, Hopper P (2001) Frequency and the emergence of linguistic structure. In: Bybee J, Hopper P (eds) Typological studies in language, vol 45. John Benjamins Publishing Company, Amsterdam, pp 1–24
Cartwright D, Harary F (1956) Structural balance: a generalization of Heider's theory. Psychol Rev 63(5):277–293
Chadwick A (2017) The hybrid media system: politics and power, 2nd edn. Oxford University Press, Oxford
Charlton N, Singleton C, Greetham DV (2016) In the mood: the dynamics of collective sentiments on Twitter. R Soc Open Sci 3(6):160162
Crystal D (2002) Language and the internet. IEEE Trans Prof Commun 45:142–144
Dandekar P, Goel A, Lee DT (2013) Biased assimilation, homophily, and the dynamics of polarization. Proc Natl Acad Sci 110(15):5791–5796
Delgado-Ballester E, López-López I, Bernal-Palazón A (2021) Why do people initiate an online firestorm? the role of sadness, anger, and dislike. Int J Electron Commer 25:313–337
Drasch B, Huber J, Panz S, Probst F (2015) Detecting online firestorms in social media. In: ICIS
Ferrara E, Yang Z (2015) Measuring emotional contagion in social media. PLOS ONE 10(11):e0142390
Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning. Springer series in statistics. Springer, New York
Goel S, Hofman J, Lahaie S, Pennock D, Watts D (2010) Predicting consumer behavior with Web search. In: Proceedings of the National Academy of Sciences. National Academy of Sciences Section: Physical Sciences, pp 17486–17490
Gonzales AL, Hancock JT, Pennebaker JW (2010) Language style matching as a predictor of social dynamics in small groups. Commun Res 37(1):3–19
Gumperz J (1968) The speech community. In: Duranti A (ed) Linguistic anthropology: a reader. Wiley, New York, pp 166–173
Hauser F, Hautz J, Hutter K, Füller J (2017) Firestorms: modeling conflict diffusion and management strategies in online communities. J Strat Inf Syst 26(4):285–321
Heider F (1946) Attitudes and cognitive organization. J Psychol 21:107–112
Hennig M, Brandes U, Pfeffer J, Mergel I (2012) Studying social networks. A guide to empirical research. Campus Verlag, Frankfurt
Herhausen D, Ludwig S, Grewal D, Wulf J, Schoegel M (2019) Detecting, preventing, and mitigating online firestorms in brand communities. J Mark 83(3):1–21
Jackson B, Scargle JD, Barnes D, Arabhi S, Alt A, Gioumousis P, Gwin E, Sangtrakulcharoen P, Tan L, Tsai TT (2005) An algorithm for optimal partitioning of data on an interval. IEEE Signal Process Lett 12(2):105–108
James G, Witten D, Hastie T, Tibshirani R (2014) An introduction to statistical learning. Springer, Cham
Johnen M, Jungblut M, Ziegele M (2018) The digital outcry: what incites participation behavior in an online firestorm? New Media Soc 20(9):3140–3160
Killick R, Fearnhead P, Eckley IA (2012) Optimal detection of changepoints with a linear computational cost. J Am Stat Assoc 107(500):1590–1598
Lamba H, Malik MM, Pfeffer J (2015) A tempest in a teacup? Analyzing firestorms on Twitter. In: 2015 IEEE/ACM ASONAM. New York, NY, USA, pp 17–24
Leskovec J, Rajaraman A, Ullman JD (2014) Mining of massive datasets, 2nd edn. Cambridge University Press, Cambridge
McCulloh I, Carley KM (2011) Detecting change in longitudinal social networks. J Soc Struct 12(3):1–37
McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Ann Rev Sociol 27(1):415–444
Mochalova A, Nanopoulos A (2014) Restricting the spread of firestorms in social networks. In: ECIS 2014 proceedings
Morstatter F, Pfeffer J, Liu H, Carley K (2013) Is the sample good enough? comparing data from Twitter's streaming API with Twitter's firehose. In: Proceedings of the international AAAI conference on web and social media, vol 7, no 1
Myers S, Zhu C, Leskovec J (2012) Information diffusion and external influence in networks. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, pp 33–41
Newman M (2010) Networks: an introduction. Oxford University Press Inc, New York
Newman M, Barabási AL, Watts DJ (2006) The structure and dynamics of networks. Princeton University Press, Princeton
Pariser E (2011) The filter bubble. What the internet is hiding from you. The New York Press, New York
Pennebaker JW, Boyd RL, Jordan K, Blackburn K (2015) The development and psychometric properties of LIWC2015. Technical report, The University of Texas at Austin
Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: NAACL-HLT 2018, pp 2227–2237
Pfeffer J, Zorbach T, Carley KM (2014) Understanding online firestorms: negative word-of-mouth dynamics in social media networks. J Mark Commun 20(1–2):117–128
Rost K, Stahel L, Frey BS (2016) Digital social norm enforcement: online firestorms in social media. PLoS ONE 11(6):e0155923
Ruths D, Pfeffer J (2014) Social media for large studies of behavior. Science 346:1063–1064
Scott K (2015) The pragmatics of hashtags: inference and conversational style on Twitter. J Pragmat 81:8–20
Scott AJ, Knott M (1974) A cluster analysis method for grouping means in the analysis of variance. Biometrics 30:507–512
Sen A, Srivastava MS (1975) On tests for detecting change in mean. Ann Stat 3(1):98–108
Shaikh S, Feldman LB, Barach E, Marzouki Y (2017) Tweet sentiment analysis with pronoun choice reveals online community dynamics in response to crisis events. In: Advances in cross-cultural decision making. Springer, pp 345–356
Sharp WG, Hargrove DS (2004) Emotional expression and modality: an analysis of affective arousal and linguistic output in a computer vs. paper paradigm. Comput Hum Behav 20(4):461–475
Snijders TAB (2001) The statistical evaluation of social network dynamics. Sociol Methodol 31(1):361–395
Snijders TA, Koskinen J, Schweinberger M (2010) Maximum likelihood estimation for social network dynamics. Ann Appl Stat 4(2):567–588
Stich L, Golla G, Nanopoulos A (2014) Modelling the spread of negative word-of-mouth in online social networks. J Decis Syst 23(2):203–221
Strathern W, Schönfeld M, Ghawi R, Pfeffer J (2020) Against the others! Detecting moral outrage in social media networks. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp 322–326
Strathern W, Ghawi R, Pfeffer J (2021) Advanced statistical analysis of large-scale web-based data. In: Data science in economics and finance for decision makers. Edited by Per Nymand-Andersen
Tausczik YR, Pennebaker JW (2009) The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol 29:24–54

Udupa S (2020) Artificial intelligence and the cultural problem of extreme speech. Social Science Research Council (20 December 2020)

Vermeer S, Araujo T, Bernritter S, Noort G (2019) Seeing the wood for the trees: How machine learning can help firms in identifying relevant electronic word-of-mouth in social media. Int J Res Mark 36:492–508

Vespignani A (2009) Predicting the behavior of techno-social systems. Science 325:425–428

Wagner C, Singer P, Karimi F, Pfeffer J, Strohmaier M (2017) Sampling from social networks with attributes. In: Proceedings of the WWW conference, pp 1181–1190

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge, MA

5.4 Polarization and Radicalization

5.4.1 The Polarizing Impact of Continuous Presence on User's Behavior

This peer-reviewed paper is relevant for examination.

Authors Wienke Strathern, Angelina Mooseder, Jürgen Pfeffer

In Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media (ICWSM 2022), Atlanta, Georgia, June 6-9, https://doi.org/10.36190/2022.52.

Abstract

Online political communities offer spaces where people share and exchange political views, content, and ideas. However, people seeking political exchanges online are increasingly confronted with an exclusionary intensification of discussion that no longer makes it possible to participate in constructive discourse. Online political discussion forums like r/The Donald and r/ChapoTrapHouse on Reddit have been banned recently due to the proliferation of hate speech and anti-social behavior of users. Homogeneous online discussion communities have been shown to play a key role in political polarization. Spending time in these communities tends to foster strong political positions associated with conflict. In this study we analyzed social media data from online political communities on Reddit and asked: how does presence in polarizing environments influence users' behavior? With initially equal user activity, what happens if one group of users continues to be present while the other is not? Our analysis shows that continuously present users become more active, use simpler vocabulary, and employ more abusive words in their text contributions. Our results have implications for automated moderation of polarizing online communities.

Publication Summary

Questions: The constant flow of information on social media can expose users to a wide range of topics, leading to potential polarization. In Strathern et al. (2022a) we aimed to explore the linguistic effects of this exposure and understand how it affects users' behavior and writing style. Specifically, we are interested in observing the behavior of one group of users in the absence of another group with the same level of activity. Our study examines how users' activity and language change under these conditions and whether these changes are influenced by the absence of users. Additionally, we seek to provide recommendations for automated content moderation based on our findings.

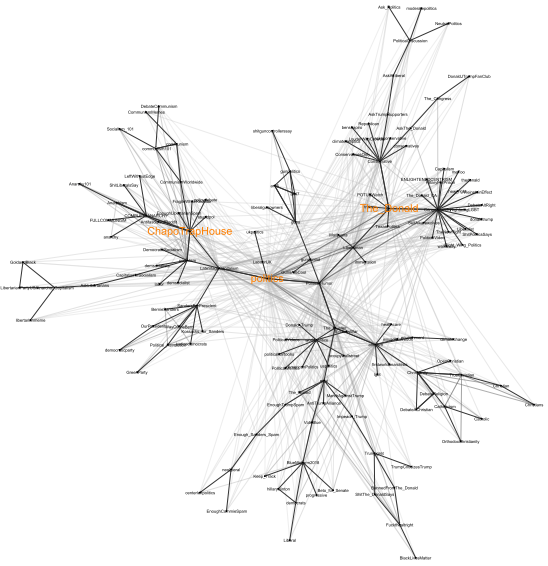


Figure 8: Subreddit network according to Strathern et al. (2022a)

Data: We used the Pushift dataset and collected posts and comments from 146 political subreddits from January 2018 to August 2019. We selected the three subreddits with the highest posting activity for our analysis: The left-wing community r/ChapoTrapHouse, the right-wing community r/The Donald, and r/politics. An analysis of user co-occurrence within political subreddits affirms the structural positions of the subreddits illustrated in Figure 8. **Methods and Analysis:** Figure 9 illustrates the experimental design developed in our study. Our goal was to examine the effects of continuous activity in online communities and to measure differences in behavior that result from an absence. To accomplish this, we identified two groups with similar activity levels at the start of our study. The first group, referred to as the "present users," remained

active throughout the analysis. The second group, the "absent users," refrained from any activity for at least 14 days. We then compared the behavioral changes of both groups between time frame A and time frame C to determine the impact of activity on their behavior.

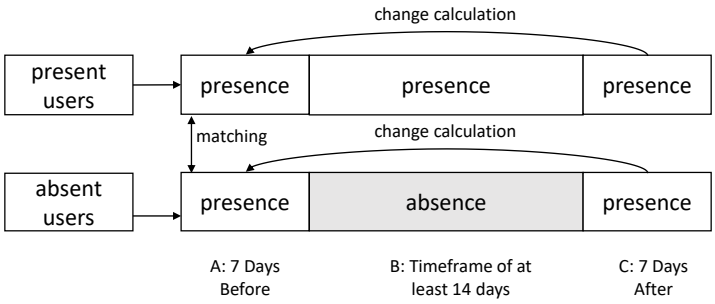


Figure 9: Overview of experimental setup according to Strathern et al. (2022a)

Results and Interpretation: Figure 10 displays each group's interquartile range of activity differences, with the median represented by the middle line. Our findings indicate that continuously present individuals increase their activity and contribute more to discussions. However, they also tend to exhibit a decrease in lexical diversity and use simpler vocabulary. Additionally, we observed an increase in the use of profanity and offensive language among continuously present individuals.

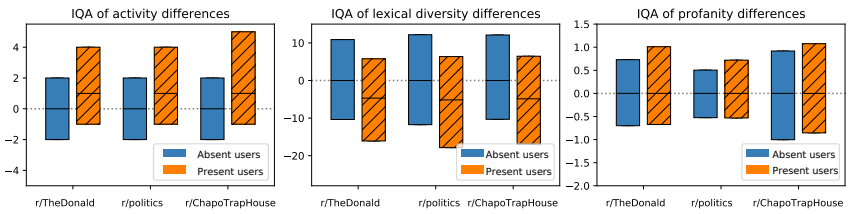


Figure 10: Interquartile range of activity, lexical diversity and profanity differences as in Strathern et al. (2022a)

Notably, those who are more active in general show more significant changes in behavior regarding activity and lexical diversity. Our study highlights the polarizing impact of continuous presence in these online environments. It prompts a discussion on how information consumption can be disrupted and whether non-consumption can impact behavior.

Author Contribution

Wienke Strathern headed the project, developed the conceptual framework and research question, developed the experimental design, conducted and wrote the literature review, chose the metrics, wrote the introduction, wrote the conclusion and discussion, manuscript writing, revisions, editing, coordinated the team. Wienke Strathern was responsible for the overall manuscript.

The Polarizing Impact of Continuous Presence on Users' Behavior

Wienke Strathern,[1] **Angelina Mooseder,**[1] **Jürgen Pfeffer**[1]
[1]School of Social Sciences and Technology, Technical University of Munich
wienke.strathern@tum.de, angelina.mooseder@tum.de, juergen.pfeffer@tum.de

Abstract

Online political communities offer spaces where people share and exchange political views, content, and ideas. However, people seeking political exchange online are increasingly confronted with an exclusionary intensification of discussion that no longer makes it possible to participate in constructive discourse. Online political discussion forums like r/The Donald and r/ChapoTrapHouse on Reddit have been banned recently due to the proliferation of hate speech and anti-social behavior of users. Homogeneous online discussion communities have been shown to play a key role in political polarization. Spending time in these communities tends to foster strong political positions associated with conflict. In this study we analyzed social media data from online political communities on Reddit and asked: how does presence in polarizing environments influence users' behavior? With initially equal user activity, what happens if one group of users continues to be present while the other is not? Our analysis shows that continuously present users become more active, use simpler vocabulary, and employ more abusive words in their text contributions. Our results have implications for automated moderation of polarizing online communities.

1 Introduction

Millions of users actively participate in online communities and social media networks offering a platform for exchanging views on any political topic. The platforms make it possible for people to network and find a community that matches their interests, convictions and political views. By offering them a space to exchange opinions, social media operators had high hopes for these platforms. However, negative online effects such as hate speech, anti-social behavior by users and offline spillover effects such as the storming of the US Capitol in January 2021, which were fomented on politically active social media forums are studied (Gallacher, Heerdink, and Hewstone 2021). The often unrestricted and anonymous environment of online discussions can become a platform for anti-social behavior, such as online abuse or harassment (Walther 2022). The drifting apart of political opinions and attitudes on platforms has been much discussed (Kubin and von Sikorski 2021). The fact that people spend inordinate amounts of time in their close-knit social media networks—sometimes referred to as filter bubbles (Pariser 2011)—is seen as a central problem in the study of polarization (Sunstein 2009). Strong political opinions are associated with conflict, misinformation and a reluctance to engage with people and perspectives beyond one's (often narrow) own views. Given the high-speed communication of social media, its polarizing effects are even stronger.

Research Questions Reddit is a community driven discussion platform for political and polarizing discourse (Kane and Luo 2019; Phadke, Samory, and Mitra 2021). In a variety of political subreddits like-minded people come together to discuss current political topics. Given its openness and speed of communication with unrestricted ways of expression, discussions can be polarizing, heated, engaging, controversial, productive, creative, diverse in their facets. Due to its speed, its unrestricted flow of information and the constantly high number of contributions, users, if present, are continuously exposed to a variety of topics. Almost half of the Reddit users are heavy Internet users[1], studies show that even short time effects can give first initial insights on changing behavior (Zhou et al. 2021). The question arises as to what presence in a polarizing environment does to people linguistically? Does it change the way users behave? Does it change the way users write? Regarding theories on polarization and radicalization in political settings (Grover and Mark 2019) we assume that prolonged dwelling in these polarizing environments affects users' behavior in terms of activity, thus interactions. Furthermore, if we assume that polarizing environments promote homogeneity, we expect this to be reflected in linguistic patterns of complexity. With respect to studies of online radicalization in polarizing environments, we assume that people show differences in behavior and that it is reflected in their textual contributions.

Methods To answer our research question methodologically, we ask: given the same level of activity, what happens if one group of users continues to be present while the other is not? In order to assess differences in behavior we measure interactions and text contributions, apply a linguistic metric

[1]https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

to determine the level of lexical diversity and apply a metric to determine the level of profane word usage in users' posts and comments.

To test our hypothesis we analyzed social media data of three famous online political communities on Reddit. We used data from three main political subreddits alongside the political spectrum to test our hypothesis: r/The_Donald, a right-wing subreddit, r/politics, a forum for all users interested in politics, and r/ChapoTrapHouse, a left-wing subreddit. The explosive nature of the topic is shown by the fact that two of the subreddits we investigated, r/ChapoTrapHouse and r/The_Donald, were banned, i.e., the discussion site was dissolved, due to hate speech, anti-social behavior, and violations against Reddits hate policy [2] [3]. The user communities for these subreddits generated an extensive chat history of roughly 17 million posts and comments. Since these are structurally large communities, we want to know whether our hypothesis applies to both the right and the left, as well as to the moderate spectrum.

We compared the behavior of users who were continuously active to the behavior of similar users who were non-active for at least 14 days to determine 1) whether users change in activity and language and 2) whether this change is correlated with absence/non-activity. To answer our research question, we a) defined groups and b) measured posting and commenting activity for each group, the level of lexical diversity and profane word usage in their text comments. By comparing two groups and their linguistic attributes we were able to extract data-driven insights about the users' activity and linguistic changes in three online political communities. In addition, we took advantage of naturally occurring variation in the degree of overall activity of users—i.e., how frequently users posted over the time period of six months—to analyze whether behavioral changes were stronger in people who are generally more engaged in a subreddit.

Contributions. Based on our analysis, we derive the following conclusions:

1. Continuously present people increase their activity, show more contributions.
2. Continuously present people decrease their lexical diversity, use simpler vocabulary.
3. Continuously present people increase their profanity, use more offensive language.
4. Continuously present users who are in general more active show greater behavioral changes regarding activity and lexical diversity.

To summarize, we observe that users who are continuously present on Reddit change in activity, vocabulary and emotions, while users who are inactive for a certain time show the same level of activity, vocabulary and emotions after their absence as before.

[2] https://en.wikipedia.org/wiki/Controversial_Reddit_communities
[3] https://en.wikipedia.org/wiki/R/The_Donald

2 Related Work

Our study is motivated by work discussing polarizing environments on social media (2.1) measuring activity and language (2.2) to assess psycho-linguistic behavioral attributes.

2.1 Polarizing Online Environments

The unrestricted way of freedom of expression in online environments can lead to hate speech and discrimination of marginalized groups, hence causing polarizing effects. A substantial body of work explores ideologically cross-cutting discussion spaces and the role of non-political spaces for political discussion. In this work the authors examine the fact that online news consumption follows a polarized pattern, observing that users' visits to news sources aligned with their own political leaning are substantially longer than their visits to other news sources (Garimella et al. 2021). Another study focused on identifying high and low consensus news posts on Twitter and presenting a method to automatically detect them (Babaei et al. 2018). In another study cross-cutting posting was tested. This determined that people are more likely to accept a news article containing conflicting views when it is delivered by a chatbot (Zarouali et al. 2021). A focus of this work was on news-link sharing. It shows that Reddit users' voting and re-sharing behaviors generally decrease the visibility of extremely biased and low factual content (Weld, Glenski, and Althoff 2021). In this paper the authors show that political conversations are less toxic in non-political subreddits (Rajadesingan, Budak, and Resnick 2021).

Several works study the effects of social media use. For example, they examine the temporary lapse from social media platforms on behavior (Allcott et al. 2020; Kovacs, Wu, and Bernstein 2021; Brown 2020) and the effects of how to make online spaces more civil (Wadden et al. 2021). Reddit has become a famous community discussion platform for political discourse. Reddit data is used to measure similarity in the commenting user bases of communities (Mamié, Horta Ribeiro, and West 2021). Another study discovers language biases encoded in the vocabulary of on- line discourse communities on Reddit (Ferrer et al. 2021). A study was conducted on analyzing online news sharing at scale to study bias a factual news on Reddit (Weld, Glenski, and Althoff 2021). This work examines the social makeup of online communities to understand the social organization of online platforms on Reddit (Waller and Anderson 2021).

2.2 Measuring activity and linguistic patterns

User behavior plays an important role in understanding social media platform effects. In their analysis (Jhaver et al. 2019) the authors characterized the removal explanations that are provided to Redditors and link them to measures of subsequent user behaviors. In order to better understand political engagement, this study analyzes the political interaction network on Reddit contradicting the echo chamber narrative (De Francisci Morales, Monti, and Starnini 2021). A further paper addresses the social effects of content ratings on Reddit (Davis and Graham 2021). Another study observed Redditors behavior in terms of how they interact with

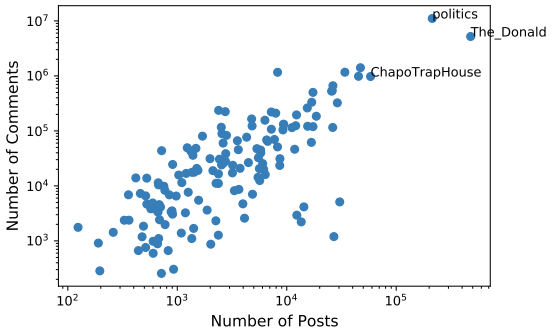


Figure 1: Correlation of number of posts and comments of 146 political subreddits

health-related messages (Silberman and Record 2021). Language, more precisely **linguistic attributes** play a key role in examining people's behavior. Some studies have used language to record psychological stress in social media (Guntuku et al. 2019). In a work exploring the nature of political discussions in homogeneous and cross-cutting communication spaces based on interaction and linguistic patterns analysis, different behavioral patterns in homogeneous and cross-cutting communications spaces are revealed (An et al. 2019). Lexical items can be used to detect early changes in behavior (Strathern et al. 2020). In order to gain insights on how people feel more comfortable online, a study measured the effect of moderation on health in online conversations based on linguistic features (Wadden et al. 2021).

Lexical diversity plays a key role in measuring complexity. It reflects the complexity of vocabulary knowledge as well as the level of language proficiency. Many indices of lexical diversity have been proposed, most of which involve statistical relationships between types and tokens and which ultimately reflect the rate of word repetition. Type-Token-Ratio (TTR) provides insights into the vocabulary knowledge. This is a syntactical index that divides the number of distinct words (types) by the total number of words (tokens), computed as a running average based on consecutive 1000-word blocks of text. High values indicate texts with a heterogeneous vocabulary and linguistic structure of moderate complexity, whereas a low level of lexical diversity indicates simple terminology and little complexity (Jarvis 2013).

Profanity is a socially offensive use of language, its main function being to express emotions, especially anger and frustration. The purpose is to express the speaker's emotional state and communicate that information to listeners (Jay and Janschewitz 2008). We apply a usage-based approach to study linguistic properties and functions assuming that frequency distributions in the contributions of individual users can, under certain circumstances, be interpreted regarding the writers' underlying mental representations (Bowers and Pleydell-Pearce 2011; Schmid et al. 2021).

3 Data

We describe the dynamics of three political subreddits, the dataset studied in this work, and basic preprocessing steps to filter out low-quality data.

In its structure and function, Reddit [4] contains elements of a discussion forum, a social network, and a news service. Users submit posts, called submissions. These can be original content, links to external content, or a combination of both. Much of the content is linked to other websites. Other users can then add comments to a post. In addition to commenting on posts, users can also rate posts and comments by upvoting them (meaning they are worthy of being seen by others) or downvoting them (meaning they should not be seen). This voting controls the display of posts and comments on the website. Overall, Reddit is divided into subreddits, which are roughly equivalent to forums or topics on other online message boards. The names of these subreddits usually describe the topic discussed (e.g., r/politics). A Pew Research Center publication gives information on the demographics of Reddit users: While only 4% of U.S. adults report using Reddit, about seven in ten of those users (78%) obtain news from the site. Overall, 2% of U.S. adults obtain news on Reddit. Both Reddit users in general and those who source news from the site tend to be young, male, and more likely to describe themselves as more liberal than the general population. About seven in ten (71%) of Reddit news users are men, 59% are between 18 and 29 years old, and 47% describe themselves as liberal, while only 13% are conservative (39% describe themselves as moderate). By comparison, of all adults in the U.S., about half (49%) are men, only 22% are between the ages of 18 and 29, and about a quarter (24%) describe themselves as liberal. Reddit news users are also heavy Internet users, with 47% reporting being online almost constantly (compared to 21% of U.S. adults overall). In this study we are interested in polarizing open online discussion forums. We would like to understand interactions and linguistic patterns. The platform offers an open space, open access and freedom of speech. There is no hierarchy – we find polarizing environments and user affiliation. Furthermore, we find informal language, a huge amount of contributions, few rules, colloquial language, and anonymity which allows for provoking behavior. Thus, the dataset is appropriate for the study of online user behavior. [5]

3.1 Dataset description

Using the Pushift Dataset (Baumgartner et al. 2020), we collected posts and comments from 146 political subreddits from January 2018 to August 2019. Fig. 1 shows the number of posts and number of comments for these subreddits (Pearson's $r = 0.75$, $p = 0.0$), with both axes being log-scaled. We choose the three subreddits with the highest posting activity for our analysis: The left-wing community r/ChapoTrapHouse, the right-wing community

[4] https://www.redditinc.com/

[5] https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

r/The_Donald, as well as r/politics, which can be described as a meeting place for politically interested users.

Co-occurrence network. An additional co-occurrence analysis of users in the political subreddits confirms the structural positions of these three subreddits. Fig. 2 is based on the original two-mode network data of a user being connected to a subreddit in case the user had posted in a particular subreddit. The subsequent transformation into a one-mode network of subreddits that are connected by shared users was then reduced for the visualization to show only the top five connections for every subreddit. The use of Visone's (Baur et al. 2002) backbone visualization procedure can reveal the different political areas in Reddit's political discussion forums. r/The_Donald can be found in an area of the network dominated by conservative subreddits and r/ChapoTrapHouse is densely connected through shared users with a set of liberal subreddits. The central position of r/politics results from the fact that this forum is populated by users from a wide range of political views.

Preprocessing. For each of these subreddits, we stored all comments and posts and removed all duplicated posts, as well as posts where the author information was missing or where the contribution was created by the AutoModerator instead of a real person. While posts could include a title as well as text, comments only included text. For each post, we combined the title and text into a text field and saved the author, the day of creation and the combined text. For each comment, we saved the author, the day of creation and the text. In the following analysis we did not differentiate between comments and posts but regarded them as objects from the same type (namely contribution). For the purpose of this study, we have limited the time frame to six months and used the Reddit data from January 1 to June 30 of 2018. A brief overview of the remaining data is given in Table 1.

subreddit	Users	Comments	Posts
r\The_Donald	126,217	5,204,877	478,614
r\politics	368,056	11,081,988	213,484
r\ChapoTrapHouse	18,719	974,558	58,286

Table 1: Dataset details

4 Study Design

Our study is motivated by the assumption that continuously present users in polarizing environments change their behavior. Continuous interactions make the user more exposed to the existing community and less to other influences, thus being repeatedly exposed to the same opinions which reinforces existing behavior and enhances further activity. We define presence as active interactions in terms of posts and comments contributions, absence in turn is defined as no activity at all for at least 14 days. Based on these assumptions we compare the behavior of present users with the behavior of similar users with absence to determine 1) whether users change with respect to their activity and language and 2) whether this change is correlated with absence.

4.1 Setup

The setup of this study is depicted in Fig. 3. We have built two groups showing a similar activity level in the beginning of our study (see time frame A). While the group of present users is present during the whole analysis period, the group of absent users is absent for at least 14 days (see time frame B). We then compare how each group has changed, by comparing the behavior in time frame C to the behavior in time frame A. In the following the selection process for the present and absent group is described in more detail.

4.2 Defining groups

We define absence on a subreddit as inactivity in posting and commenting on a subreddit for at least 14 days (i.e., a minimum 14-day difference between the creation days of two consecutive comments by the same author). To build our *absent group* for a subreddit, we collected all cases where a user was absent from the subreddit during the six months of our data. For each user, we analyze the time frame starting with the 7 days before an absence and ending with the 7 days after the absence. This is our investigation period for the absent user. To ensure that all users were active in the 7 days before and after absence (and not absent again), we included only those users, whose first contribution in the investigation period was created less than 14 days after the preceding contribution and whose last contribution in the investigation period was created less than 14 before the following contribution. A user could have multiple entries in the absent group when being absent multiple times. For each absent user, we collected the contributions in the 7 days before absence as well as the contributions in the 7 days after absence.

We then built the set of *present users*. To ensure that the comparison between present and absent users would not be influenced by inherent activity differences in both groups, we decided to choose for each absent user a matching present user with a similar activity level. To furthermore ensure that time and external events have no effect on the comparison, we decided to analyze a present user in the same time frame as his/her matching absent user. The matching process worked as follows: for each absent user, we randomly selected a user who was present during the whole investigation period of the absent user and whose number of contributions in the 7 days before the absence of the treatment user had a minimal difference to the number of contributions of the absent user in this time frame. These users represented the present group. A user could have multiple entries in the present group, when he/she was chosen as partner for multiple treatment cases. For each present user, we collected the contributions in the 7 days before the absence of his/her matching absent user (time frame A in Fig. 3) as well as the contributions in the 7 days after the absence of his/her matching absent user (time frame C in Fig. 3) . Both groups for r/The_Donald consisted of 17,157 cases, for r/politics of 52,071 cases and for r/ChapoTrapHouse of 2,788 cases.

4.3 Measuring Activity

Methods. To analyze whether continuous interactions in a polarizing environment has an affect on posting and commenting activity, we did the following: for each user in each

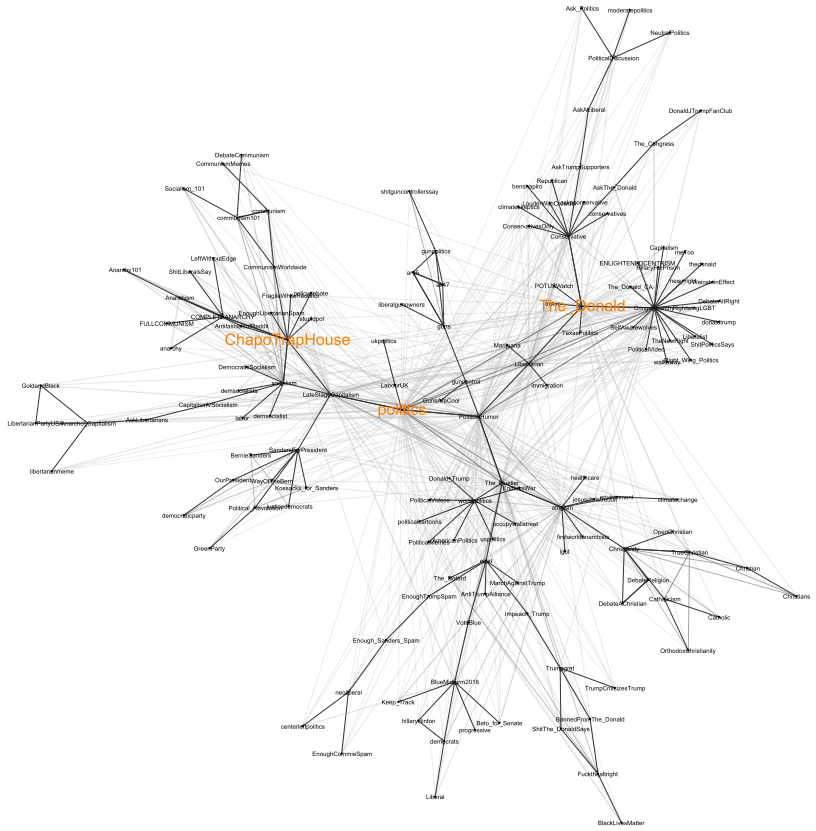


Figure 2: Visone's (Baur et al. 2002) backbone visualization of the co-occurrence network of users being active in the 146 political subreddits.

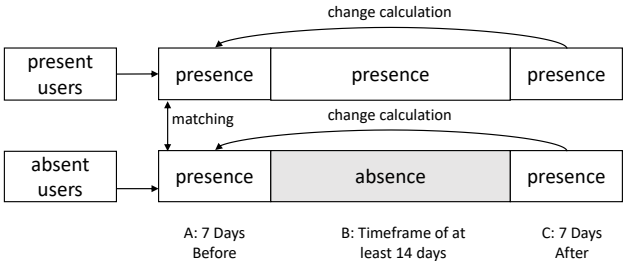


Figure 3: Overview of experimental setup.

group, we calculated the number of contributions he/she has made in the 7 days after (the own or matching user's) absence (time frame C in Fig. 3) and subtracted the number of contributions he/she has made in the 7 days before (the own or matching user's) absence (time frame A in Fig. 3). We then compared both groups with regard to their activity differences. The values of activity differences in both groups were not normally distributed. Therefore, we used the nonparametric Mann-Whitney U test to assess whether the difference values in the present group were statistically different from the difference values in the absent group. In a next step, we conducted the nonparametric Wilcoxon signed-rank test for each group to investigate whether the activity in that group is different in the 7 days after an absence/presence than in the 7 days before. As we conducted the test for each of the three subreddits, we use a Bonferroni adjusted significance level of 0.05/3=0.017.

Results. The first chart of Fig. 4 shows the interquartile range of activity differences for each group, with the middle line representing the median. The statistical values of all tests can be found in Table 2. On all subreddits, continuous interactions has an effect on activity, as the difference values in the absent group are statistically significantly lower than the difference values in the present group. Present users (=users, who were continuously engaged in a subreddit) on all subreddits significantly increased in their level of activity. Absent users in contrast did not significantly change in their level of activity.

Interpretation. This means that users who are constantly engaged in a subreddit not only remain equally active, but become more active.

4.4 Measuring Language

Methods. In order to understand whether continuously active/present users change in their style of language, we now wanted to test for features of linguistic attributes. We test this again for the three online political communities.

To analyze whether continuous interaction in a polariz-

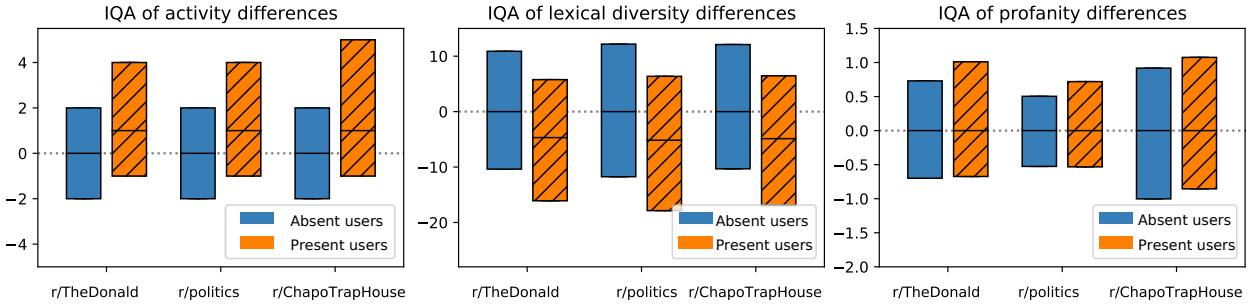


Figure 4: Interquartile range (central lines represent median) of activity, lexical diversity and profanity differences.

Differences in	Subreddit	Absent Users	Present Users	Between Groups
Activity	r\The_Donald	Z= -2.17, p= .029	Z= -30.60, p<.001 ↗	U_{min}= 117047361, p<.001, r= .21
	r\politics	Z= -2.18, p= .029	Z= -46.20, p<.001 ↗	U_{min}= 1105221144, p<.001, r= .19
	r\ChapoTrapHouse	Z= -0.05, p= .962	Z= -13.27, p<.001 ↗	U_{min}= 3101638, p<.001, r= .20
Lexical Diversity	r\The_Donald	Z= -2.14, p= .033	Z= -32.12, p<.001 ↘	U_{min}= 102863734, p<.001, r= .18
	r\politics	Z= -2.62, p= .009 ↗	Z= -58.37, p<.001 ↘	U_{min}= 941222516, p<.001, r= .18
	r\ChapoTrapHouse	Z= -2.21, p= .027	Z= -12.58, p<.001 ↘	U_{min}= 2728456, p<.001, r= .19
Profanity	r\The_Donald	Z= -1.29, p= .197	Z= 16.78, p<.001 ↗	U_{min}= 121043421, p<.001, r= .04
	r\politics	Z= -0.39, p= .694	Z= 32.63, p<.001 ↗	U_{min}= 1105253057, p<.001, r= .03
	r\ChapoTrapHouse	Z= -0.67, p= .504	Z= 5.42, p= .260	U_{min}= 3242894, p= .009, r= .04

Table 2: Tests for trends in differences in activity, lexical diversity or profanity before and after a period of absence/presence in each group as well as between groups. The Z-values represent the z-standardized test statistics of the Wilcoxon signed-rank tests and the U_{min}-values represent the test statistics for the Mann-Whitney U tests. Arrows indicate a statistically significant increase or decrease in activity, lexical diversity or profanity, based on the Bonferroni adjusted significance level of .017 .

ing subreddit has an affect on language use in the comments and posts made on this subreddit, we first needed to extract all users, for whom a text comparison before and after absence/presence is possible. All absent users had at least one contribution before and one after absence. All present users in our data set have made at least one contribution before the absence of their matching absent partner (thus, a user was selected, whose contribution behaviour before absence was the most similar to the absent user's). However, present users could have zero contributions after absence, which would make a text comparison for them impossible. Therefore, we excluded all cases in the present group, where a present user had no contributions in the 7 days after absence, as well as their matching cases in the absent group. Furthermore, we excluded all cases and their matching partners if the users wrote no words in any contributions in the time frame before or after absence. This was for example the case, when users contributed only pictures or messages including only symbols and/or numbers. Both groups for r/The_Donald included 15,850 cases, for r/politics 4,7847 cases and for r/ChapoTrapHouse 2,596 cases. As comments and posts were short in general, we combined the texts of all contributions a user has made in the 7 days before (the own or matching users) absence (time frame A in Fig. 3) to a single text. We did the same for all contributions in the 7 days after (the own or matching user's) absence (time frame C in Fig. 3). For each text we removed special characters, punctuation, empty spaces and numbers. To measure lexical diversity, we calculated for each text the percentage of unique words by all words. For each user we subtracted the diversity percentage of a text before (the own or matching user's) absence from the diversity percentage of a text after (the own or matching user's) absence. To measure profanity, we used a list of profane words, provided by the python package better-profanity[6], and used them as a profanity dictionary. For each text we calculated the percentage of profane words by all words. For each user we subtracted the profanity percentage of a text before (the own or matching user's) absence from the profanity percentage of a text after (the own or matching user's) absence. We compared both groups with regards to their lexical diversity and profanity differences, using Mann-Whitney U tests. Furthermore, we assessed for each group whether their difference values would follow a symmetric distribution around zero, using Wilcoxon signed-rank test. As we conducted the test for each of the three subreddits, we used a Bonferroni adjusted significance level of 0.05/3=0.017.

Results for lexical diversity. In the middle part of Fig. 4 we see the interquartile range of lexical diversity differences for each group, with the middle line representing the median. The statistical values of all tests can be found in Table 2. On all subreddits, continuous interaction has an effect on lexical diversity, as the difference values in the absent group were

[6]https://pypi.org/project/better-profanity/

Differences in	Subreddit	Absent users	Present Users
Activity	r\The_Donald	r= -.02, p= .014	r= .29, p<.001
	r\politics	r= -.01, p= .043	r= .27, p<.001
	r\ChapoTrapHouse	r= .01, p= .784	r= .27, p<.001
Lexical Diversity	r\The_Donald	r= .01, p= .334	r= -.15, p<.001
	r\politics	r= .01, p= .015	r= -.15, p<.001
	r\ChapoTrapHouse	r= .03, p= .207	r= -.12, p<.001
Profanity	r\The_Donald	r= -.02, p= .019	r= .02, p= .013
	r\politics	r= .00, p= .799	r= .02, p<.001
	r\ChapoTrapHouse	r= .01, p= .609	r= -.01, p= .489

Table 3: Correlations between overall activity and the differences in activity, lexical diversity or profanity before and after a period of absence/presence

statistically significantly higher than the difference values in the present group. Present users (=users, who were continuously engaged in a subreddit) on all subreddits decreased in their lexical diversity over time. In contrast, absent users on r/politics significantly increased in their lexical diversity after an absence, while absent users on the other two subreddits did not change significantly.

Results for profanity The lower part of Fig. 4 shows the interquartile range of profanity differences for each group, with the middle line representing the median. The statistical values of all tests can be found in Table 2. On all three subreddits, continuous interaction has an effect on profanity, as the difference values in the absent group were statistically significantly lower than the difference values in the present group. Present users (=users, who were continuously engaged in a subreddit) on r/The_Donald and r/politics significantly increased in their level of profanity. While the interquartile range also indicates an increase of profanity for present users on r/ChapoTrapHouse this was not statistically significant. Absent users in contrast, did not change their level of profanity.

Interpretation. Our results demonstrate that when users are present and remain constantly active, the diversity in their language decreases and their use of profanity increases.

4.5 Measuring Overall Activity

In the following, we will test whether the observed effects are influenced by activity, i.e. are the effects stronger for users, who are in general more active?

Methods. To assess the influence of overall activity on the different patterns that present and absent users had shown, we calculated for each subreddit for each user in our groups the number of contributions he/she has made on this subreddit in the six-months observation period. The number of contributions was not normally distributed. We therefore tested for a monotonic relationship between the number of contributions and the difference values in each group using Spearman's rank correlation coefficient, with a Bonferroni adjusted significance level of 0.05/3=0.017.

Results for activity. In Table 3 we can see the statistical values of the correlations. On all three subreddits there is a weak positive correlation between the activity differences and the number of contributions in the present group: The more a user is generally active over the whole time period the more he/she increases in activity if he/she remains present and continuously active. In contrast, there is no correlation between the activity differences and the number of contributions in the absent group on r/ChapoTrapHouse and on r/politics, and a significant, but negligible negative correlation between the activity differences and the number of contributions in the absent group on r/The_Donald.

To visualize how this would affect the differences between present and absent group, we sorted all users of each group on each subreddit according to the number of contributions and built subgroups of size equal to 10% of overall group size. We then contrasted the interquartile range of present users with the interquartile range of absent users in each subgroup. The resulting image is shown in Fig. 5. The figure shows that in the group of present users on all three subreddits more active users have a higher increase in activity than less active users.

Results for lexical diversity. The statistical values of the correlations can be found in Table 3. On all three subreddits there is a weak negative correlation between the lexical diversity differences and the number of contributions in the present group: The more a user is generally active the more he/she decreases in lexical diversity when being continuously present in a subreddit. In contrast, there is no correlation between the lexical diversity differences and the number of contributions in the absent group on r/ChapoTrapHouse and on r/The_Donald, and a significant, but negligible positive correlation between the activity differences and the number of contributions in the absent group on r/politics. The effect of general activity on differences between present and absent user is visualized in Fig. 6. The figure shows that in the present groups of all three subreddits more active users have a higher decrease in lexical diversity than less active users.

Results for profanity. The statistical values of the correlations can be found in Table 3. There is a significant, but negligible positive correlation between the profanity differences and the number of contributions in the present group on r/politics and r/The Donald, as well as no correlation on r/ChapoTrapHouse. Similarly, there is no correlation between the profanity differences and the number of contributions in the absent group on all three subreddits.

Interpretation. We can state that the effect of continuous presence on activity and lexical diversity is influenced by

overall activity. This means that people who are generally more active show a greater increase in activity and a greater decrease in diversity after an activity period than people who are in general less active.

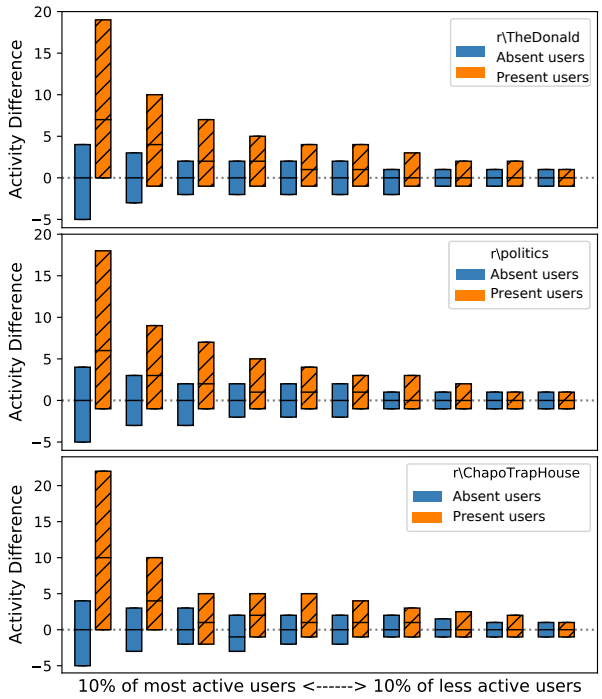


Figure 5: Activity differences of users grouped by overall activity for three subreddits.

5 Discussion and Conclusion.

To investigate the question of how presence in polarizing environments influences people linguistically, we analyzed social media data from three political subreddits on Reddit. We compared the behavior of continuously engaged users to the behavior of similar users who were absent for a certain time to determine 1) whether users change regarding their activity and language and 2) whether this change is influenced by absence.

For all three online political communities we found that given the same initial level of activity, users who are continuously present increase their activity and their language changes compared to users with a non-activity. We were able to determine this equally for three of the biggest online political communities on Reddit.

As a result, we can conclude that:

1. Continuously present people increase their activity, decrease their lexical diversity and increase their level of profanity.
2. The effect of continuous presence on activity and lexical diversity is stronger for people who are in general more active.

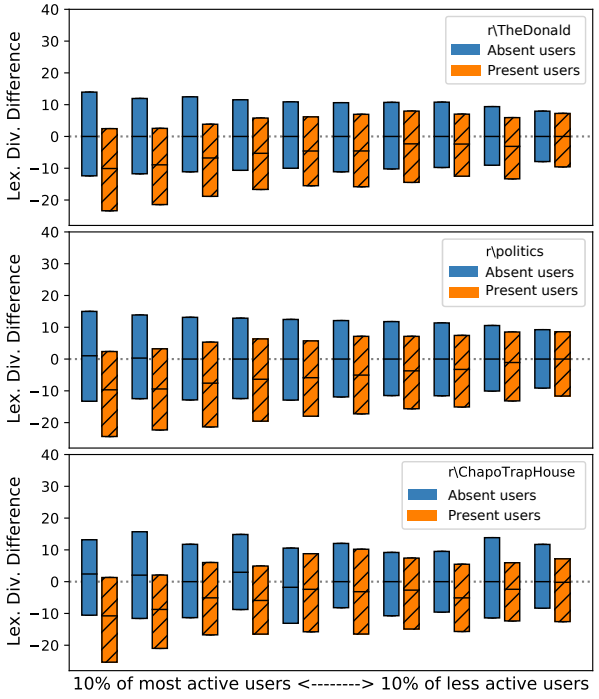


Figure 6: Lexical diversity differences of users grouped by overall activity for three subreddits.

In other words, when dwelling in polarizing environments people engage more in discussions, become simpler in terminology and more angry.

5.1 Limitations.

As we rely on the Pushift Dataset (Baumgartner et al. 2020), we cannot assess whether there were any missing data due to data collection errors. Using social media to study human behaviour is flawed by many challenges (Ruths and Pfeffer 2014). We are aware, however, that there were posts and comments in our dataset which were deleted by users or moderators before being retrieved by Pushift. This data was marked with a ”[deleted]” in the author field and had therefore to be excluded from the analysis. From the data that were available, we leveraged word lists implemented in Python in order to gain insights into the general wording of users. Furthermore, we did not identify any obvious idiosyncrasies with regard to this platform which would prevent our findings from being generalized to other online political forums although a broader comparative study would help to determine this question. Our research represents an attempt at understanding what presence in polarizing environments does on user behavior. Follow up research could leverage contextualized word embeddings (Peters et al. 2018) to identify which mentions of potentially harmful words demand attention. In addition, future work could complement our linguistically-driven approach by comparing lexical choices (Wang and Culotta 2019).

5.2 Implications and Future Work.

In our analysis we compared the behavior of present users to the behavior of similar users with absence to determine whether users' behavior changes in polarizing environments. We found that users participate more, become simpler in vocabulary and use more offensive language. Spending inordinate amounts of time in close-knit environments is seen as a central problem and our results confirm slight changes. Based on this we would like to enrich mass media research on user activity and on short-term intervention techniques. Media consumption leads to changes in behavior and being active online in communities is one form of media consumption. News that is shared and active social interaction in form of discussions influence consumption behavior and this in turn influences the selection of news and interaction. That said, reinforcement effects in behavior and attitudes can occur through selective exposure (Berelson, Gaudet, and Lazarsfeld 1944; Katz 2001).

The question we want to pose here is how the mass media stream of information consumption can be interrupted and whether non-consumption in turn can influence behavior. Hence, we open the discussion for possible (automatic) content moderation techniques to counteract reinforcement effects. In particular, we think of time outs as in sports or the *dead cat* in British parliament discussions[7]. However, little is known about the effectiveness of concrete interventions such as blocking single users temporarily in online political communities. The core motivation and function of short-term interventions could be to slow down activity of single group participants to balance emotions and avoid escalation. Can a time-out have a de-escalating effect? Future work could include experiments which actively induce a user's absence to measure the effect of time-outs caused by others. Furthermore, it would be interesting also to examine the long-term effects of online political conversations. When do these discussions lead to constructive discourses and conversations not resulting in a ban from these platforms and, given that conversations often turn on a particular instance in the discourse, what are the micro-level linguistic traits that enhance long-term stability?

[7] https://en.wikipedia.org/wiki/Dead_cat_strategy

References

Allcott, H.; Braghieri, L.; Eichmeyer, S.; and Gentzkow, M. 2020. The Welfare Effects of Social Media. *American Economic Review* 110(3): 629–676.

An, J.; Kwak, H.; Posegga, O.; and Jungherr, A. 2019. Political discussions in homogeneous and cross-cutting communication spaces. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 68–79.

Babaei, M.; Kulshrestha, J.; Chakraborty, A.; Benevenuto, F.; Gummadi, K. P.; and Weller, A. 2018. Purple Feed: Identifying High Consensus News Posts on Social Media. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 10–16.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, 830–839.

Baur, M.; Benkert, M.; Brandes, U.; Cornelsen, S.; Gaertler, M.; Köpf, B.; Lerner, J.; and Wagner, D. 2002. Visone : Software for Visual Social Network Analysis. In Mutzel, P., ed., *Graph drawing : 9th international symposium*, 463–464. Springer.

Berelson, B.; Gaudet, H.; and Lazarsfeld, P. F. 1944. *The People's Choice. How the voter makes up his mind in a presidential campaign.* Duell, Sloan & Pearce.

Bowers, J. S.; and Pleydell-Pearce, C. W. 2011. Swearing, Euphemisms, and Linguistic Relativity. *PLoS ONE* 6(7): e22341.

Brown, A. J. 2020. "Should I Stay or Should I Leave?": Exploring (Dis)continued Facebook Use After the Cambridge Analytica Scandal. *Social Media + Society* 6(1).

Davis, J. L.; and Graham, T. 2021. Emotional consequences and attention rewards: the social effects of ratings on Reddit. *Information, Communication & Society* 24(5): 649–666.

De Francisci Morales, G.; Monti, C.; and Starnini, M. 2021. No echo in the chambers of political interactions on Reddit. *Scientific Reports* 11(1): 2818.

Ferrer, X.; Nuenen, T. v.; Such, J. M.; and Criado, N. 2021. Discovering and Categorising Language Biases in Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 140–151.

Gallacher, J. D.; Heerdink, M. W.; and Hewstone, M. 2021. Online Engagement Between Opposing Political Protest Groups via Social Media is Linked to Physical Violence of Offline Encounters. *Social Media + Society* 7(1).

Garimella, K.; Smith, T.; Weiss, R.; and West, R. 2021. Political Polarization in Online News Consumption. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 152–162.

Grover, T.; and Mark, G. 2019. Detecting Potential Warning Behaviors of Ideological Radicalization in an Alt-Right Subreddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 193–204.

Guntuku, S. C.; Buffone, A.; Jaidka, K.; Eichstaedt, J. C.; and Ungar, L. H. 2019. Understanding and Measuring Psychological Stress Using Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 214–225.

Jarvis, S. 2013. Capturing the Diversity in Lexical Diversity. *Language Learning* 63: 87–106.

Jay, T.; and Janschewitz, K. 2008. The pragmatics of swearing. *Journal of Politeness Research* 4,2.

Jhaver, S.; Birman, I.; Gilbert, E.; and Bruckman, A. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction* 26(5): 31:1–31:35.

Kane, B.; and Luo, J. 2019. Do the Communities We Choose Shape our Political Beliefs? A Study of the Politicization

of Topics in Online Social Groups. In *IEEE International Conference on Big Data*, 3665–3671.

Katz, E. 2001. Lazarsfeld's Map of Media Effects. *International Journal of Public Opinion Research* 13(3): 270–279.

Kovacs, G.; Wu, Z.; and Bernstein, M. S. 2021. Not Now, Ask Later: Users Weaken Their Behavior Change Regimen Over Time, But Expect To Re-Strengthen It Imminently. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14.

Kubin, E.; and von Sikorski, C. 2021. The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association* 45(3): 188–206.

Mamié, R.; Horta Ribeiro, M.; and West, R. 2021. Are Anti-Feminist Communities Gateways to the Far Right? Evidence from Reddit and YouTube. In *13th ACM Web Science Conference 2021*, 139–147. Association for Computing Machinery.

Pariser, E. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin Press.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, 2227–2237.

Phadke, S.; Samory, M.; and Mitra, T. 2021. What Makes People Join Conspiracy Communities? Role of Social Factors in Conspiracy Engagement. *Proc. ACM Hum.-Comput. Interact.* 4(CSCW3).

Rajadesingan, A.; Budak, C.; and Resnick, P. 2021. Political Discussion is Abundant in Non-political Subreddits (and Less Toxic). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 525–536.

Ruths, D.; and Pfeffer, J. 2014. Social Media for Large Studies of Behavior. *Science* 346(6213): 1063–1064.

Schmid, H.-J.; Würschinger, Q.; Fischer, S.; and Küchenhoff, H. 2021. That's Cool. Computational Sociolinguistic Methods for Investigating Individual Lexico-grammatical Variation. *Frontiers in Artificial Intelligence* 3.

Silberman, W. R.; and Record, R. A. 2021. We Post It, U Reddit: Exploring the Potential of Reddit for Health Interventions Targeting College Populations. *Journal of Health Communication* 26(6): 381–390.

Strathern, W.; Schoenfeld, M.; Ghawi, R.; and Pfeffer, J. 2020. Against the Others! Detecting Moral Outrage in Social Media Networks. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 322–326.

Sunstein, C. R. 2009. *Going to extremes: How like minds unite and divide*. Oxford University Press.

Wadden, D.; August, T.; Li, Q.; and Althoff, T. 2021. The Effect of Moderation on Online Mental Health Conversations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 751–763.

Waller, I.; and Anderson, A. 2021. Quantifying social organization and political polarization in online platforms. *Nature* 600(7888): 264–268.

Walther, J. B. 2022. Social media and online hate. *Current Opinion in Psychology* 45.

Wang, Z.; and Culotta, A. 2019. When Do Words Matter? Understanding the Impact of Lexical Choice on Audience Perception Using Individual Treatment Effect Estimation. *AAAI* 33(1): 7233–7240.

Weld, G.; Glenski, M.; and Althoff, T. 2021. Political Bias and Factualness in News Sharing across more than 100,000 Online Communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 796–807.

Zarouali, B.; Makhortykh, M.; Bastian, M.; and Araujo, T. 2021. Overcoming polarization with chatbot news? Investigating the impact of news content containing opposing views on agreement and credibility. *European Journal of Communication* 36(1): 53–68.

Zhou, X.; Rau, P.-L. P.; Yang, C.-L.; and Zhou, X. 2021. Cognitive Behavioral Therapy-Based Short-Term Abstinence Intervention for Problematic Social Media Use: Improved Well-Being and Underlying Mechanisms. *Psychiatric Quarterly* 92(2): 761–779.

5.4.2 Are your Friends also Haters? Identification of Hater Networks on Social Media

Authors Maximilian Wich, Melissa Breitinger, Wienke Strathern, Marlena Naimarevic, Georg Groh, Jürgen Pfeffer

In WWW '21 Companion: Companion Proceedings of the Web Conference 2021, Ljubljana, Slovenia, Association for Computing Machinery, pp. 481-485, ISBN: 9781450383134, https://doi.org/10.1145/3442442.3452310.

Abstract

Hate speech on social media platforms has become a severe issue in recent years. To cope with it, researchers have developed machine learning-based classification models. Due to the complexity of the problem, the models are far from perfect. A promising approach to improving them is to integrate social network data as additional features in the classification. Unfortunately, there is a lack of datasets containing text and social network data to investigate this phenomenon. Therefore, we develop an approach to identify and collect hater networks on Twitter that uses a pre-trained classification model to focus on hateful content. The contributions of this article are (1) an approach to identify hater networks and (2) an anonymized German offensive language dataset that comprises social network data. The dataset consists of 4,647,200 labeled tweets and a social graph with 49,353 users and 122,053 edges.

[5]https://creativecommons.org/licenses/by/4.0/

Publication Summary

Questions: In Wich et al. (2021) we present a method to classify hate speech. Based on theories and literature the outcome was a classification schema with which we can capture different forms of hate. The issue of hate speech on social media has become a significant concern, and researchers have developed machine learning models to classify such content. However, the complexity of the problem has made it challenging to create effective models. Therefore, researchers are exploring incorporating social network data as additional features to improve the models. Unfortunately, the lack of datasets containing both text and social network data has hindered this research. To address this gap, we present a new method for detecting and collecting networks of individuals who engage in hate speech on Twitter. **Data:** Our method uses a pre-trained classification model to identify offensive language and collect a new anonymized dataset of 4,647,200 labeled tweets. This dataset also includes a social graph of 49,353 users and 122,053 edges. **Methods and Analysis:** Our research has two primary goals: (1) to develop a methodology for identifying hate speech networks and (2) to create a valuable dataset of offensive language in German that includes social network data. We first train an offensive language classification model using a publicly available dataset to achieve these goals. Then, we select initial seed users who have exhibited hateful behavior and collect data on their social networks, using our classifier to identify and collect data based on the level of offensiveness. **Results and Interpretation:** Finally, we manually annotate a sample of the collected data to evaluate the effectiveness of our approach. Our method involves leveraging four types of social relationships: Friends network, Mutual network, Retweet network (in-degree), and Retweet network (out-degree). The Friends network includes all the users followed by the seed user, while the Mutual network represents the intersection of the seed user's friends and followers. The Retweet network (in-degree) consists of all users who have retweeted the seed user, and the Retweet network (out-degree) includes all users retweeted by the seed user.

Author Contribution

Wienke Strathern compiled the German right-wing dataset and developed a classification schema.

Are Your Friends Also Haters? Identification of Hater Networks on Social Media

Data Paper

Maximilian Wich
Department of Informatics, Technical University of Munich
Germany
maximilian.wich@tum.de

Melissa Breitinger
Department of Informatics, Technical University of Munich
Germany
melissa.breitinger@tum.de

Wienke Strathern
Bavarian School of Public Policy
Technical University of Munich
Germany
wienke.strathern@tum.de

Marlena Naimarevic
Bavarian School of Public Policy
Technical University of Munich
Germany
marlena.n@arcor.de

Georg Groh
Department of Informatics, Technical University of Munich
Germany
grohg@in.tum.de

Jürgen Pfeffer
Bavarian School of Public Policy
Technical University of Munich
Germany
juergen.pfeffer@hfp.tum.de

ABSTRACT

Hate speech on social media platforms has become a severe issue in recent years. To cope with it, researchers have developed machine learning-based classification models. Due to the complexity of the problem, the models are far from perfect. A promising approach to improve them is to integrate social network data as additional features in the classification. Unfortunately, there is a lack of datasets containing text and social network data to investigate this phenomenon. Therefore, we develop an approach to identify and collect hater networks on Twitter that uses a pre-trained classification model to focus on hateful content. The contributions of this article are (1) an approach to identify hater networks and (2) an anonymized German offensive language dataset that comprises social network data. The dataset consists of 4,647,200 labeled tweets and a social graph with 49,353 users and 122,053 edges.

CCS CONCEPTS

• **Computing methodologies** → **Language resources**; • **Human-centered computing** → **Collaborative and social computing theory, concepts and paradigms**.

KEYWORDS

hate speech, abusive language, dataset, network analysis, machine learning, classification

ACM Reference Format:
Maximilian Wich, Melissa Breitinger, Wienke Strathern, Marlena Naimarevic, Georg Groh, and Jürgen Pfeffer. 2021. Are Your Friends Also Haters? Identification of Hater Networks on Social Media: Data Paper. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3442442.3452310

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3452310

1 INTRODUCTION

The rise of social media platforms (e.g., Facebook and Twitter) does not only have positive effects on society. A phenomenon showing the dark side of social media is the spread of hate speech [3]. Hate speech is a severe issue because it is not limited to the online world but it can also *spill over* into the offline world, e.g. by causing physical crime [17]. Consequently, the identification of hate speech is an important societal challenge.

Since the users on social media produce enormous amounts of data, it is impossible to manually monitor their content. That is why machine learning models have been developed to automatically detect hate speech. Even if the results look promising, the models have limited accuracy [2, 12].

One challenge is that hate speech is a broad and complex phenomenon and comprises various sub-types (e.g., anti-Semitism, misogyny, racism), making automatic detection difficult. One idea is to integrate additional data into the classification model besides the textual data [10, 11]. The hypothesis behind this is that characteristics about the user and its social network provide additional clues helping to detect hate speech. It is grounded on the fact that according to [7] a high portion of hateful and offensive content is produced by small subnetworks. The problem that the research community is facing here is a lack of datasets to investigate this hypothesis. There are already a lot of abusive language datasets available.

Therefore, we have two research objectives: (1) we aim to develop an approach to identify and collect hater networks on a social media platform (in our case Twitter) and (2) we aim to release the collected data (social media posts and social network data of the authors).

For this purpose, we train an offensive language detection model on a publicly available dataset. In the second step, we select a set of hateful seed users that serves as a starting point. Then, we collect their social networks depending on the offensiveness of the content by pseudo labeling the collected data with our classifier. In the fourth step, we annotate a sample of the gathered data to evaluate our approach.

Contributions:

- Approach: We provide a methodology to identify and gather hater networks on Twitter.
- Dataset: We release an offensive language dataset in German that contains 4,647,200 labeled tweets, 49,353 users and 122,053 edges of the social graph. The 4,647,200 labels are pseudo labels produced by a classification model. Furthermore, human annotators annotated 1,356 tweets for evaluation purposes (included in the dataset). To protect users' privacy, we anonymized the data and replaced all usernames with anonymous identifiers.

2 RELATED WORK

Researchers in the hate speech detection community have investigated the relevance of social network data [6] for hate speech classification. Chatzakou et al. [1] integrated user-based and network-based features into their classification model in addition to the textual data. They showed that the additional features improve the classification performance. But the network-based features were limited to aggregated metrics for each user (e.g., number of followers and friends), meaning that the dataset did not contain any information about relations. Other researchers [4, 5] picked up Chatzakou et al.'s [1] approach to integrating aggregated network metrics and confirmed their findings. In contrast to them, [10] used the actual edges of the follower network in form of a node2vec graph embedding to improve the hate speech classification. For this purpose, they used the dataset from Wassem and Hovy [14] and enriched it with social network data. The problem with this approach is that most of the hateful tweets in the dataset were produced by only a few users [15], meaning the network data is not representative. Ribeiro et al. [11] applied a network-centric approach to collect data and to investigate the relevance of network data for hate speech detection. They crawled a sample of Twitter's retweet network and tweets of the discovered users, starting from a seed user. Then, they annotated a sample of the data, trained a classifier using textual and network data, and evaluated the model. Unfortunately, they released only the social graph and the tweets as averaged word embeddings, making it very hard to use this dataset in other models. Their approach, however, is similar to our one - except that we consider more network types and integrate a classification model in our process to crawl the networks more targeted.

3 METHODOLOGY

Our approach consists of 4 phases, as depicted in Figure 1. In the first phase, we train an offensivness classification model. In the second phase, we select the seed users whose social networks are gathered based on the content's offensiveness. Thirdly, we crawl the social networks using an offensiveness classification model to filter offensive users (haters). In the fourth phase, we manually annotate a sample of the collected tweets to evaluate our approach.

3.1 Training Classification Model

We need a classification model to detect offensive language in the tweets for identifying hater networks. As the basis, we use a pre-trained German BERT model [9]. In the first step, we fine-tune the language model of the pre-trained BERT with around 4 million German tweets, which we preprocess beforehand. In the second step, we add a classification head to the model and train it to distinguish between offensive and non-offensive languages. For the training, the datasets of GermEval Shared Task on the Identification of Offensive Language 2018 [16] and GermEval Task 2, 2019 shared task on the identification of offensive language [13] are used. Since both datasets have the same labeling schema, they can be merged to one dataset. The term offense in the context of these datasets covers a wide range of aspects so that a classifier trained on this data is suitable to identify haters. It comprises "abusive language, insults, as well as merely profane statements" [16, p.2].

3.2 Selecting Seed Users

In the second phase, we select the seed users that serve as a starting point for the network crawling phase. In total, we select 9 seed users from different sources: (1) GermEval 2019 dataset, (2) German right-wing dataset, and (3) manual exploration of Twitter. By doing so, we ensure to have already classified haters and avoid an author bias. Due to the limitations of the Twitter API, we cannot start with a large number of seeds. Otherwise, crawling would take too long.

GermEval 2019 Shared Task 2. The first one is the dataset of GermEval 2019 Shared Task 2 containing 8,952 tweets labeled as offensive or non-offensive that is also used for training the classifier. We select the top 500 users that the largest amount of offensive tweets stems from. After that, we collect from these 500 users their most recent timelines via the Twitter API, limiting the number of tweets to 50. Then, all collected tweets are classified to assign each user an offensiveness score o_u that is calculated as follows:

$$o_u = \frac{1}{1 - \log\left[\frac{\sum_{i=1}^{n} p_{i,c_i=1}}{\sum_{i=1}^{n} p_{i,c_i=1} + \sum_{i=1}^{n} p_{i,c_i=2}}\right]} \in [0,1] \tag{1}$$

$p_{i,c_i=c}$: probability of tweet i for class c

Subsequently, the users with $o_u \geq 0.5$, i.e. offensive users or haters, are manually reviewed with respect to user activeness. Finally, 4 users from list of the most offensive and active users are selected as seed profiles.

German right-wing dataset. The second source is a dataset that we have collected, containing German-speaking tweets from right-wing Twitter users. Since the data is not labeled, we classify all tweets with our classifier and apply the same procedure as the one for the GermEval dataset — computing the offensiveness scores, ranking the user accordingly, investigating the user activity of the top ranked. Finally, we select the top 5 of most offensive and active users, while two of them appear already in the seed list from GermEval. The dataset itself was collected as follows: In a first step, we searched Twitter for users whose profile information included two German right wing parties. In a second step, we read about a 100 tweets to study the topics being discussed on Twitter by these two parties. Reading the tweets we filtered seven main categories to which the content could be referred to: ethnicity, nationality, sexuality, gender, religion, disability, class. Next, we manually collected Twitter account names of people who frequently took action in these discussions, i.e., actively posted and interacted. For each party we collected 500 followers. In addition to the names, we filtered

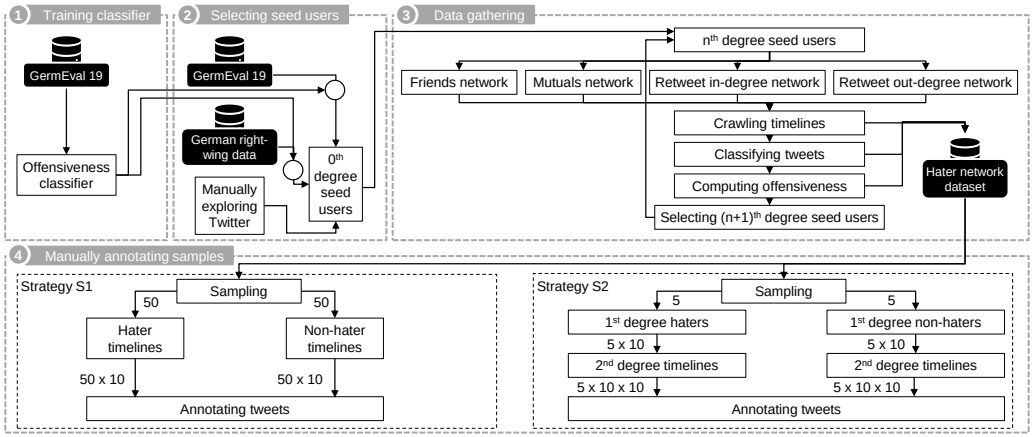


Figure 1: Methodology

also the associated profile information. Based on these information, we followed these users on Twitter and collected their posted tweets from February 22 to April 6, 2020, 45 days in total. To further understand content and language of these users, we evaluated qualitatively the top 1,000 tweets that had been re-tweeted most often. By doing so we observed that 90-95% of these tweets could be classified according to the above mentioned categories. Secondly, we took a closer look at the language being used and observed that 90% of these tweets contained offensive words. More precisely, the tweets contained clear offensive words and they were used in the context of directed aggressiveness against a group of one of the above categories. These categories are predominantly topics of hate speech. Hate speech, mostly likely, is used against a certain group or community. Regarding the time period, it should be noted that the first Corona case in Germany became known on January 28, 2020. It can be said in retrospect that the next four weeks were the media starting point of Covid19 reporting. The continuous increase of infected persons started four weeks later, February 25, 2020 — right after the start of our data collection period. From the 1,000 manually collected accounts, there was some overlap between the two right-wing party supporters. 886 accounts remained. Of these, some were no longer active, and we were ultimately able to filter out 858 users. We followed them and in total, we were able to collect about 9,000-10,000 tweets per day. The majority of the tweets (90-95%) were retweets. The data was collected on the basis of UTC-0 timezone.

Manual exploration. During our explorative research on Twitter, we identified 2 more hateful profiles that we add to our seed list.

3.3 Data Gathering

After selecting the 0^{th} seed users, we iteratively collect their social network, as depicted in phase 3 in Figure 1. For this purpose, we use 4 different types of social relations:

- Friends network: users followed by seed user
- Mutual network: intersection of friends and followers of a seed user
- Retweet network (in-degree): retweeters of a seed user
- Retweet network (out-degree): users retweeted by a seed user

We do not consider all types of social relations that are provided by Twitter. We exclude the follower network of seed users because the follower network, in general, could be extensive. The reason is that everyone can nearly follow everyone without permission, making this kind of relationship also less meaningful. The mention network meaning one user mentions another user in a tweet is also not considered since the in-degree mention network (users mentioned a seed user) is not accessible via the standard API.

To collect the retweet network, we extract the 500 most recent tweets of a user and analyze whom they have retweeted and who has retweeted the tweets of the user. The result is a list of usernames that have a relationship to the seed users. In the next step, we gather the 100 most recent tweets from their timeline to classify them with the hate classifier and calculate the users' offensiveness score.

Since we want to collect data from hater networks and avoid that the amount of data to collect grows exponentially, we cannot crawl the social networks of all collected users. Therefore, we have to limit this number. We do this by selecting all intersecting haters — an intersecting hater has relations to at least two seed users — and 50 other users with the highest offensiveness score o_u. Haters with a score of 1.0 are excluded because manual exploration has shown that these are either bots or users with only a few tweets. Regarding the non-hater seeds, we define a range for o_u between 0.25 and 0.5 for intersecting non-haters, aiming to choose seeds that are close to haters. A further restriction is a limit of a maximum of 1,000 followers. It aims to exclude popular profiles that interact with many non-hateful users.

These identified haters serve as seed users for the next cycle. In this paper's scope, we apply this cycle two times, meaning that we collect the 1^{st} and 2^{nd} degree hater network.

3.4 Manual Annotation

Since a pre-trained offensiveness classification model classifies the collected tweets, we want to evaluate the classification performance

by manually annotating a sample of the data. To increase the portion of offensive and hateful content in our sample, we apply two different sampling strategies:

- S1: We randomly sample 10 tweets from 50 haters and 50 non-haters — in total 1,000 tweets.
- S2: Firstly, we randomly select 5 1^{st} degree haters and 5 1^{st} degree non-haters. Secondly, we sample 10 tweets from 50 users belonging to the social networks of the 1^{st} degree haters. We also apply this for the non-haters. In total, S2 comprises 1,000 tweets.

Besides increasing the portion of offensive content, sampling the data equally from haters and non-haters helps us to test whether the haters' network contains more offensive content than the others. Applying two different sampling strategies aims to get a diverse sample from the dataset.

The sampled data are annotated by three annotators with expert knowledge in hate speech. Most of the data is annotated by two persons. The third person annotates only these tweets that received diverging annotations from the other two annotators. Since the annotators are allowed to skip a tweet and tweets containing only link(s) are ignored, some sampled tweets are not annotated and others have only one annotation instead of two or three. The inter-rater reliability of the annotators is measured with Krippendorff's Alpha [8].

4 RESULTS

4.1 Classification Model

Our fine-tuned BERT model for identifying offensive language in German tweets reached a macro F1 score of 78.6%. It is 1.5 pp better than the best model submitted to GermEval 2019 [13]. The other evaluation metrics can be found in Table 1.

Table 1: Evaluation metrics of trained BERT classifier

Acc.	Prec.	Recall	Micro F1	Macro F1	Weighted F1
0.821	0.753	0.654	0.82	0.786	0.817

4.2 Collected Data

Starting from the 9 seed users, we partially captured the 1^{st} and 2^{nd} degree network of these users between May 15, 2020 and August 15, 2020. Due to the size of the network and our goal to identify hater networks, we focused on offensive content and offensive users. In total, we collected 49,353 users, the mentioned social relations of these users (friends network, the intersection of follower and friends, retweet in- and out-degree network), and 4,647,200 tweets. 396 (0.8%) of the users were classified as haters ($o_u \geq 0.5$) and 289,780 of the tweets (6.2%) as offensive. Further details can be found in Table 3.

Table 4 shows how many users were gathered depending on the network type and the subnetwork and how large the hater percentage was. In this context, subnetwork means a part of the collected social network. For example, "Degree 1 (H)" comprises all users that have any kind of relations to the 0^{th} degree seed haters. Degree 2 (H and NH) refers to the subnetwork that was collected based on the hate and non-hate seed users of degree 1. Note: Since 0^{th} degree contains only haters, there is no Degree 1 (H).

The first finding is that the subnetworks that have only haters as seed — Degree 1 (H) and Degree 2 (H) — have for all types of networks a higher percentage of haters than the others. The second finding is that the percentage of haters also depends on the type of network. While the retweet in-degree has on average the lowest percentage, the retweet out-degree network seems to be the best network for identifying connected haters.

4.3 Evaluation of Classifications

To evaluate the quality of the pseudo labels that are assigned to the gathered tweets by the classifier, three annotators annotated 1,356 tweets containing 270 offensive ones. The inter-rater reliability in form of the Krippendorff's alpha is 48.9%. It is not the best one, but it is comparable to other hate speech datasets (e.g., [18] with $\alpha = 0.45$). The data to be annotated was sampled by two strategies - 1,000 tweets from S1 and 1,000 tweets S2. Since annotators could skip tweets (e.g., tweets containing only URLs, missing context), S1 produced 857 annotated tweets, S2 499.

To measure the classification performance, we calculated the classification metrics between the pseudo labels provided by the classifier and our annotations. The results can be found in Table 2. The macro F1 score of the classifier on all annotated tweets (S1 and S2) is 75.3%, which is only 3.9 pp lower than on the original test set. The macro F1 score on the S2 data is only 65.9%. This could be related to the fact that dataset is smaller and more imbalanced than the S1 dataset. All in all, the classification performance on the GermEval 2019 test set (Table 1) and on the annotated test set of S1 and S2 (Table 2) are comparable.

Table 2: Classification performance on the manually annotated test data (total and split into strategy S1 and S2)

	S1 and S2	S1	S2
Accuracy	0.828	0.812	0.856
Precision	0.555	0.620	0.324
Recall	0.693	0.728	0.522
Micro F1	0.828	0.812	0.856
Macro F1	0.753	0.769	0.659
Weighted F1	0.835	0.817	0.870
Test data	1356	857	499
– Offensive	270	224	46
– Non-offensive	1086	633	453

5 DISCUSSION

We presented an approach of identifying and collecting hate networks on Twitter and showcased the utility of our approach. We found that the out-degree retweet network is the best of our four selected social relations to uncover hater networks, which partially confirms the finding from [11]. Unfortunately, we could not consider all kinds of social relations offered by Twitter due to missing endpoints. A type that is also interesting is the mention network

Table 3: Overview of gathered data by network degree

	Degree 0	Degree 1	Degree 2	Total	Hater/offensive
Number of users	9	14,084	35,260	49,353	396 (0.8%)
Number of tweets	700	1,367,441	3,279,059	4,647,200	289,780 (6.2%)

Table 4: Number and percentage of classified haters by network type, network degree, and split between hater (N) and non-hater (NH) seeds

	Degree 1 (H)		Degree 2 (H)		Degree 2 (NH)		Degree 2 (H and NH)	
	Total	Hater per.	Total	Hater per.	Total	Hater per.	Total	Hater per.
Friends	3,250	1.45%	12,423	1.57%	28,547	0.37%	36,933	0.71%
Mutuals	1,796	1.61%	6,410	1.61%	7,003	0.73%	11,581	1.11%
Retweet In-Degree	10,332	0.57%	4,062	1.08%	896	0.11%	4,590	0.98%
Retweet Out-Degree	2,419	2.77%	1,070	3.83%	4,757	0.21%	5,488	0.89%

because it reflects which users interact. In general, our approach should be applicable to other social networks that allow extracting social relations.

A point of criticism can be that our dataset mainly contains pseudo labels provided by a classification model. Firstly, it was not possible to manually annotate all 4.6 million tweets due to limited resources. Secondly, our manually annotated test data showed that the classifier provides valid and reliable classification performance to some extent because the metrics on the annotated sample are comparable to the ones on the test set. Thirdly, the focus of this paper was to provide a hate speech dataset with social network data so that other researchers can integrate this additional data into hate speech detection.

A possible improvement of our approach for future work is to work with several classifiers trained on different datasets to cover more aspects of hate speech (e.g., personal attack, sexism misogyny, anti-Semitism). Besides that, increasing the number of annotators and annotated data would also improve our findings' reliability.

6 CONCLUSION

We developed an approach to identifying and collecting hater networks on Twitter that applies a pre-trained classification model to focus on offensive users. We showed that our method produces the desired results. Furthermore, we collected a dataset comprising around 4,647,200 million tweets from 49,353 users (including social relations) that the research community can use to investigate social network data's relevance in hate speech detection. All tweets were pseudo-labeled, and a small sample was manually annotated. An additional finding was that the retweet out-degree network is the most appropriate network type of the investigated networks to detect hater networks.

RESOURCES

The code of our approach is available under https://github.com/mawic/hater-network-identification. If you are interested in the dataset, please contact us via e-mail or https://www.in.tum.de/social/team/maximilian-wich/.

REFERENCES

[1] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proc. 2017 ACM on Web Science Conference*. 13–22.

[2] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proc. 11th ICWSM Conf.*

[3] Maeve Duggan. 2017. *Online harassment 2017.* Pew Research Center.

[4] Elise Fehn Unsvåg and Björn Gambäck. 2018. The Effects of User Features on Twitter Hate Speech Detection. In *Proc. 2nd Workshop on Abusive Language Online*. 75–85.

[5] Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proc. 10th ACM Conference on Web Science*. 105–114.

[6] Marina Hennig, Ulrik Brandes, Jürgen Pfeffer, and Ines Mergel. 2012. *Studying Social Networks. A Guide to Empirical Research.* Campus Verlag.

[7] Philip Kreißel, Julia Ebner, Alexander Urban, and Jakob Guhl. 2018. Hass auf Knopfdruck. Rechtsextreme Trollfabriken und das Ökosystem koordinierter Hasskampagnen im Netz. *Institute for Strategic Dialogue* (2018).

[8] K. Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology.* Sage.

[9] MDZ Digital Library. 2020. dbmdz BERT models. https://github.com/dbmdz/berts (accessed on 22.4.2020).

[10] Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author Profiling for Abuse Detection. In *Proc. 27th International Conference on Computational Linguistics*. 1088–1098.

[11] Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgílio A. F. Almeida, and Wagner Meira. 2018. Characterizing and Detecting Hateful Users on Twitter. *arXiv preprint arXiv:1803.08977* (2018).

[12] Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proc. 5th Intl. Workshop on Natural Language Processing for Social Media*. 1–10.

[13] Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval Task 2, 2019 shared task on the identification of offensive language. In *Proc. 15th KONVENS*. 354–365.

[14] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proc. NAACL student research workshop*. 88–93.

[15] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 602–608.

[16] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proc. 14th KONVENS*.

[17] Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2020. Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. *The British Journal of Criminology* 60, 1 (2020), 93–117.

[18] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proc. 26th International Conference on World Wide Web*. 1391–1399.

5.4.3 Identifying Different Layers of Online Misogyny

This peer-reviewed paper is relevant for examination.

Authors Wienke Strathern, Jürgen Pfeffer

In Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media, 5-8 June, Cyprus, Greece.

Abstract

Social media has become an everyday means of interaction and information sharing on the Internet. However, posts on social networks are often aggressive and toxic, especially when the topic is controversial or politically charged. Radicalization, extreme speech, and in particular online misogyny against women in the public eye have become alarmingly negative features of online discussions. The present study proposes a methodological approach to contribute to ongoing discussions about how women, their experiences, and their choices are attacked in polarized social media responses. Based on a review of theories on and detection methods for misogyny, we present a classification scheme that incorporates eleven different explicit and implicit layers of online misogyny. We also apply our classes to a case study related to online aggression against Amber Heard in the context of her allegations of domestic violence against Johnny Depp. We finally evaluate the reliability of Google's Perspective API – a standard for detecting toxic language – for determining gender discrimination as toxicity. We show that a large part of online misogyny, especially when verbalized without expletive terms but instead more implicitly, is not captured automatically.

Publication Summary

Questions: In Strathern and Pfeffer (2023), we investigate different forms of misogyny on social media and compare automated toxicity metrics for group-based hate. Our focus is on identifying distinct linguistic forms of hate, which can be used to go beyond defined categories and explore more abstract levels of misogyny. **Data:** To accomplish this, we collected 240,000 tweets from the Twitter handle @realamberheard in 2019, 2020, and 2021. We then extracted a subset of 1,000 top retweets for the initial analysis and an additional 5,000 top retweets for the annotation process. **Methods and Analysis:** To develop a classification scheme for online misogyny, we reviewed related literature and identified explicit and implicit misogynistic language. Our classification includes various aspects of hate against women, which sometimes overlap and can be difficult to distinguish. Our goal was not to provide unambiguous definitions, but to cover different forms of misogyny and explore their linguistic characteristics. We applied schema to a case of online misogyny and tested it against a the toxicity metric provided by Google (Google Perspective API). **Results and Interpretation;** Results in

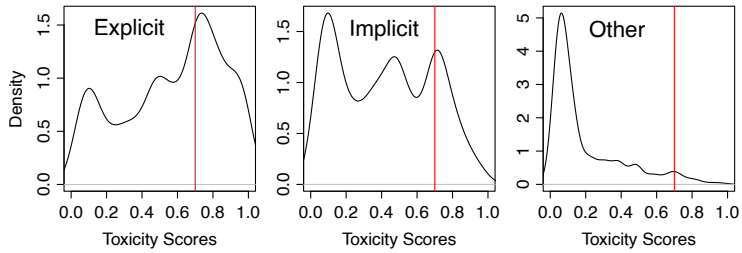


Figure 11: Distribution of toxicity scores from Google's Perspective API for tweets with explicit or implicit misogyny according to Strathern and Pfeffer (2023)

Figure 11 show that the metric performed well in detecting our explicit misogyny forms but poorly in detecting our implicit forms. We mapped the toxicity scores with our labeling to identify structural patterns from the co-occurrence network. We build a co-occurrence network where the nodes are the 12 categories, and the edge value is the number of co-occurrences (=common occurrence of codes within a tweet). Insights from the map in Figure 12 are pretty significant as the network visualizes the proximity of specific categories. It offers a more qualitative comparison of stereotypical hating: statements that demonstrate power are associated with inferiority and insults. A skeptical attitude is associated with abusive terms of inferiority, imputation, gendered personal attacks, and insults.

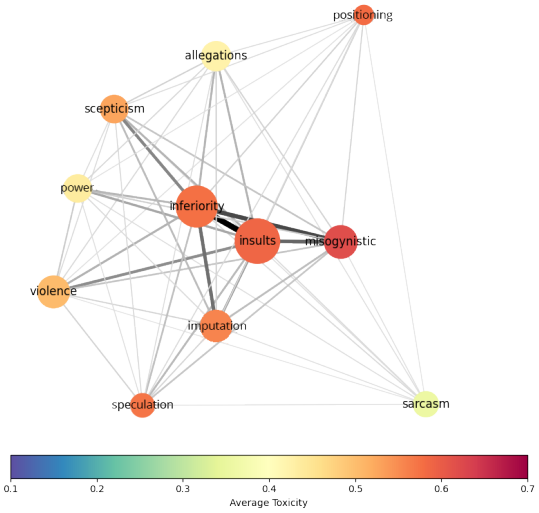


Figure 12: Co-occurrences of categories as in Strathern and Pfeffer (2023)

Statements of speculation and doubt are associated with sarcastic and gender-attacking language.

Author Contribution

Wienke Strathern headed the project, developed the conceptual framework and study design, conducted the literature review, worked out the theory and the study design, built and applied the classification schema, annotated the dataset, did the statistical analysis, and wrote the overall manuscript, revisions, editing.

Identifying Different Layers of Online Misogyny

Wienke Strathern,[1] **Jürgen Pfeffer**[1]

[1]School of Social Sciences and Technology, Technical University of Munich
wienke.strathern@tum.de, juergen.pfeffer@tum.de

Abstract

Social media has become an everyday means of interaction and information sharing on the Internet. However, posts on social networks are often aggressive and toxic, especially when the topic is controversial or politically charged. Radicalization, extreme speech, and in particular online misogyny against women in the public eye have become alarmingly negative features of online discussions. The present study proposes a methodological approach to contribute to ongoing discussions about the multiple ways in which women, their experiences, and their choices are attacked in polarized social media responses. Based on a review of theories on and detection methods for misogyny, we present a classification scheme that incorporates eleven different explicit as well as implicit layers of online misogyny. We also apply our classes to a case study related to online aggression against Amber Heard in the context of the allegations of domestic violence she made against Johnny Depp. We finally evaluate the reliability of Google's Perspective API—a standard for detecting toxic language—for determining gender discrimination as toxicity. We show that a large part of online misogyny, especially when verbalized without expletive terms but instead more implicitly is not captured automatically.

1 Introduction

In May 2016 actress, model, and activist Amber Heard went public and accused her then-husband, actor Johnny Depp, of intimate partner violence. She described a turbulent relationship and reported that "Johnny verbally and physically abused me throughout our relationship"[1]. She publicly posted a picture of injuries and filed for divorce. This sparked a firestorm on social media and online news sites, with commentators offering wildly differing opinions as to what happened and who was to blame. Of course, it is not possible for an outsider to know exactly what happened in this incident or what the dynamics were in the relationship. However, many were quick to make accusations and blame one or the other.

In recent years more attention has been paid to the role of women in society, unfortunately also because of cases of real hatred against them.[2] In accordance with the Pew Research Center report on online harassment (Vogels 2021), women and men are similarly often abused or threatened online. However, women are more likely than men to report being sexually harassed (16% vs. 5%) or stalked (13% vs. 9%) online. Young women are particularly often affected by sexual harassment on the Internet—33% of women under 35 say they have been sexually harassed online. With the constant growth of social media and microblogging platforms, hatred of women is becoming more prevalent, creating numerous examples of how misogyny can spread almost uncontrolled (Jane 2017b; Ging and Siapera 2018, 2019).

Misogyny refers to hatred or prejudice against women and is manifested linguistically through various means, such as marginalization, bias, animosity, intimidation or violence, and objectification (Fersini, Rosso, and Anzovino 2018; Anzovino, Fersini, and Rosso 2018). A study reveals the sheer scale and nature of online abuse faced by women and provides a resource to researchers and engineers interested in exploring the potential of machine learning in content moderation.[3] In order to handle hateful content and protect people, automated systems are being used extensively to identify potentially problematic content. But a series of Failure-to-Act reports uncovers the dark side of social media platforms, more often experienced by women who are active on social media: "how harassment, violent threats, image-based sexual abuse can be sent by strangers, at any time and in large volumes, directly into your DMs without consent and platforms do nothing to stop it"[4]. Machine learning algorithms are deployed to scan content and flag it for human moderators. For instance, the Perspective API developed by Google Jigsaw was used to flag potentially toxic content for review on Wikipedia and in the New York Times comments section.[5] One challenge is to capture the linguistic specifics of hate speech, polarizing and offensive statements. Udupa observed that users of online social media platforms have managed to bypass automatic hate speech detection methods by using creative indirect forms of linguistic expression. According to Strathern et al. alternative methods to recog-

[1]https://www.chicagotribune.com/entertainment/ct-johnny-depp-amber-heard-statement-20160531- story.html

[2]https://onlineviolencewomen.eiu.com/

[3]https://decoders.amnesty.org/projects/troll-patrol/findings

[4]https://counterhate.com/research/hidden-hate/

[5]https://perspectiveapi.com/case-studies/

nize moral slurs could be successfully implemented.

Since hate is expressed in many different ways, automated methods can lack context sensitivity when determining implicit hate. To shed light on this discrepancy, we first examine which scientific theories and methods deal with the topic of misogyny. In the second step, we examine more closely how, based on theory and empirical work, classes of misogyny are built according to which content of hate speech can be assigned. In this, we assume that, in addition to a large amount of explicit hate speech, there is also a significant proportion of implicit misogynistic hate. Consequently, another goal of our study is to examine how well automated approaches to detect toxic language can identify misogyny. We collected 240,000 tweets from 2019–2021 containing the tweet handle @realamberheard and selected the top 5,000 most retweeted tweets to label and score them according to the classes identified in the literature. We then had these 5000 tweets analyzed by the Google Perspective API toxicity metric. A major outcome of this study is that online misogyny cannot be satisfactorily identified with this automated toxicity identification tool.

2 Review of Theories and Methods on Misogyny

Our study is motivated by work dealing with a) misogyny, its modeling and detection, b) the classification of hate speech and c) the verification of hate speech detectors.

2.1 Misogyny

As per Allen, there is no universally accepted definition of misogyny. When studying online anti-feminist language, different terms have been used, including "gender hate speech" (Jane 2015), "gender trolling" (Mantilla 2013), "cyber harassment" (Citron 2014), "technological violence" (Ostini and Hopkins 2015), "e-bile," and "gender cyber hatred" (Jane 2017a), as summarized by McGuirk. According to Code, misogyny can manifest in sexual and physical violence, exclusion, promotion of patriarchy, belittlement, or marginalization of women. Zuckerberg has supplemented this framework with specific forms of online misogyny. Jane identifies technological determinism as a paradigm of flaming. However, research on flaming does not show that online abuse is gender-specific (Lee 2016). In contrast, Herring and Martinson found that the "gendered nature" of online abuse messages and hate speech is significant when examining gender differences in communication styles. Online misogyny can have real-world consequences that require further investigation. Citron and Norton hypothesize that the gendered nature of online harassment and digital abuse is critical to women's online identity. Megarry has studied the psychological effects of online misogyny, including pseudonymous involvement and pullback, which limit women's online engagement.

The case of Amber Heard was the subject of a study by Whiting et al.. They conducted their study from a psychological perspective on the subject of domestic violence. The authors examined the commenting behavior of users on various social media platforms. To better understand typical types of social media reactions to allegations of domestic violence, the authors performed a content analysis on Facebook. Five main categories were extracted, namely victim blaming, perpetrator blaming, couple blaming, withholding judgment, and mixed reactions to the process. The respective main topics also contain subtopics on reactions to the allegations.

2.2 Modeling Misogyny

Determining and classifying misogyny in comments is a major challenge for humans and computers. There are various definitions and approaches to modeling this complex social and linguistic phenomenon. Fersini, Rosso, and Anzovino developed a machine learning classification approach to model misogyny. The main categories are based on gender studies theory and contain classes that are used to determine comments. The classes are: stereotyping and objectification, dominance, derailment, sexual harassment, threats of violence, and discrediting. The categorization starts after an a priori distinction of whether a tweet is classified as misogynistic or not. In a study by Farrell et al. a misogyny model was developed to examine the flow of extreme language in online communities on Reddit. Based on feminist language criticism, the author created nine lexicons that capture specific misogyny rhetoric (physical violence, sexual violence, hostility, patriarchy, stoicism, racism, homophobia, disparagement, and inverted narrative), and used these lexicons to examine how language evolves within and between misogynist groups. Recent work by Guest et al. presents a hierarchical taxonomy for online misogyny and an expert-labeled data set that allows automatic classification of misogyny content. The taxonomy consists of misogynistic content, broken down into misogynistic pejoratives and treatment, misogynistic disparagement, and gendered personal attacks.

2.3 Detecting Online Misogyny

In addition to modeling misogyny and detecting hate speech, we find studies examining how politically and socially active women are treated in current public debates. To gain insight into gender discrimination, various automated methods are used. In a study by Rheault, Rayment, and Musulan, the authors applied machine learning models to predict rudeness directed at Canadian politicians and US senators on Twitter. In particular, they test whether women in politics are more affected by online abuse, as recent media reports suggest. Another article by Beltran et al. examined gender insults towards Spanish female politicians. In an analysis of tweets written by citizens, the authors found evidence of gender slurs and note that mentions of appearance and infantilizing words are disproportionately common in texts addressing female politicians in Spain. The results show how citizens treat politicians differently depending on their gender. Fuchs and Schäfer presented the results of an exploratory analysis of misogynistic and sexist hate speech and abuse against female politicians on Twitter, using computer-assisted corpus linguistic tools and methods, supplemented by a qualitative in-depth study of abuse by four prominent female politicians in Japan. Studies suggest that voters evaluate candi-

dates from the perspective of gender stereotypes and test how this affects attitudes and voting behavior (Bauer 2015; Ditonto, Hamilton, and Redlawsk 2014; Herrnson, Lay, and Stokes 2003; Lawless 2015).

2.4 Hate Speech Classification

The annotation of hate speech is important for automated classification tasks. The classification scheme and its underlying assumptions are crucial for annotation. There are different approaches to this process such as predefined word lists or more complex models. One of the main difficulties is the definition of hate speech and its interpretation and therefore correct application. Recently, the Gab Hate Corpus was published (Kennedy et al. 2022), which uses a specially developed coding typology for annotating hateful comments. It was developed based on a synthesis of hate speech definitions drawn from legal precedents, hate speech coding classifications, and definitions from sociology and psychology. Moreover, the system includes a hierarchical clustering technique to identify dehumanizing and aggressive language, markers for targeted groups, and rhetorical features. Ben-David and Fernández researched the circulation of explicit hate speech and subtle forms of discrimination on Facebook. They contend that hate speech and discrimination cannot solely be attributed to the users' intentions and behaviors. It is also influenced by the interplay between the platform's policies, technological capabilities, and communicative practices of its users. The difficult task of capturing implicit and explicit statements was addressed in a study by Gao, Kuppersmith, and Huang. The writers suggested a technique for identifying online hate speech that employs a weakly supervised two-path bootstrapping method. This approach utilizes extensive unmarked data to overcome some constraints of supervised hate speech classification procedures, including dataset bias and the prohibitive expense of annotation. The implicitness of linguistic statements is also the subject of a work by Frenda, Patti, and Rosso. The authors proposed a number of statistical and computational analyses that support reflections on indirect propositions that focus on the creative and cognitive aspects of implicitness. In a more recent work by ElSherief et al., implicit statements were used for machine learning tasks to introduce a theoretically based taxonomy of hate speech. The research conducted by Wiegand, Ruppenhofer, and Eder focuses on identifying implicitly abusive language, meaning language that conveys abusive intent without using explicitly abusive words. Their position paper outlines the challenges in learning implicit abuse due to the limitations of current datasets and proposes changes in the dataset design to overcome these obstacles.

2.5 Bypassing Hate Speech Detection

Tricking or recalibrating automated methods results from the observation that the underlying assumptions of common machine methods do not adequately define group-specific hatred. That is, there seems to be a discrepancy between methods for operationalization tasks and the complexity of social processes. Against this background there are ways to trick hate speech detection methods or to test them for their measurement accuracy and validation. Both, cultural and associated linguistic peculiarities are thus taken into account. There are studies that try to capture culture- and language-specific hatred, which machines have difficulty recognizing. Zannettou et al. focused on examining the spread of antisemitic content. The authors carried out a large-scale quantitative analysis to discover abnormalities in language use. The results show that there are several distinct facets of antisemitic language, ranging from slurs to conspiracy theories, drawing on biblical literature and narratives expressed differently in the language. In this context, antisemitism is considered as a manifestation of hate speech, and the writers devised a technique to address it. Another investigation by Gröndahl et al. examined the efficacy of previously proposed models and datasets for categorizing hate speech. The findings revealed that none of the pre-existing models achieved satisfactory results when tested on a different dataset. The authors assert that the characteristics indicative of hate speech are not consistent across different datasets. The results show that the definitions of hate speech do not seem to be consistent and that they need further differentiation and context sensitivity. Another study by Hiruncharoenvate, Lin, and Gilbert examined ways to circumvent the observation of the state in the Chinese language, which suppresses free speech. In China, political activists use homophones (two words that are written differently and have different meaning but sound the same, e.g., brake/break) of censored keywords to avoid detection by keyword-matching algorithms. The authors claim that it is possible to expand this idea in a way that makes them difficult to counteract. One result of this work is to mathematically (and almost optimally) change the content of a post by replacing censored keywords with homophones. So, by tricking the system with linguistic creativity, they bypass the derived rules for automatic speech recognition on Weibo.

3 Overview of Misogyny Classes from the Literature

Based on the theories and methods discussed above, we have developed a classification scheme for online misogyny that covers most of the aspects discussed in the related literature. These classes include explicit and implicit misogynistic language and are presented in the following. Some of these classes are close to each other in their definitions and are not always easy to distinguish. The case study in the second part of this article will show that they significantly overlap when used for coding real-world messages. The goal of identifying misogyny classes was not to identify unambiguous definitions, but to cover a wide variety of aspects of hate against women.

3.1 Explicit Misogyny

In explicit misogynistic statements users openly attack, insult, or even threaten a woman (Waseem et al. 2017; Gao, Kuppersmith, and Huang 2017). Based on the literature presented above, we have identified the following four subcategories of explicit misogyny.

Call for action/violence. This class implies verbal

threads that intend to punish a target physically. Statements in which users call for deletion, prison, boycott, or sending the target to a psychiatric institution (Fersini, Rosso, and Anzovino 2018).

Personal insult, denigration. Personal insults and denigration intended to cause harm to a target verbally. Statements containing harmful wishes, demeaning, threatening, denigrating, inciting, defaming, use of slur words (Fersini, Rosso, and Anzovino 2018; Guest et al. 2021; Farrell et al. 2019).

Gendered personal attack. Gendered personal attacks refer to stereotypes of women. Verbal (misogynistic) attacks draw on these stereotypes. Statements that contain misogynistic speech and swearwords, revenge porn, or are sexually motivated because the target is a woman (Fersini, Rosso, and Anzovino 2018; Guest et al. 2021; Farrell et al. 2019).

Weakness of character, intellectual inferiority. Making negative judgments of a woman's moral and intellectual worth using explicit slur words. Statements that call a woman controlling, psychotic, a liar, hypocritical, narcissistic, or manipulative (Fersini, Rosso, and Anzovino 2018; Guest et al. 2021; Farrell et al. 2019).

3.2 Implicit Misogyny

Implicit statements of misogyny include cynicism and sarcasm, skepticism and distrust, insinuation, accusations, speculation and questioning of credibility, a demonstration of power, and taking a position (Waseem et al. 2017; Gao, Kuppersmith, and Huang 2017; ElSherief et al. 2021; Frenda, Patti, and Rosso 2022).

Cynicism, sarcasm. Cynicism and sarcasm represent a very derogatory attitude of a person towards others. It is expressed in an indirect form and is spiteful and bitter. Statements in which in a subliminal way, a rejecting attitude is shown (Whiting et al. 2019).

Skeptical attitude, distrust. That includes "facts" or other details to undermine a woman's account. Doubtfulness about a woman's claims or accusations. Questions whether the target had lied before and therefore cannot be trusted (Whiting et al. 2019).

Imputation. Imputation is understood as the assumption that the target behavior is motivated by flawed motivations. That includes statements that show a moral judgment, and comments where a woman is described as revenge-seeking, vindictive, attention-seeking, monetarily driven (Whiting et al. 2019).

Allegation. The category implies actions in which the evidence and allegations are challenged suggesting intentionally motivated actions. Statements of users that offer facts that refute a woman's account in spite of evidence (Whiting et al. 2019).

Speculation, denying credibility. This category includes an investigative-style attitude. Speculations and doubts about the target's behavior. In users' comments on the case, e.g., of domestic violence and its severity, we find claims about how the case might affect future reporting, users offering life stories to undermine the target's account, together with claims to personal expertise, the intent to prove something, credibility from experience, and special predictive power (Whiting et al. 2019).

Demonstration of power. The category implies a power relation between one gender and the other. Statements in which support for the man is demonstrated (Fersini, Rosso, and Anzovino 2018).

Taking position. Taking position or 'flipping the narrative' encapsulates terms and expressions that refer to the relationship between the target and the perpetrator. Statements on who is the 'perpetrator' and who is the 'victim' (Fersini, Rosso, and Anzovino 2018; Guest et al. 2021; Farrell et al. 2019).

3.3 Examples for Misogyny Classes

In order to study the prevalence of these misogynistic classes on social media, we have collected and analyzed messages addressing Amber Heard's Twitter account @realamberheard in a case study in the next section. Here, in Table 1, we show sample tweets to exemplify these classes. Since the content contains explicit hate speech and profanity, we have redacted the texts.

4 Case Study

To assess the importance of the misogyny classes presented in the article, we conducted a case study using Twitter data related to the celebrity domestic violence abuse case between Amber Heard and Johnny Depp. In the following, we describe the data and the annotation process as well as present quantitative results showing the prevalence of our explicit and implicit misogyny classes in the data.

Kennedy et al. documented that the annotation of hate speech has been shown to lead to a high level of disagreement between the annotators, see also Ross et al.. According to Mostafazadeh Davani et al. this is due to a combination of factors, including differences in understanding of the definition of hate speech, interpretation of the annotated texts, or assessment of the harm done to certain groups, i.e. inconsistent application of the definition of hate speech to different social groups.

Data. By utilizing the Twitter Academic API (Pfeffer et al. 2023) we collected 266,579 original tweets (excluding re-tweets) in January of 2022 that contained the account @realamberheard in the tweet texts. This resulted in 266,579 tweets (2019: 64,334 tweets, 2020: 117,231 tweets, 2021: 85,014 tweets). For the annotation process, we extracted 5,000 tweets that have been retweeted most often.

4.1 Annotation Process

For our case study we employed two annotators, a graduate student who is also a co-author on this paper and was instrumental in developing the misogyny classes (annotator 1), as well as an undergraduate student who was new to the topic (annotator 2). The annotators were briefed with an introduction to the topic in general and then presented with the misogyny classes. All the information presented together with coding examples was also shared in a coding manual. The manual also includes detailed descriptions of the individual coding steps and further explanations of the definition

Class	Example Tweet
Call for action/violence	Oh @realamberheard .... You ignorant witch. We ALL already know you're the guilty one here. Johnny's innocence has been proven. You're just trying to buy time, before you (hopefully) have you sit your scronny ass in a jail cell. You speak nothing but venomous lies. #JohnnyDepp
Personal insult, denigration	Seriously, how fucking sick you have to be to pull a "prank" like this on someone ? What kind of gross bitch would think pooping in people's bed is funny ? Well, apparently @realamberheard does. #JusticeForJohnnyDepp
Gendered personal attack	Not a johnny Depp fan but @realamberheard claims have more holes than swiss cheese. I dont understand females who can't make their own money and want to pocket off someone elses. It's hard to find a victim that no one sides with in todays world but I think we all call bs on AH.
Weakness of character, intellectual inferiority	Look what headline just poped up on sky news! @realamberheard you dirty little Lier! #AmberHeardIsALiar #JusticeForJohnnyDepp
Cynicism, sarcasm	@realamberheard Yes, the excitement around #JusticeLeague was huge ... definitely nothing to do with you though. Imagine being in a 4 hour movie for 5 minutes and being the most insufferable part of it.
Skeptical attitude, distrust	I just noticed the 'actor/ activist' claims in your biog @realamberheard !! Well, you certainly are an actress for real!! Only trouble is that the majority of your acting seems to be done OFF stage!! And you have set 'activism' back decades dear!! Ugh, you are some piece of work!
Imputation	@realamberheard @realamberheard Put your hand down and stop exploiting Evan's story to sway the public perception back in your favor. Don't act like you didn't break bread and hang out with Marilyn Manson for years after his relationship with ERW/ your o
Refutation	Listen bitch, I just saw a video about you demanding Depp supporter info for some legal implications!!If you want any info about me just DM me and I'll be MORE than happy to bring you upto speed!! @realamberheard I am allowed my opinion and you are scum (&u better pay my airfare!)
Speculation, denying credibility	@realamberheard You do not represent women nor survivors. I stand with Johnny Depp, Kate James, Jennifer Howell, Lily-Rose Depp, Hilda Vargas, Samantha McMillen, Katherine Kendall, Trinity Esparza and ALL THE OTHER women and men who knows your true color
Demonstration of Power	Justice for Johnny Depp outside @wbpictures studio where @realamberheard is currently filming @aquamanmovie #JohnnyDepp #JusticeForJohnnyDepp #JOHNNY #AmberHeard
Taking up a position	@realamberheard is not a victim, she is the perpetrator.

Table 1: Misogynic classes and example tweets

	Misogyny Class	Frequency	All	Misogyny
Explicit	Call for Action	681	13.6%	20.4%
(35.6%)	Personal Insult	1,649	33.0%	49.5%
	Gendered Personal Attack	730	14.6%	21.9%
	Intellectual Inferiority	1,325	26.5%	39.8%
Implicit	Cynicism/Sarcasm	367	7.3%	11.0%
(30.3%)	Skepticism/Distrust	461	9.2%	13.8%
	Imputation	556	11.1%	16.7%
	Allegation	546	10.9%	16.4%
	Speculation	305	6.1%	9.2%
	Demonstration of Power	459	9.2%	13.8%
	Taking up a Position	181	3.6%	5.4%
N			5,000	3,331

Table 2: Frequencies and proportions of misogyny classes in all 5,000 annotated tweets as well as proportions in 3,331 misogynistic tweets.

of the classes and the coding method according to the literature.

We analyzed the entire tweet at the sentence and word level, including the use of emoticons and content on the websites following URLs appearing in tweets. We looked at images, memes, or quotes, and watched linked videos. Each tweet was rated by the annotators based on all of its content. If the tweet contained statements supporting Amber Heard was neutral, or contained advertising, we annotated this tweet as *other* and ignored the tweet in the subsequent analytical steps. We used the eleven misogyny classes for annotation. After the annotation process, we created the explicit/implicit annotation from the eleven classes following the categorization described above. A single tweet could be annotated with multiple misogyny classes. If a tweet contained multiple sentences where one was implicit and one was explicit, we chose the explicit class due to the fact that a Tweet with explicit misogynistic content will be perceived as being explicit in its entirety.

Coding 11 classes with multiple overlapping definitions will lead to low levels of completely identical annotations. However, when comparing the explicit/implicit/other classes among the two annotators, the overall level of agreement between the annotators was acceptable. We can report the following values for Krippendorff's alpha (Krippendorff 2011): explicit 0.779, implicit 0.736, other 0.867.

4.2 Prevalence of the Misogyny Classes

For further analysis of this article, the annotator 1 manually compared the annotations from both annotators for all 5,000 tweets and harmonized the annotations into a single mapping of tweets to misogyny classes. The frequencies and proportions of the classes in the overall dataset as well as in the misogynistic tweets can be seen in Table 2. Shockingly, two-thirds of the most retweeted tweets addressing Amber Heard's Twitter account have been classified into explicit (35.6%) or implicit (30.3%) classes of misogyny. While explicit and implicit classes can overlap within tweets, the meta-classes explicit/implicit are mutually exclusive (see above).

5 Comparing Misogyny Classes with Google's Perspective API

Google's Perspective API is one of the standards for identifying toxic language on online platforms and is described as "the product of a collaborative research effort by Jigsaw and Google's Counter Abuse Technology team exploring machine learning as a tool for better discussions online."[6]. In this section, we will test how well the toxicity scores of this API are capable of identifying online misogyny as operationalized with our eleven classes to get an understanding of how useful these approaches can be in automatically identifying online misogyny.

We worked out the different attributes and evaluation methods of the API as the first step for comparison. In the

[6]https://www.perspectiveapi.com/research/

second step, we applied the API to the same dataset of 5,000 tweets. For each tweet, the API specifies a range of values for each of its categories. In the third step, we compared the values using statistical methods and applied network analysis to show the co-occurrence of classes and their average toxicity value reported by the Perspective API.

5.1 Attributes of Perspective API

The Perspective API predicts the perceived impact of a comment on a conversation by evaluating the comment with a set of emotional concepts known as 'attributes', namely toxicity, severe toxicity, identity attack, offense, threat, and profanity. The returned values are in the range [0.1] and are an indicator of the likelihood that something will be perceived as toxic. The higher the score, the more likely it is that the patterns in the text are similar to the patterns in comments that others have identified as toxic. The values are intended to allow developers/users to set a threshold and ignore values below that value. Values around 0.5 indicate that the model does not know if it is similar to toxic comments. The Google recommended threshold setting is 0.7. These thresholds are central to interpretation.

	Misogyny Class	**Average Toxicity**
Explicit (35.6%)	Call for Action	0.504
ø0.572	Personal Insult	0.589
	Gendered Personal Attack	0.619
	Intellectual Inferiority	0.577
Implicit (30.3%)	Cynicism/Sarcasm	0.356
ø0.493	Skepticism/Distrust	0.527
	Imputation	0.557
	Allegation	0.423
	Speculation	0.572
	Demonstration of Power	0.436
	Taking up a Position	0.581
Other (34.1%)	Marketing/PR	0.193

Table 3: Categories and Average Toxicity for Explicity and Implicity.

5.2 Measuring Toxicity for Misogyny Classes

To measure the average toxicity for the misogyny classes, we compare Google's probability score to our manual coding by summing up the codes divided by the number of tweets in each meta-class. The results show that the average toxicity score by Google for our category of explicit misogyny is 0.572. For our category of implicit misogyny, the average score by Google is 0.493. These numbers already are a strong indicator that toxicity, as identified with the Perspective API, is a poor predictor of our variable of online misogyny, and in particular of implicit hate against women. Table 3 reveals the average toxicity scores for each class. In Figure 1 we can further see the density distribution of toxicity scores for each of the meta-classes of tweets with explicit or implicit misogyny as well as others.

In the *other* sub-figure we can clearly see that there are almost no tweets that have been identified by the Perspective API as toxic that we have not also classified as misogynistic—consequently, the automated coding does not create false positives. The explicit language used for the

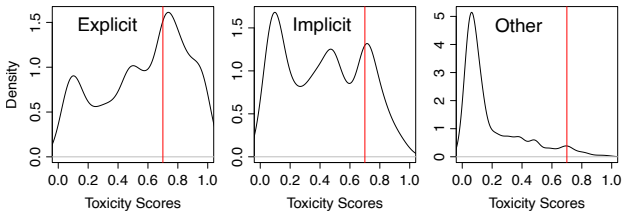


Figure 1: Distribution of toxicity scores from Google's Perspective API for tweets with explicit or implicit misogyny as well as tweets without misogynistic content.

classes that we have summarized with the meta-class *explicit* can be identified by the Perspective API to a certain degree, and the peak of the score distribution is above the standard threshold of 0.7. In other words, tweets coded with explicit misogyny contain text patterns that are similar to the patterns in comments that have been identified as toxic when the Perspective API models have been trained.

Unfortunately, the picture looks different when looking at the distribution of scores for the implicit misogyny classes. Here, the resulting toxicity scores are almost evenly distributed, having more scores with very low values than with very high values. Consequently, the tweets coded with implicit misogynistic classes do not reflect text patterns that are similar to the patterns that have been identified as toxic in the Perspective API's training data.

5.3 Co-Occurrence Network of Misogyny Classes

In addition to statistical analysis, we built a co-occurrence network that maps manual coded classes and the average toxicity scores by the Perspective API (3). Nodes represent the eleven classes and the edge value is the number of co-occurrences, i.e., the co-occurrence of classes within a tweet. The edge color is the edge value, and the node size is the proportion of the number per code divided by the number of tweets. The node color is the average toxicity value from the Perspective API where blue means low and red means high toxicity values.

In the centre, we can find the dominant four explicit classes which are identified to a certain degree as being toxic. The classes are well connected with each other. Explicit abusive statements come with similar forms of abusive language. For implicit statements, the picture looks different. In the periphery, we can find the seven classes of our meta-class implicit. Implicit misogynistic statements occur more with various forms of explicit abusive language and less among each other. In many cases, something is said implicitly, but it co-occurs with an explicit abusive statement. As mentioned above, we decided to code a tweet as explicit if both classes occurred. But the network analysis reveals the co-occurrence of explicit and implicit abusive language against women within one statement. It offers a more qualitative comparison of stereotypical hating: statements that contain a demonstration of power are associated with inferiority and insults. A skeptical attitude is associated with abusive terms of inferiority, imputation, gendered

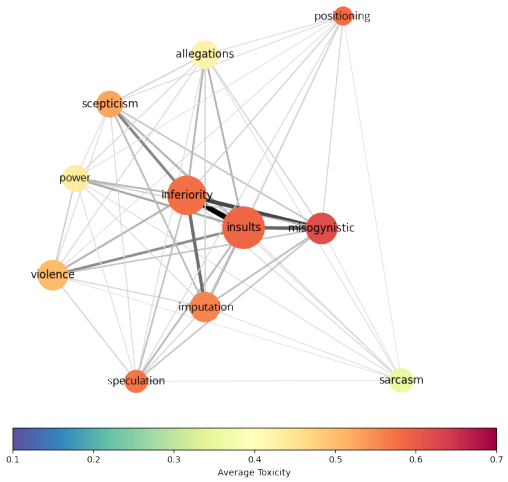


Figure 2: Co-Occurrences of Categories within a Tweet

personal attacks, and insults. Statements of speculation and doubt are associated with sarcastic and gender-attacking language. Despite the proximity of all classes, the network reveals a distinction between explicit and implicit misogyny.

5.4 Interpretation and Conclusion

We asked how well an automated approach like Google's Perspective API performs in detecting misogyny. Based on our study, two things become apparent: Google's text model does recognize explicit misogyny in the text patterns as toxic. However, the model does not recognize implicit misogyny in text patterns as toxic. The interpretation of the following tweets underlines the challenges of detecting and understanding implicit/indirect hate: "@realamberheard It's the way you think that posing this is going to change public perception of you. We heard what you did in your own words. A failure in the system isn't uncommon, so thank you for proving that male victims will never be taken seriously." A user recapitulates what has happened, draws conclusions for men, and thanks the target person for that in a very calm manner. But reading the tweet with contextual knowledge makes one understand that the thankful gesture is a cynical one. No keyword of hate can be found here; the words are all positive, but the underlying assumption is an accusation against Amber Heard and against her gender. None of the scores indicates harm in this tweet: Toxicity: 0.28, Severe Toxicity: 0.17, Identity Attack: 0.26, Offense: 0.07, Threat: 0.21 and Profanity: 0.14.

In another tweet, a user comments on what has happened and concludes that this behavior is not acceptable. The tweet contains a link to a screenshot in which impressions of what happened are reflected. Again, there is no harmful word, it all sounds positive in isolation, but clearly implies that this user is rejecting the behavior of the woman and at the same time accusing her of what she has done: "@realamberheard I had to translate to really understand where you're coming from. And no I wouldn't encourage my daughter or sister to do what you did (URL redacted)". But here as well, the scoring is very low. Toxicity: 0.20, Severe Toxicity: 0.12, Identity Attack: 0.11, Offense: 0.07, Threat: 0.16 and Profanity: 0.14.

The following example can exemplify how the toxicity score can be influenced by a single word that is interpreted as negative, even though the tweet could be interpreted as being funny: "@realamberheard @USNatArchives She will forever be known as the lady who pooped on Johnny Depp's bed." Toxicity: 0.69, Severe Toxicity: 0.15, Identity Attack: 0.74, Offense: 0.65, Threat: 0.34, and Profanity: 0.74.

There may be several reasons for this discrepancy to detect misogyny. One reason could be that there was no misogynistic content in the training texts for the human annotators. Or misogyny was never defined as an annotation class, hence, annotator could not label it. Annotators could not be informed / trained on the topic of misogyny and, therefore, could not recognize and annotate it in the texts. Although we do not know how the data sets were constructed and the model trained, we can summarize that Google's Perspective API struggles with identifying text patterns containing implicit misogynistic statements.

6 Discussion

In this manuscript, we have presented a classification scheme that incorporates 11 classes of misogyny and have described a data set that contains misogynistic content labels from Twitter. We have also provided a detailed coding book and a data set with all of the labels. The data set benefits from a detailed classification scheme based on the existing literature on online misogyny. The involvement of trained annotators and an adjudication process also ensures the quality of the labels.

We applied the classification scheme to a case related to online aggression against Amber Heard in the context of her allegations of domestic violence against Johnny Depp. For 5,000 tweets, we identified online misogyny operationalized with our eleven classes for two-thirds of the tweets, one-third as explicit misogyny, and one-third as implicit misogyny. Finally, we evaluated the reliability of Google's Perspective API for determining implicit misogyny and found that this approach can identify explicit misogyny to a certain extent, but fails with identifying implicit misogyny.

Ethical considerations and limitation. Ethical considerations must be taken into account with regard to the training and supervision of the annotator. An undergraduate student was the annotator, who underwent two steps: first, reading the typology and coding manual, and second, conducting a test on about 50 messages that had already been annotated and validated by one of the authors. Kennedy et al. pointed out the pressing concern that annotators may experience trauma or similar negative effects such as desensitization when annotating hate speech. On the basis of our own annotation experiences, we would like to highlight these thoughts. While no studies have investigated the repercussions of continuous, daily exposure to hate speech on human moderators, existing evidence suggests that being exposed to violent language and images online can adversely impact mental health, as demonstrated by Kwan et al.. We

also provided the annotator with Kennedy's suggested written guide [7] to help detect changes in cognition and avoid secondary trauma. It advises the user to take breaks and not imagine traumatic situations. The annotator was instructed to remain in communication with the study's author if she experiences any symptoms of PTSD, which are also outlined in the guide. The guide aims to normalize negative emotions resulting from work, offer education regarding trauma, identify signs of traumatic stress, and establish a support system as a preventative measure against secondary traumatic stress.

A limitation of this study is the fact that we do not know whether the Perspective API's text models contained misogynistic content and we do not know whether the data sets contained implicit/indirect forms of hate. Furthermore, we do not know whether the annotators were informed or trained on the topic of misogyny or implicit/indirect forms of hate. However, our results show that there may be a lack of information on misogyny according to existing definitions.

Google's Perspective API is a prominent tool for recognizing hate speech that uses machine learning to reduce toxicity, which is an important step towards addressing the challenge of online abuse and harassment. The API calculates the probability that a comment is perceived as toxic, reflecting Google's ambitious goal to prevent online toxicity and protect marginalized voices in conversation: "Toxicity online poses a serious challenge for platforms and publishers. Online abuse and harassment silence important voices in conversation, forcing already marginalized people offline".[8]

To evaluate the tool's effectiveness, we believe it is legitimate to directly compare it to misogyny, which represents abuse and hate according to the definition of toxicity. Given that misogyny is often subtle and has various layers, it is necessary to observe and document specific situations to collect as many characteristics as possible. We hope that by taking this approach, we can encourage the developers to adjust the tool's performance and better address the issue of online toxicity.

Implications and Future Work. Given real-world online aggression against women, it is probable that Google's toxicity model would not identify it. Thus, a huge fraction of implicit misogyny texts would stay be left in place and would not be deleted or otherwise acted upon. Misogynous behavior and target classification still remain a very challenging problem. One approach may be to create lexicons capturing specific misogynistic rhetoric and improve annotation scheme. Another challenge is to capture the peculiarities of implicity or indirect forms of hate in language. Language is very context-sensitive, and a negative tone can be expressed without a clear negative key word. Moreover, implicit sentences depend decisively on the non-linguistic accompanying signals. With our work, we would like to enhance existing research on investigating linguistic distinction between implicit and group-specific hate rhetoric. Furthermore, as we have seen from the network perspective, aside from the technical solution questions arise on how and why these different sub-classes are closely connected. From a gender perspective, we ask why are these stereotypes so consistent over time?

Given the still increasing number of users and posts in social media, automated annotation based on machine learning is inevitable. There is no other way to handle the vast volume of text. At the same time, it becomes apparent that the proportion of aggressive misogynistic speech is increasing sharply. An assessment and, if necessary, the deletion of unacceptable statements is imperative for the protection of people. Especially with regard to women, their protection is of immense importance to enable participation in public discourse and avoid withdrawing because of fear of being attacked or marginalized. However, the key to better handling the problem is to better understand the phenomenon of misogyny.

[7] https://www.apa.org/ptsd-guideline/ptsd.pdf

[8] https://perspectiveapi.com/

7 Acknowledgments

The author(s) gratefully acknowledge the financial support from the Technical University of Munich - Institute for Ethics in Artificial Intelligence (IEAI). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the IEAI or its partners.

The labeled dataset, the classification schema and the coding manual are publicly available: https://strathern.de/wp-content/uploads/2023/05/data_misogyny.zip.

References

Allen, A. 2021. Feminist Perspectives on Power. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Anzovino, M.; Fersini, E.; and Rosso, P. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In *In International Conference on Applications of Natural Language to Information Systems*, 57–64.

Bauer, N. M. 2015. Emotional, Sensitive, and Unfit for Office? Gender Stereotype Activation and Support Female Candidates. *Political Psychology*, 36.

Beltran, J.; Gallego, A.; Huidobro, A.; Romero, E.; and Padró, L. 2021. Male and female politicians on Twitter: A machine learning approach. *European Journal of Political Research*, 60: 239–251.

Ben-David, A.; and Fernández, A. M. 2016. Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain. *International Journal of Communication*, 10.

Citron, D. K. 2014. Hate Crimes in Cyberspace - Introduction. *Harvard University Press*.

Citron, D. K.; and Norton, H. 2011. Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age. *Boston University Law Review*, 91.

Code, L. 2003. *Encyclopedia of Feminist Theories*. Routledge.

Ditonto, T. M.; Hamilton, A. J.; and Redlawsk, D. P. 2014. Gender Stereotypes, Information Search, and Voting Behavior in Political Campaigns. *Political Behavior*, 36: 335–358.

ElSherief, M.; Ziems, C.; Muchlinski, D.; Anupindi, V.; Seybolt, J.; De Choudhury, M.; and Yang, D. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 345–363. Association for Computational Linguistics.

Farrell, T.; Fernandez, M.; Novotny, J.; and Alani, H. 2019. Exploring Misogyny across the Manosphere in Reddit. In *Proceedings of the 10th ACM Conference on Web Science*.

Fersini, E.; Rosso, P.; and Anzovino, M. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *EVALITA@CLiC-it*.

Frenda, S.; Patti, V.; and Rosso, P. 2022. Killing me softly: Creative and cognitive aspects of implicitness in abusive language online. *Natural Language Engineering*. Publisher: Cambridge University Press.

Fuchs, T.; and Schäfer, F. 2021. Normalizing misogyny: hate speech and verbal abuse of female politicians on Japanese Twitter. *Japan Forum*, 33(4).

Gao, L.; Kuppersmith, A.; and Huang, R. 2017. Recognizing Explicit and Implicit Hate Speech Using a Weakly Supervised Two-path Bootstrapping Approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*.

Ging, D.; and Siapera, E. 2018. Special issue on online misogyny. *Feminist Media Studies*, 18.

Ging, D.; and Siapera, E., eds. 2019. *Gender Hate Online: Understanding the New Anti-Feminism*. Springer International Publishing.

Gröndahl, T.; Pajola, L.; Juuti, M.; Conti, M.; and Asokan, N. 2018. All You Need is "Love": Evading Hate Speech Detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, 2–12.

Guest, E.; Vidgen, B.; Mittos, A.; Sastry, N.; Tyson, G.; and Margetts, H. 2021. An Expert Annotated Dataset for the Detection of Online Misogyny. In *Association for Computational Linguistics*, 1336–1350.

Herring, S. C.; and Martinson, A. 2004. Assessing Gender Authenticity in Computer-Mediated Language Use: Evidence From an Identity Game. *Journal of Language and Social Psychology*, 23(4). Publisher: SAGE Publications Inc.

Herrnson, P. S.; Lay, J. C.; and Stokes, A. K. 2003. Women Running "as Women": Candidate Gender, Campaign Issues, and Voter-Targeting Strategies. *The Journal of Politics*, 65(1).

Hiruncharoenvate, C.; Lin, Z.; and Gilbert, E. 2015. Algorithmically Bypassing Censorship on Sina Weibo with Nondeterministic Homophone Substitutions. *Proceedings of the International AAAI Conference on Web and Social Media*, 9: 150–158. Number: 1.

Jane, E. 2017a. Gendered cyberhate: A new digital divide? In *Theorizing Digital Divides*. Routledge.

Jane, E. 2017b. *Misogyny Online: A Short and (Brutish) History*. Sage Publications.

Jane, E. A. 2015. Flaming? What flaming? The pitfalls and potentials of researching online hostility. *Ethics and Information Technology*, 17: 65–87.

Kennedy, B.; Atari, M.; Davani, A. M.; Yeh, L.; Omrani, A.; Kim, Y.; Coombs, K.; Havaldar, S.; Portillo-Wightman, G.; Gonzalez, E.; Hoover, J.; Azatian, A.; Hussain, A.; Lara, A.; Cardenas, G.; Omary, A.; Park, C.; Wang, X.; Wijaya, C.; Zhang, Y.; Meyerowitz, B.; and Dehghani, M. 2022. Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56.

Krippendorff, K. 2011. *Computing krippendorff's alpha-reliability*. Retrieved from: https://repository.upenn.edu/ascpapers/43.

Kwan, I.; Dickson, K.; Richardson, M.; MacDowall, W.; Burchett, H.; Stansfield, C.; Brunton, G.; Sutcliffe, K.; and Thomas, J. 2020. Cyberbullying and Children and Young People's Mental Health: A Systematic Map of Systematic Reviews. *Cyberpsychology, Behavior and Social Networking*, 23(2).

Lawless, J. L. 2015. Female Candidates and Legislators. *Annual Review of Political Science*, 18.

Lee, H. 2016. Behavioral Strategies for Dealing with Flaming in An Online Forum. *The Sociological Quarterly*, 46(2): 385–403.

Mantilla, K. 2013. Gendertrolling: Misogyny Adapts to New Media. *Feminist Studies*, 39(2): 563–571.

McGuirk, O. 2021. Where Have All The Good Men Gone? An Exploration of Misogyny and Anti-Feminist Discourse in the 'Mansophere'. *Thesis, Dun Laoghaire Institute of Art, Design, and Technology*.

Megarry, J. 2014. Online incivility or sexual harassment? Conceptualising women's experiences in the digital age. *Women's Studies International Forum*, 47: 46–55.

Mostafazadeh Davani, A.; Atari, M.; Kennedy, B.; Havaldar, S.; and Dehghani, M. 2020. Hatred is in the Eye of the Annotator: Hate Speech Classifiers Learn Human-Like Social Stereotypes. In *31st Annual Conference of the Cognitive Science Society (CogSci)*.

Ostini, J.; and Hopkins, S. 2015. Online harassment is a form of violence. *The Conversation*, 8: 1–4. Publisher: The Conversation Media Trust.

Pfeffer, J.; Mooseder, A.; Lasser, J.; Hammer, L.; Stritzel, O.; and Garcia, D. 2023. This Sample seems to be good enough! Assessing Coverage and Temporal Reliability of Twitter's Academic API. In *Proceedings of the 17th International AAAI Conference on Web and Social Media*.

Rheault, L.; Rayment, E.; and Musulan, A. 2019. Politicians in the line of fire: Incivility and the treatment of women on social media. *Research & Politics*, 6(1).

Ross, B.; Rist, M.; Carbonell, G.; Cabrera, B.; Kurowsky, N.; and Wojatzki, M. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In Beißwenger, M.; Wojatzki, M.; and Zesch, T., eds., *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, 6–9.

Strathern, W.; Schoenfeld, M.; Ghawi, R.; and Pfeffer, J. 2022. Identifying Lexical Change in Negative Word-of-Mouth on Social Media. *Social Network Analysis and Mining*, 59(12).

Udupa, S. 2020. Artificial Intelligence and the Cultural Problem of Extreme Speech. *Social Science Research Council, (20 December 2020), online*.

Vogels, E. A. 2021. The State of Online Harassment. *Pew Research Center*.

Waseem, Z.; Davidson, T.; Warmsley, D.; and Weber, I. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, 78–84. Vancouver, BC, Canada: Association for Computational Linguistics.

Whiting, J.; Olufowote, R. D.; Cravens-Pickens, J.; and Witting, A. B. 2019. Online Blaming and Intimate Partner Violence: A Content Analysis of Social Media Comments. *The Qualitative Report*, 24.

Wiegand, M.; Ruppenhofer, J.; and Eder, E. 2021. Implicitly Abusive Language – What does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 576–587. Association for Computational Linguistics.

Zannettou, S.; Finkelstein, J.; Bradlyn, B.; and Blackburn, J. 2020. A Quantitative Approach to Understanding Online Antisemitism. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 786–797.

Zuckerberg, D. 2018. *Not All Dead White Men. Classics and Misogyny in the Digital Age*. Harvard University Press.

5.5 Summary

The preceding chapters considered various aspects of how to approach communicative behavior in conflicts methodologically and theoretically. The theoretical part presented the basics of social media networks, research on social processes, and language and communication. In Chapter 3 we presented the methods for data analysis. In the previous sections of this chapter, the published paper were presented. This section provides an overview of how the published articles are embedded into the overall questions addressed in this thesis.

Structural Level

How can we map change in communication conflicts? What are the properties of negative word-of-mouth?

Chapter 5 presented work on the phenomenon of firestorms on social media platforms, which occur when a rumor or scandal spreads rapidly and gains significant attention, resulting in reputational damage to individuals, brands, or governments. In Strathern et al. (2020b), we were interested in detecting the outbreak of firestorms and identifying their starting point, emotional evolution, and linguistic features. The study analyzed 21 Twitter firestorms using network analysis and text statistics to identify features that indicate change or anomaly and linguistic peculiarities of firestorm tweets. Results show that the maximum in-degree feature best detects a change. In firestorm tweets, personal pronouns, netspeak, and affective words categorized as negative were significantly different. The study sheds light on the importance of understanding firestorms' dynamics and linguistic features (Strathern et al., 2020b). Results have shown that a change in the lexicon at a certain point indicates interpersonal differences in language use, including a switch in pronouns, use of Netspeak, and a decrease in positivity with an increase in negativity. To identify these underlying changes, we built a change detection model. For the model we applied a dictionary-based approach and statistical testing to extract linguistic features and map the trajectory of emotions. The challenge was in mapping emotions and translating language into numerical data. While emotions did change, switching pronouns and using Netspeak were stronger indicators of a shift in language use. The time factor was also an important consideration in the design of the testing process. These prevention techniques for emerging conflicts can help identify and intervene before hate speech and polarization can escalate. We have used individual

tweets to generate aggregated time series data. However, the challenge lies in finding the right balance between the number of tweets required to produce a signal and the temporal resolution needed for accurate event prediction. It is crucial to avoid a time span that is too large for high temporal resolution. Additionally, the change point's proximity to the time series's characteristic time depends on the chosen interval. To create time series data, we employed statistical methods, such as the elbow criterion and the penalty parameter method, which are particularly effective for change point detection. Our approach worked well for both streaming and linear data. We used the Linguistic Inquiry and Word Count (LIWC) tool to track personal pronoun usage in tweets per second, and we aggregated these data points every half-hour to keep memory compact. We then formed a sequence of 48 numbers over a 24-hour period, with the list being updated by one at the front and one down at the back. It allowed us to monitor and track anomalies in the data stream continuously. We assumed the change point with the history of the last 24 hours and the deviation. Within a half-hour interval, we observed the composition of personal pronouns and track their changes over time. We also described how the characteristics of the tweets have changed over time by analyzing the history, current value, and the new half-hour interval. Updating the time interval created a kind of streaming effect, leading to the identification of change points. Through this approach, we developed a method to detect anomalies in the data stream and created a procedure to help identify significant events and changes (Strathern et al., 2020b). In Strathern et al. (2022b) we extracted linguistic features such as pronouns, Netspeak and network features to compare the performance of different prediction models. Results show that all models perform well, the ones containing network features are slightly stronger. Nevertheless, adding linguistic features increases the overall performance. Negative communication is characterized by pronoun switch and the use of Netspeak. Change can be mapped by using these properties. Comparing these two methods, we can conclude that the detection model better reflects the conceptual idea of change processes. In contrast, the prediction model reflects better the pure accuracy of algorithmic performance rather than the idea of understanding underlying change in conflicts.

What are the structural properties of polarized communities? How can we identify interpersonal differences and similarities in language use and style?

Chapter 5 presents work on the effects of polarization in political communities

on Reddit (Strathern et al., 2022a). To conduct this study, we collected data from three subreddits, two of which were prohibited, and compared user activity and language usage. To investigate the polarizing effects, we posed the question of whether a temporal absence from these subreddits would change the level of activity, diversity, and use of profanity in language. We designed an experimental setup, created groups based on the activity level, and observed changes in activity after an absence compared to those who remained continuously active. We analyzed changes in language use for each group, using metrics such as lexical diversity and the frequency of profanity words. The findings indicate that users who remain continuously active increase their activity and decrease their lexical diversity, while users who take a break from the subreddit reduce their activity and increase diversity. These initial findings provide a basis for further research on interventions that could slow down or interrupt certain processes. Future research could explore intervention techniques in more detail on social media platforms (Strathern et al., 2022a). Results indicate that structural properties of polarized environments can be well captured by the experimental setup considering activity and metrics such as lexical diversity.

Content Level

How can different linguistic layers of hate be identified? What are the distinct properties of communicative aggression?

Chapter 5 also addresses the issue of misogynistic language on Twitter (Strathern and Pfeffer, 2023). The study investigates whether automated methods to detect toxic language can identify misogyny. The catalyst for this research was the online harassment of actress Amber Heard, who faced a Twitter firestorm after accusing her then-husband Johnny Depp of domestic violence. To answer this question, we conducted a literature review of theories and methods to detect misogyny. Additionally, we analyzed a sample of 500 top hateful tweets against Amber Heard to develop a classification schema that identifies implicit and explicit misogyny. We then applied this schema to a dataset of the 5,000 top-most retweeted tweets and manually annotated it. Finally, we compared our manual coding with a toxicity measurement from Google's Perspective API. The results indicate that while Google's API performs well in detecting explicit hate, it performs poorly in detecting implicit misogyny. Our study employed a design that included content analysis, statistical testing, a classification schema, and

mixed methods. We focused on individual cases to create a taxonomy for abusive language. However, we faced challenges in delimiting explicit from implicit abusive language and comparing our findings with existing toxicity parameters. For instance, Google annotates content based on the perception of hate, whereas we were interested in the actual presence of abusive language. This discrepancy highlights the methodological challenge of determining what constitutes abusive language (Strathern and Pfeffer, 2023). Nevertheless, our approach followed our theoretical considerations in Chapter 2. A limitation of this study is that we do not know whether the Perspective API's text models contain misogynistic content. We do not know whether the data sets contained implicit/indirect forms of hate. Furthermore, we need to determine whether the annotators were informed or trained on misogyny or implicit/indirect forms of hate. However, our results show that there may be a need for more information on misogyny according to existing definitions. It is an important area for future research in understanding the prevalence and impact of misogyny in online discourse (Strathern and Pfeffer, 2023). In Wich et al. (2021), we examined the properties of users that use offensive language. For that purpose, we classified users whose profile information indicated support for two German right-wing parties and collected their tweets. Based on a content analysis of tweets, we developed a classification schema with seven main categories expressing distinct hate forms: ethnicity, nationality, sexuality, gender, religion, disability, and class. Different layers of hate can be captured well by a in-depth content analysis and by building up on theories. Developing classification schema that were rooted in zooming in specific hateful situations are a valid complement to automated rather pure quantitative methods.

Chapter 6

Discussion and Outlook

6.1 Social Media Data for the Study of Human Behavior

Through a number of studies on the structure and dynamics of conflict-ridden online communications, Chapter 5 presented methods for analyzing large-scale data and addressed the research questions of this thesis. This chapter examines these questions more thoroughly. It is primarily concerned with evaluating quantitative methods for assessing the emotional valence of opinions and attitudes expressed in social media texts. In general, the methods have proved effective in determining the changing positive and negative valences of social media texts. The issue of monitoring online hate speech based on user-generated content has become increasingly important for various entities as a response to the alarming rise of online abuse and hateful language that target vulnerable groups on social media platforms. This type of abusive language encompasses a range of hostile messages that intimidate or incite violence and hatred towards certain communities and can even be found in other forms of online text. Despite various approaches proposed in recent years to identify and monitor hateful content, the problem persists due to the complexity of abusive language and its implicit forms. Our investigation focused on the methodological approaches to automatic identification of abusive language across different situations to examine how hate speech or negative communication is expressed and processed. Our analysis of classification methods revealed significant challenges related to the implicit nature of abusive language, which can be expressed through different forms of language. To address these challenges, we proposed solutions applicable to various situations, recognizing the difficulty in automatically inferring creative aspects of abusive language. Sarcasm, for instance, can disguise hurtful messages,

especially in informal and short texts like those on Twitter, which can affect the accuracy of the detection systems. Our hypothesis was that incorporating linguistic knowledge into detection models can help capture implicit meaning and improve the detection of hateful messages. An alternative approach is to consider the change process in situations of conflict. A sudden twist is captured best with behavioral data such as pronoun use. Here we discuss the most prominent research finding - the methodological examination of distinct linguistic markers present in the text - and suggest directions for future research.

6.2 Modeling Social Media Communication

The primary objective of this research has been to investigate the online occurrence of communication conflicts from a diverse range of methodological angles. Recognizing the intricacy of this phenomenon, we employed various approaches and directions to gain a comprehensive understanding of it. To detect the features of negative communication, we leveraged automated classification tasks. We also conducted experimental tests to examine the language characteristics of online communication conflicts. Furthermore, we built a model to validate existing tools. Our multifaceted approach enabled us to gain a nuanced understanding of the nature and impact of negative communication online. The focus of the analyses has been on the characteristics of language and on how opinions and attitudes spread through interaction in social networks. Insights into the structure and dynamics of interaction networks are provided based on empirical data collected from social media networks. The specific goal of the thesis is to gain a better understanding of communication behavior and content in opinion-forming processes characterized by conflicts, with a focus on linguistic markers and patterns. The theoretical part introduced the concepts relating to networks, social behavior, and language behavior used to analyze the data. Subsequent chapters compare various methods of quantitative text and content analysis, experiments, and techniques of network analysis for measuring opinion processes. For this purpose, network data, particularly text data from social media users, were collected, formatted for evaluation, and analyzed for content, structure, and dynamics. The main focus of these analyses is the comparison of different language situations at different time points. Another focus is on the content and linguistic differentiation of text data and the methods of preparing them for quantitative analysis. Following the analyses, properties from text analysis

were used to develop a model for capturing processes of change. In addition, insights from theoretical social concepts and language behavior were used to experimentally test effects on communication behavior. Furthermore, theoretical findings from the experiment were applied to develop a classification scheme for capturing misogyny in content and language and to compare it with automated methods. The specific results of the work established the structural significance of similarity in behavior, as well as its possible differentiation through language. In this thesis, we have endeavored to represent and analyze human behavior using social media data in various ways. Our primary interest has been to explore how human behavior can be measured through speech, text and network data, and how we can extract content from this data. To achieve this, we have applied various methods as elucidated in Chapter 4. Our goal has been to measure and map behavior with empirical methods. The thesis also presents a critical discussion of methods for analyzing text quantitatively. The core consideration of our study was to approach the content of social media posts in a way that would enable us to gain insights into behavior in specific situations. However, the data format itself has presented a tremendous challenge. Social media text data is often unclear, as it does not consist of edited, topic-specific, consistent texts framed within a fixed context. Social media comments are expressions of opinions and, as such are comparatively unstructured and unclear. People text, comment and post their opinions on topics, people, personal interests, personal feelings, political actions, well-being, sports, music, religion, politics, and sex. As a reflection of people's interest, the range of topics is inexhaustible. Social media allows for the exchange of ideas on any interest and niche topic. Everything is discussed and shared as long as it finds an audience interested in it. As we explain in Chapter 2, given the niche-character of social media discussions, echo effects are prevalent on social media. Social media mirrors human behavior as seen in the offline world, but transposed to the online world. Behaviors are similar, but different mechanisms promote and faciliate the spread of social media processes. We have elaborated on this in detail in Chapter 2. The difference is that platform operators can capture behavior and provide traceability. Each platform has its own formal construct that enables or disables a certain behavior. Social actions such as commenting, liking, sharing, mentioning, following, and blocking, which are formalized in technical functions such as buttons (Marres, 2017) reinforce mechanisms. The platform's own rules and regulations define the forms of possible and permissible behavior, and the specific purposes of online

communications. For example, Twitter is best suited for communication from one to many. It resembles shouting aloud in a room: if someone hears it, perceives it, likes or dislikes it, a reaction will follow. However, interactions are limited to a particular topic, and users come and go, contribute, and then disappear from a conversation. Future research is required on identifying the types of communication structures and the forms of discussion progress. In this thesis, we have concentrated on comments on content with the purpose of expressing one's opinion, rather than elaborated exchanges. When working with Reddit, we found that the platform architecture allows for a much stronger communal exchange than other platforms. However, for the analysis of behavioral data, it was necessary to examine both the structural mechanisms of the platforms and the content level of the comments, and to use appropriate data and methods. Based on theory and empirical findings our approach to social media communication structure is represented in Figure 13.

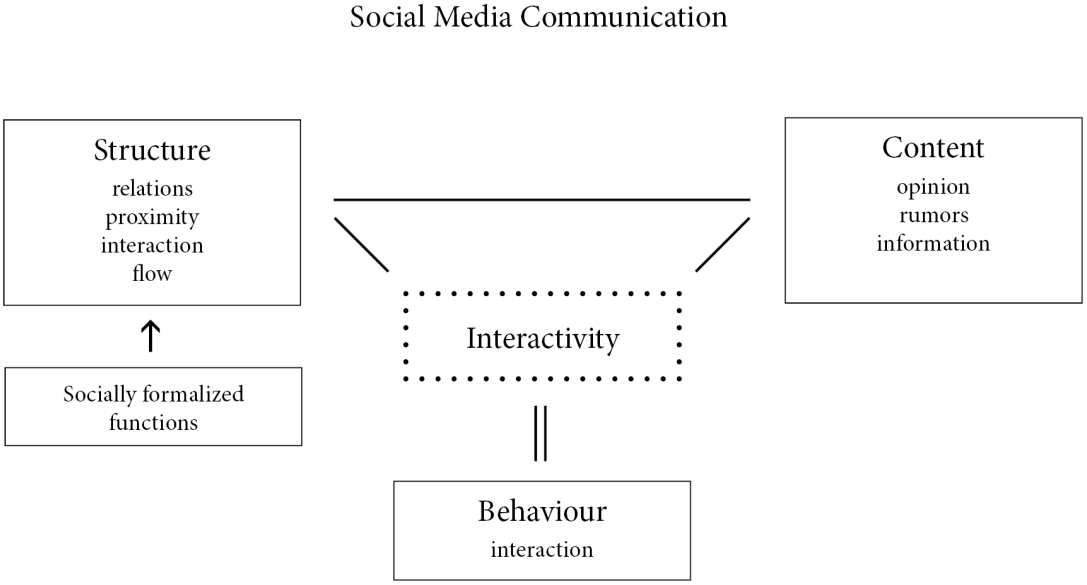


Figure 13: Model of social media communication

We refer to Borgatti's work on social media networks and Labianca's extension to this approach. The core distinction lies in structure and content as elaborated in Chapter 2. For methodological considerations we have added interactivity as a core mechanism that enhances certain behaviors. Following Marre's work on digital sociology, we have also added the socially formalized functions that enable structural characteristics such as relations, proximity, interactions and flows. The platform architectures, with their various socially formalized functions, al-

low for a quantitative approach. When collecting data on social media behavior, the important question is what types of information should and can be extracted from the available data. Therefore, it is not entirely correct to say that the data is unstructured. Metadata, in particular, imposes a structured format on the data, as explained in Chapter 4. While we may not find structured discussions on social media within a specific framework, such as political speeches or panel discussions, we do find discussions revolving around a specific topic. Zooming in on a specific topic enables us to retrieve social media data for a specific purpose, and then develop a specific measurement approach, as explained in the introduction (Lazer et al., 2021). As discussed, we approach the data on two levels: structure and content. On the structural level, we are interested in the functions and mechanisms that the system offers. The question we are interested in are: What types of relationships can be built through the platform? What types of interactions are possible? What are the functions through which contents are spread? How can comments be made? The user's framework of action is determined by these socially formalized functions described earlier. It should be noted that the system architecture provides users with immense freedom to communicate in their own way. Chapter 3 on Netspeak illustrates how technology and language can be combined to create something new. This means that when users try out new things, it is also possible for platforms to adjust and adapt their socially formalized functions. This involves a constant process of change and transformation which is reflected in the data structures adopted. This means that the system architecture sets the structural framework within which users operate. Social interaction is characterized by strong dynamics, as discussed in Chapter 2. The interactivity between humans, machines, and content has a strong process character, which can be observed both on a structural level and a content level. On the content level, we are interested in how the opinions in the texts are expressed linguistically. The system architecture also influences linguistic expressions, as the form of communication affects language. Twitter allows only 280 characters, while Reddit allows for extended exchanges with undefined character lengths. This results in different communication formats and observation cases. As discussed in Chapter 2, certain fundamental principles come into play in human interaction. These social processes allow for developments to emerge, which are also marked by strong negative dynamics. Negative dynamics can lead to extreme positions or polarization in people, groups, and topics. In Chapter 2, we discuss the structural properties of language in aggres-

sive communication. Language is our tool for capturing negative processes. In Strathern et al. (2020b, 2022b), we see that variation in language is an indicator of social processes. This leaves room for considerations on whether and how interventions can and should be conducted in times of conflict. In Strathern and Pfeffer (2023), we also see the different layers that language can have. Language is a driving force in social processes. Although we cannot demonstrate that it is a driving force, we can show how processes can be mapped through language, and that leaves room for further research questions. As described at the beginning, we look at behavior from two perspectives: interactions on the structural level, and language on the content level. This corresponds to the characteristics of the platforms and their technical conditions. This perspective allows for, and requires, various methods. As explained in Chapter 3, the methods we use come from quantitative text analysis, experimental methods, quantitative content analysis, network text analysis, and social network analysis. In addition, machine learning techniques have been applied. One of our initial assumptions has been that negative processes induce mood shifts and people become angry. For this reason, we have analyzed emotions in tweets. As described in Chapter 3, the analysis is based on a metric that combines various word types and categories to determine the level of positivity and negativity. Developed by a linguist and psychologist and further refined by several collaborators, the method has considerable potential. It is an established technique which we first used in analyses of the ways emotions develop (Strathern et al., 2020b, 2022b). For this purpose, as described in this chapter, we use a dataset consisting of 21 firestorms. Since text comments are hard to analyze automatically, we need to enrich text with further information. The first task is to process the words numerically in order that they can be quantitatively measured. This methodological approach involves a non-trivial step, as it always involves an assumption in order to make the data interpretable. In the processing stage, a new level of abstraction is added to the data that influences both the model and the interpretation later on. We attempt to capture the content level, categorically by annotating the words in the text data. By adding information in this way, complexity is simultaneously reduced. The annotation process is an important step in the processing of text and language. However, on the content level, enriching information is essential and we emphasize the need for data processing steps as part of the overall method. Important decisions are made with consequences for the interpretability of data. The data-driven approach is not free from assumptions (Lazer et al.,

2021; Ruths and Pfeffer, 2014). In order to capture the multidimensionality of social media processes and data, we apply techniques from complex system theory. According to the definition commonly used in systems theory, complex systems are characterized by a large number of system elements and different relations between these elements. This extremely flexible definition can be specified by indicating typical (but neither necessary nor sufficient) characteristics of complex systems. These include, for example, non-linearity, emergence, self-organization, heterogeneity, path-dependence, feedback, and the existence of attractors. What these properties have in common is that they cannot be grasped immediately and are difficult to formalize. Understandably, modeling such systems often (but not necessarily) leads to complex models. The actual complexity of the model is determined, in addition to the complexity of the reference system, by the modeling goal that reflects the problem to be solved. In general, with every modeling of a complex system, there will also be a reduction in complexity, since the causal relationships in the system itself are "too complex" for the human observer. In addition to the simple aggregation of system elements and the abstraction of selected system attributes, modeling complex systems must also attempt to consider the aforementioned properties in a simplifying manner. All of these simplifications involve modeling assumptions. Since the exact nature of nonlinearity, emergence, self-organization, etc. is a priori typically just as unknown as the "right" aggregation and abstraction, the modeler usually makes commonly accepted assumptions about these properties. The assumptions are often no more than plausible and sometimes even counterfactual. Let us recall our central research questions and hypothesis - that variation in language is an indicator of change/processes. In Strathern et al. (2020b) and Strathern et al. (2022b) we assume that users' emotions change in the course of a firestorm. To establish this empirically we have to transform 280 characters into a numerical value. We apply a sentiment dictionary to determine the emotional value of each tweet. In addition, each tweet is analyzed in terms of its lexical structure, as described in Chapter 3. The distribution of parts of speech and word classes provides initial descriptive insights. Based on the dictionary, tweets are classified according to text properties. Each tweet is segmented into word units. Based on these values we compared different time slots. Text properties do not reflect much content and in fact the emotional tone does not change greatly from one time slot to another. However, the purpose of the technique to add linguistic information to text. Since we are also interested in change and prediction, the

second purpose is to extract features that are prevalent and useful for automated detection. We calculate differences in pronoun usage in different time periods and establish that during firestorms the number of pronouns are significantly higher. At other times users tend to write only about themselves. Accordingly, during times of conflict, users change their perspective and the type of pronouns they use. This is somehow an intuitive result, one we can confirm through the parameters of our model. It is not so much the affects that are conveyed through words, but rather the perspective of users. The significant finding was that at a certain point, users switch pronouns and instead of writing about themselves, they target another person. We developed a model with which we can represent our findings. Given the unstructured format of text data and the huge amount of data, the high number of pronouns and the type of pronouns used turns out to be a strong indicator of the user's sentiments and attitudes. For our model we defined a line that represents our baseline and which is interrupted at the point where the number of tweets is at its highest. And for this window, we observe a change in language. We could have looked at the spike in the number of tweets a little earlier, but we defined this time window for our model and used pronouns as features because there was a significant difference in times before compared to times during firestorms. The fact that in hate speech and other "negative communications" in Twitter people use more third-person pronouns than normally is not in itself surprising because normally Twitter users write about themselves and what they are doing. What is surprising is the fact that this turned out to be an effective and reliable indicator of hate speech and Twitter firestorms when Twitter texts were subjected to computer analysis. It was surprising because hitherto, and presumably no one else had not realised that this correlation, unsurprising in itself once explained, could be used to get the computer to identify hate speech and firestorms in very large volumes of Twitter text. This part, incidently, is the most interesting for me – the account of computational methods used to track and analyze text; interesting because the meaning or semantic aspect of written text is something computers are not good at. In sentiment analysis the decision as to what is positive or negative is a binary choice. There is little space for contextualization. For an initial impression of tone in tweets it is a valid method. However, dictionary methods become more interesting when combined with network information (structural level). Another interesting finding on the content level is the use of the linguistic variant Netspeak. The combination of keyboard language, emoticons, and brevity allows

for creative developments. There are several research questions on which future research could built on. Based on our classification we extracted lexical items, with which we can map smallest changes in text and speech. In the beginning I assumed it would be possible to make stronger statements about collective behavior. However, this cannot be conclusively asserted. The switch of pronouns is what we defined as an indication of behavioral change. In contrast to the method used in Strathern et al. (2020b), we used several linguistic features and asked whether we can predict the start of a firestorm. We built a baseline model with several basic features, a mention-network model, amd a retweet-network model. The linguistic model extends the basic model by including linguistic features, the mean value of extracted features from the psycho-linguistic dictionary (mainly pronouns, Netspeak, positivity, negativity). As we can see here, we use different information from the structural and content levels in order to predict the onset of a firestorm. We asked whether we can predict the start and can positively confirm this. Though the network features are slightly stronger, adding linguistic information increases the predictive power. Especially in terms of the interpretability of models, the process of using and enriching information is an important step as it already contains personal assumptions. For the scope of these studies – defining lexical properties, detecting change and prediction – social network analysis, machine learning and text statistics are powerful computational methods. We processed text data with a dictionary, classifying each tweet to make assumptions about the underlying speech structure. Due to brevity and affective nature of the tweet comments this is a valid and handy method. Even if the dictionary does not capture the context, a firestorm - in its nature – has little context. A firestorm reflects a binary yes no acting of user, to me it seems justifiable to use a classification schema that simply accounts for positivity and negativity. As described in Chapter 2 interactions between people are subject to constant change, and the goal of this work has been to measure behavioral change as such. In order to do so, we need to define different situations in which we assume we can identify differences and similarities. We define polarization as a dynamic process, which is assumed to be in operation. We further assumed that activity is a central behavioral aspect in social media. Platforms are made for constant interaction, whether with content or with other users. Another assumption is that a temporal absence from these platforms – an inactivity – reduces activity. Reduced activity would, on this assumption, diminish echo chamber effects that occur when people are constantly exposed to the same

people and content. Hence, in Strathern et al. (2022a) we developed a quasi-experimental setup to test these predefined possible effects. For this purpose, we analyzed communication behavior in political subreddits. We selected three subreddits, one left-leaning, one right-leaning and a neutral one. We are interested in their interactivity and their use of language before and during an absence. We built two groups with each having the same level of activity. We calculated who was absent, who was not and matched them accordingly. The biggest limitation of the quasi-experiment is that group membership is not randomized. And the setting here is, of course, quite different from typical laboratory settings. The notable thing about this quasi-experiment is that the behavioral data – while writing the thesis – was freely available. As mentioned above, access to data has become more and more limited. The data from Reddit contains a large amount of behavioral data. In addition, there are posts and comments containing news, opinions, pictures, links. For the linguistic analysis, we were only interested in text data. This means that from the vast pool of data, a database of the social network communication, we define what we want to extract. In our case, we were interested in interactions, absences, and text data. For a text analysis, we segmented text data based on a linguistic metric with which we can measure the level of lexical diversity. We assumed that activity means being exposed more to the same content and people in consequence leading to less diversity. Furthermore, we counted the number of swear words in these comments. We compared the test and control group for differences in activity and language use. Users who are constantly active become less diverse in their language use and more active in terms of posting whereas users who are absent become more diverse in their language use and show less activity. We developed an experimental setup and applied a linguistic metric with which we were able to capture this change. This is of course just one way of representing communication behavior with social media data. The vast amount of data cannot be handled manually, hence different techniques need to be applied. From a linguistic point of view there are strong limitations on the extent to which statements about validity can be made. Consequently, in aggregating posts and comments, we do not consider text and sentence length and discursive structure. Nor do we consider the sequence of threads which would otherwise be of interest. Platform-independent studies would certainly be of interest. Even if more platforms refuse to give researchers further access to data, research on platform effects and social processes is still relevant given the high number of users. The system architecture is crucial

for communication processes. However, the distinction between structural and content-related levels is important as as we outlined above. Another important distinction is that between the platform characteristics and their linguistic and discursive design, and this has implications for methodology and measurement procedures. In this context, we have used methods in our work that allow for a distinct observation and analysis. The diversity of methods and their application in this work should be emphasized at this point. Through social network analysis, we can capture relationships and structures, and through quantitative methods of text analysis, we can capture their linguistic and content-related features. In our latest study (Strathern and Pfeffer, 2023), we focused on evaluating existing models for hate speech and their automated detection. One observation was that these models are one-dimensional in their assumptions. The question was how we could approach this methodologically. The starting point was the Google Perspective API, a toxicity tool that calculates the probability that a text will be perceived as toxic. This tool is freely available to users and companies for evaluating text content. The classification scheme is publicly available on the Google website. As discussed in Chapter 2, aversion, violence, and hate manifest themselves linguistically and communicatively in various ways and levels. A clear definition is not yet available, since it depends on the type and manner of the attack, as well as on the perception of the attacked. The schema in Chapter 2 by Culpeper illustrates the distinct properties of different forms of aggression, while also explaining the perception of violence. The starting point was an online firestorm on Twitter in the context of abuse allegations between a famous celebrity couple, Amber Heard and Johnny Depp. A zooming-in in a specific situation allows for a deeper understanding of what hate can be and is a good complement to large-scale analyses (Lazer et al., 2021). A detailed analysis can reveal specific properties that require context knowledge and linguistic expertise. We did a quantitative content analysis, referring back to the literature and theory, to capture the different layers. Through semantic, syntactic, and categorical differences, we were able to identify the multidimensionality of linguistic violence. As explained in Chapter 2, the various linguistic levels can be applied to a text like a template, systematized, and categorized. We came up with different categories and annotated 5000 tweets. By choosing these categories we came up with a classification that models our understanding of misogyny. Through a comparison using the perspective API we ascertained that these dimensions only partially overlap. Our model tries to reflect the multidi-

mensionality in a differentiated consideration of language. To simplify, we also subsume the micro-categories into two macro-categories, which enables us to capture the different levels. We consider distinct properties of implicit and explicit expressions. This is a way to reduce the dimensions of text data in order to make them interpretable and to capture complex concepts such as misogyny. The resulting taxonomy serves knowledge representation and makes data interpretable and measurable. It is a step towards quantifying content. In summary, it can be said that text data poses a great challenge due to their context sensitivity. Their dimensions need to be reduced in order to work with them methodically. Our approach was primarily to split social media data, which contain more than just text data, into structural data and content data in order to make statements about behavior combining the different data formats. Our hypothesis was that variation in language indicates social processes, and we tried to map it. We applied various techniques such as enriching the data with information from dictionaries, using a metric, and developing categories independently. The methodical approach consisted of observing a phenomenon in the real world, understanding the social concept behind it, collecting data, applying quantitative methods to represent the phenomenon or measure behavior. We used techniques from machine learning, content analysis, and network analysis to exemplify the dichotomy of structure and content of social media data. For the development of models in complex systems, interdisciplinary exchange was particularly important in this work. Over time, occasional exchange turned into a co-construction process. As described in the introduction, I come from language and literature studies, my approach to text and language is primarily descriptive, explanatory, theory-driven, and focused on communication. Working with computer scientists has expanded my approach to language and text. Structured access to high-scale texts is only possible through computer-assisted methods. My focus has been on the critical examination of common computer-assisted and quantitative methods for text analysis. We reflected on the use of these methods and their applicability from a linguistic and text perspective. According to our questions of how to use computer-based methods, we have outlined one way of making text data measurable by applying different techniques to process information given in text. Depending on the purpose there are different ways to gain knowledge from text. We have exemplified three approaches: binary classification based on dictionaries for the purpose of extracting features, word distribution segmentation with linguistic metrics to test effects, and the use of a conceptual schema

for categorization to validate existing models based on syntax, semantics and pragmatics. As mentioned above the main purpose is to add information to text data. We discuss some thoughts by Capurro and Hjørland (2020) on the use of the term "information": Semiotics, or the linguistic theory of signs, introduced the distinction between syntax, semantics, and pragmatics in the 1930s, prepared mainly by the work of American philosopher Charles S. Peirce (1839-1914). In his book "Signs, Language, and Behavior," American linguist Charles William Morris (1901-1979) presented a three-dimensional semiotics (syntax, semantics, pragmatics), based on Peirce's work. He believed in the informative effect of signs (Morris, 1955). With his theory of semantic information, Dretske (1986) assumes that symbols only "contain" information when they are in a causal relationship with the fact to which they refer. However, what Dretske does not consider is that such statements are not absolute, but theory-dependent. In other words, the meaning of a statement depends on the context that determines it (Zoglauer, 1996). According to MacKay (1969), information refers to anything that enhances our understanding of the external world, leading to a more accurate mental model of reality. This means that the informational content of a statement is descriptive in nature.

6.3 Future Questions

The last section of this paper deals with questions that have remained unanswered so far or with new questions that have arisen. The phenomenon of rumor spreading and moral outrage on social media platforms has become increasingly prevalent in recent years. While some may dismiss these events as mere digital noise, they have real-world consequences and can significantly impact individuals and communities. Regarding our work on hate speech, another direction for research would be to understand the implicit actions of a harasser better. Therefore, it is crucial to comprehend how they convey, stage, and conceal the adversarial context of their attack. It can involve different tactics, such as using irony or sarcasm. In sociology, constructing the context of interactions defines systemic power. The challenge lies in how the harasser uses implicitness to structure and stage the exercise of their power. Clues to the structure of the interaction context may not be present in the harasser's words and sentences but could be found in upstream thread traces. It can help identify the various ways the harasser takes a position in the structure they create. An interlocutory approach could be used

to analyze a tweet and its utterance, using different theories to specify how actors contextualize their statements and stage the "scene" projected by a tweet. It could include examining opposed or common reference groups, status differences, and normative choices. These dimensions vary in appropriateness judgments and can help to differentiate between sarcasm and irony. This approach is based on a symbolic interactionist perspective on knowledge-building and is compatible with Lorraine Code's thinking about epistemic (ir)responsibility (Code, 1987).

Bibliography

L. A. Adamic and N. Glance. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, pages 36–43, 2005. doi: 10.1145/1134271.1134277.

K. Adamzik. *Textlinguistik: Grundlagen, Kontroversen, Perspektiven.* De Gruyter, 2016.

S. Aday, H. Farrell, M. Lynch, J. Sides, J. Kelly, and E. Zuckerman. Blogs and bullets: New media in contentious politics. Technical report, United States Institute of Peace, 2010.

A. Alexiadou, L. Haegeman, and M. Stavrou. *Noun Phrase in the Generative Perspective.* De Gruyter Mouton, 2008.

K. Allan. *Natural Language Semantics.* Blackwell Publishers Ltd, Oxford, 2002.

H. Allcott and M. Gentzkow. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211–236, 2017.

G. W. Allport and L. J. Postman. The psychology of rumor. *Journal of Clinical Psychology*, 3(4):402–402, 1947.

G. W. Allport, K. Clark, and T. Pettigrew. *The nature of prejudice.* Addison-Wesley Publishing Company, Massachusetts, 1954.

O. Araque, J. F. Sanchez-Rada, A. Carrera, C. A. Iglesias, J. Tardio, G. Garcia-Grao, S. Musolino, and F. Antonelli. Making Sense of Language Signals for Monitoring Radicalization. *Applied Sciences*, 12(17), 2022.

J. S. Armstrong and S. Patnaik. Using Quasi-Experimental Data To Develop Empirical Generalizations For Persuasive Advertising. *Journal of Advertising Research*, 49(2):170–175, 2009.

J. L. Austin. *How to do things with words.* Oxford University Press, 1962. The Williams James Lectures by J. L. Austin (edited by J. O. Urmson).

C. A. Bail, T. W. Brown, and M. Mann. Channeling Hearts and Minds: Advocacy Organizations, Cognitive-Emotional Currents, and Public Conversation. *American Sociological Review*, 82(6):1188–1213, 2017.

C. A. Bail, L. Argyle, T. Brown, J. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky. Exposure to opposing views on social media can increase political polarization. *PNAS*, 115(37):9216–9221, 2018a.

C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018b.

E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

P. Barberá. How social media reduces mass political polarization. Evidence from Germany, Spain, and the US. *Job Market Paper, New York University*, 46, 2014.

J. Bartlett, M. King, and J. Birdwell. *The edge of violence*. Demos, London, 2010.

Y. Benkler. *The wealth of networks: How social production transforms markets and freedom*. Yale University Press, 2006.

B. Berelson. *Content Analysis in Communication Research*. Free Press, 1952.

A. Bonvin and A. Lambelet. Algorithmic and subjective measures of lexical diversity in bilingual written corpora: a discussion. *Corela. Cognition, représentation, langage*, 2017.

S. Bordag. A comparison of co-occurrence and similarity measures as simulations of context. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 52–63. Springer, 2008.

S. P. Borgatti, M. G. Everett, and J. C. Johnson. *Analyzing Social Networks*. SAGE Publications, 2018.

R. Borum. Radicalization into Violent Extremism I: A Review of Social Science Theories. *Journal of Strategic Security*, 4(4):7–36, 2011.

A. Bramson, P. Grim, D. J. Singer, S. Fisher, W. Berger, G. Sack, and C. Flocken. Disambiguation of social polarization concepts and measures. *The Journal of Mathematical Sociology*, 40:80–111, 2016.

N. Bubenhofer and J. Scharloth. Kulturwissenschaftliche Orientierung in der Computer- und Korpuslinguistik. In *Sprache - Kultur - Kommunikation / Language - Culture - Communication. Handbuch zu Linguistik als Kulturwissenschaft / Handbook of Linguistics as a Cultural Discipline. Edited by Jäger, Ludwig and Holly, Werner and Krapp, Peter and Weber, Samuel and Heekeren, Simone*. De Gruyter, 2016.

R. E. Burton and R. W. Kebler. The 'half-life' of some scientific and technical literatures. *American Documentation*, 11:18–22, 1960.

H. Bußmann. *Lexikon der Sprachwissenschaft*. Alfred Kröner Verlag, Stuttgart, 2008.

K. G. Bühler. *Theory of Language*. John Benjamins Publishing Company, 2011.

D. T. Campbell. *Methodology and epistemology for social science: Selected papers*. University of Chicago Press, 1988.

R. Capurro and B. Hjørland. The concept of information. *Annual Review of Information Science and Technology, (Ed.) B. Cronin*, 37:343–411, 2020.

K. Carley. Network text analysis: The network position of concepts. In *Ed. Roberts, C.W. A: Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts.*, pages 79–100. Routledge, 2000.

A. Chadwick, J. Dennis, A. P. Smith, J. Dennis, and A. P. Smith. Politics in the Age of Hybrid Media: Power, Systems, and Media Logics. *The Routledge Companion to Social Media and Politics*, 2015.

Y.-T. Chang, H. Yu, and H.-P. Lu. Persuasive messages, popularity cohesion, and message diffusion in social media marketing. *Journal of Business Research*, 68 (4):777–782, 2015.

P.-L. Chen, Y.-C. Cheng, and K. Chen. Analysis of social media data: An introduction to the characteristics and chronological process. *Big Data in Computational Social Science and Humanities*, pages 297–321, 2018. doi: 10.1007/978-3-319-95465-3_16.

J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. How Community Feedback Shapes User Behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 41–50, 2014.

L. Code. *Epistemic Responsibility*. Albany: Published for Brown University Press by University Press of New England, 1987.

K. Cohen, F. Johansson, L. Kaati, and J. C. Mork. Detecting linguistic markers for radical violence in social media. *Terrorism and Political Violence*, 26, 2014.

M. D. Conover, J. Ratkiewicz, M. Francisco, B. Goncalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 89–96, 2011. doi: 10.1609/icwsm.v5i1.14126.

T. D. Cook and D. T. Campbell. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Houghton Mifflin, 1979.

E. W. Corsevski. The physical side of linguistic violence. *Peace Review*, 10(4): 513–516, 1998. doi: 10.1080/10402659808426195.

L. J. Cronbach. Beyond the two disciplines of scientific psychology. *American Psychologist*, 30(2):116–127, 1975.

C. Crossett and J. A. Spitaletta. Radicalization. relevant psychological and sociological concepts. Technical report, Prepared for the U.S. Army Asymmetric Warfare Group, The John Hopkins University, 2010. https://info.publicintelligence.net/USArmy-RadicalizationConcepts.pdf.

D. Crystal. *Language and the Internet*. Cambridge University Press, 2001.

D. Crystal. *The Cambridge Encyclopedia of Language*. Cambridge University Press; 3. Edition, 2010.

J. Culpeper. *Impoliteness: Using Language to Cause Offence*. Cambridge University Press, 2011.

R. M. Dailey, C. M. Lee, , and B. H. Spitzberg. Communicative aggression: Toward a more interactional view of psychological abuse. In B. H. Spitzberg and W. R. Cupach, editors, *The dark side of interpersonal communication*, pages 297–326. Routledge, 2007.

P. Dandekar, A. Goel, and D. T. Lee. Biased assimilation, homophily, and the dynamics of polarization. *PNAS*, 110(15):5791–5796, 2013.

J. Davis. Presidential campaigns and social networks: How Clinton and Trump used Facebook and Twitter during the 2016 election. *Senior Theses. 75*, 2017.

L. de Vries, S. Gensler, and P. S. H. Leeflang. Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing. *Journal of Interactive Marketing*, 26:83–91, 2012.

D. Deavours, W. Heath, K. Miller, M. Viehouser, S. Palacios-Plugge, and R. Broussard. Reciprocal journalism's double-edged sword: How journalists resolve cognitive dissonance after experiencing harassment from audiences on social media. *Journalism*, 2022. doi: 10.1177/14648849221109654.

D. S. DeRue, J. D. Nahrgang, J. R. Hollenbeck, and K. Workman. A quasi-experimental study of after-event reviews and leadership development. *Journal of Applied Psychology*, 97:997–1015, 2012.

P. DiMaggio, J. Evans, and B. Bryson. Have American's Social Attitudes Become More Polarized? *American Journal of Sociology*, 102(3):690–755, 1996.

J. DiNardo. Natural experiments and quasi-natural experiments. In *The New Palgrave Dictionary of Economics*, pages 1–12. Palgrave Macmillan UK, 2016. doi: 10.1057/978-1-349-95121-5_2006-1.

F. Dretske. *Minds, Machines and Meaning.* In: C. Mitcham, A. Hunning Ed.: Philosophy and Technology II, Dordrecht, 1986.

S. Du and S. Gregory. The Echo Chamber Effect in Twitter: does community polarization increase? In *International Workshop on Complex Networks & Their Applications*, page 373–378, 2016.

N. Ellison and D. M. Boyd. Sociality through social network sites. *The Oxford Handbook of Internet Studies, W.H. Dutton (Ed.)*, pages 151–172, 2013.

M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding. Peer to peer hate: Hate speech instigators and their targets. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), 2018. doi: https://doi.org/10.1609/icwsm.v12i1.15038.

J. Esteban and D. Ray. On the measurement of polarization. *Econometrica*, 62 (4):819–51, 1994.

L. Festinger. *A Theory of Cognitive Dissonance*. Stanford University Press, 1957.

M. P. Fiorina, S. J. Abrams, and J. C. Pope. *Culture war? The myth of a polarized America*. New York: Pearson Longman, 2010.

S. Fishman. Community-Level Indicators of Radicalization: A Data and Methods Task Force. Technical report, Report to Human Factors/Behavioral Sciences Division, Science and Technology Directorate, US. Department of Homeland Security, College Park, MD., 2010.

S. Frenda, V. Patti, and P. Rosso. Killing me softly: Creative and cognitive aspects of implicitness in abusive language online. *Natural Language Engineering*, pages 1–22, 2022. doi: doi:10.1017/S1351324922000316.

I. Gagliardone, D. Gal, T. Alves, and G. Martinez. *Countering Online Hate Speech*. UNESCO Publishing, 2015.

K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*, pages 913–922, 2018.

R. K. Garrett, D. Carnahan, and E. K. Lynch. A turn toward avoidance? selective exposure to online political information, 2004–2008. *Political Behavior*, 35(1):113–134, 2013.

H. A. Gaspar, C. Daase, N. Deitelhoff, J. Junk, and M. Sold. *Radicalization and Political Violence – Challenges of Conceptualizing and Researching Origins, Processes and Politics of Illiberal Beliefs*, volume 14. Peace Research Institute Frankfurt (PRIF), Frankfurt, 2020.

R. Ghawi and J. Pfeffer. Extraction Patterns to Derive Social Networks from Linked Open Data Using SPARQL. *Information*, 11(7), 2020. doi: 10.3390/info11070361.

J. L. Gillin and J. P. Gillin. *Cultural Sociology*. Macmillan, 1950.

A. L. Gonzales, J. T. Hancock, and J. W. Pennebaker. Language Style Matching as a Predictor of Social Dynamics in Small Groups. *Communication Research*, 37(1):3–19, 2010.

S. M. Goodreau. Advances in exponential random graph (p*) models applied to a large social network. *Social Networks*, 2(29):231–48, 2007.

M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78 (6):1360–1680, 1973.

B. Gribbons and J. Herman. True and Quasi-Experimental Designs. *University of Massachusetts Amherst*, 5(14), 1996. doi: 10.7275/fs4z-nb61.

J. Groshek and K. Koc-Michalska. Helping populism win? Social media use, filter bubbles, and support for populist presidential candidates in the 2016 US election campaign. *Information, Communication & Society*, 20(9):1389–1407, 2017.

A. Gruzd and J. Roy. Investigating political polarization on twitter: A canadian perspective. *Policy & Internet*, 6(1):28–45, 2014.

F. Hamborg and K. Donnay. NewsMTSC: A dataset for (multi-)target-dependent sentiment classification in political news articles. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1663–1675, 2021.

B. D. Harris, C. V. Morgan, and B. G. Gibbs. Evidence of political moderation over time: Utah's immigration debate online. *New Media & Society*, 16(8): 1309–1331, 2014.

F. Heider. Attitudes and cognitive organizations. *Journal of Psychology*, 17: 107–112, 1946.

M. Hennig, U. Brandes, J. Pfeffer, and I. Mergel. *Studying Social Networks: A Guide to Empirical Research*. Frankfurt: Campus Verlag, 2012.

M. Hohmann, K. Devriendt, and M. Coscia. Quantifying ideological polarization on a network using generalized Euclidean distance. *Science Advances*, 9(9): eabq2044, 2023. doi: 10.1126/sciadv.abq2044.

S. Hong. Who benefits from Twitter? Social media and political competition in the US House of Representatives. *Government Information Quarterly*, 30(4): 464–472, 2013.

S. Hong and S. H. Kim. Political polarization on twitter: Implications for the use of social media in digital governments. *Government Information Quarterly*, 33(4):777–782, 2016.

M. Horta Ribeiro, S. Jhaver, S. Zannettou, J. Blackburn, G. Stringhini, E. De Cristofaro, and R. West. Do Platform Migrations Compromise Content Moderation? Evidence from r\The_Donald and r\Incels. In *Proceedings of the ACM on Human-Computer Interaction*, volume 5, 2021.

P. N. Howard, A. Duffy, D. Freelon, M. M. Hussain, W. Mari, and M. Maziad. Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring? Technical report, Social Science Research Network, 2011.

S. Iyengar and K. S. Hahn. Red media, blue media: Evidence of ideological selectivity in media use. *Journal of communication*, 59(1):19–39, 2009.

S. Jarvis. Capturing the Diversity in Lexical Diversity. *Language Learning*, 63 (s1), 2013.

F. Karim. Social media use and its connection to mental health: A systematic review. *Cureus*, 12(6), 2020. doi: 10.7759/cureus.8627.

T. Kleiner. Public opinion polarisation and protest behaviour. *European Journal of Political Research*, 57(4):941–962, 2018.

K. Krippendorff. *Content Analysis. An Introduction to Its Methodology*. Sage Publications, 2013.

P. R. Kroeger. *Analyzing Grammar: An Introduction*. Cambridge University Press, 2005. doi: 10.1017/CBO9780511801679.

E. Kubin and C. von Sikorski. The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, 45(3):188–206, 2021.

U. Kuckartz. Qualitative Text Analysis: A Systematic Approach. In G. Kaiser and N. Presmeg, editors, *Compendium for Early Career Researchers in Mathematics Education*, pages 181–197. Springer International Publishing, 2019.

S.-M. Kühnel and D. Krebs. *Statistik für die Sozialwissenschaften. Grundlagen, Methoden, Anwendungen.* 7. Aufl. Reinbek, Hamburg: Rowohlt, 2014.

G. Labianca, G. Kane, Alavi, and S. Borgatti. What's Different about Social Media Networks? A Framework and Research Agenda. *MIS Quarterly*, 38, 2013.

H. D. Lasswell. The structure and function of communication in society. *New York: Harper and Row*, 1948. L. Bryson (Ed.) The communication of ideas, pp. 37-51.

E. Lawrence, J. Sides, and H. Farrell. Self-segregation or deliberation? blog readership, participation, and polarization in american politics. *Perspectives on Politics*, 8(1):141–157, 2010.

D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. V. Alstyne. Computational Social Science. *Science*, 323(5915): 721–723, 2009.

D. Lazer, E. Hargittai, D. Freelon, S. Gonzalez-Bailon, K. Munger, K. Ognyanova, and J. Radford. Meaningful measures of human society in the twenty-first century. *Nature*, 595(7866):189–196, July 2021.

D. M. J. Lazer, A. Pentland, D. J. Watts, S. Aral, S. Athey, N. Contractor, D. Freelon, S. Gonzalez-Bailon, G. King, H. Margetts, A. Nelson, M. J. Salganik, M. Strohmaier, A. Vespignani, and C. Wagner. Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062, 2020.

J. Lehmann, M. Mittelbach, and S. Schmeier. Quantifizierung von Emotionswörtern in Texten. Technical report, DARIAH-DE Working Papers. 24, 2017.

L. Lemnitzer and H. Zinsmeister. *Korpuslinguistik. Eine Einführung.* Tübingen: Narr, 2015.

Z. C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

D. M. MacKay. *Information, Mechanism and Meaning.* MIT, 1969.

N. Marres. *Digital sociology: the reinvention of social research*. Polity, 2017.

L. Mason. "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science*, 59(1): 128–145, 2015.

P. Mayring. Qualitative Inhaltsanalyse – Abgrenzungen, Spielarten, Weiterentwicklungen. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 20(3), 2019.

C. McCauley and S. Moskalenko. Mechanisms of political radicalization: Pathways toward terrorism. *Terrorism and Political Violence*, 20(3):415–433, 2008.

I. McCulloh and K. M. Carley. Social network change detection. Technical report, Center for Computational Analysis of Social and Organizational Systems, CASOS Technical Report, 2008.

I. McCulloh, J. Lospinoso, and K. M. Carley. Social network probability mechanics. In *Proceedings of the World Scientific Engineering Academy and Society 12th International Conference on Applied Mathematics*, 2007.

M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27:415–444, 2001.

A. Medvedev, R. Lambiotte, and J.-C. Delvenne. The Anatomy of Reddit: An Overview of Academic Research. *Springer Proceedings of Complexity*, pages 183–204, 2019. doi: 10.1007/978-3-030-14683-2_9.

Y. Mehmood and V. Balakrishnan. An enhanced lexicon-based approach for sentiment analysis: a case study on illegal immigration. *Online Information Review*, 44(5):1097–1117, 2020.

S. Milgram. *The individual in a social world: Essays and experiments*. London: Pinter & Martin, 1977.

F. M. Moghaddam. The Staircase to Terrorism: A Psychological Exploration. *American Psychologist*, 60(2):161–169, 2005.

G. A. Morgan, J. A. Gliner, and R. J. Harmon. Quasi-experimental designs. *Journal of the American Academy of Child and Adolescent Psychiatry*, 39(6): 794–796, 2000.

C. W. Morris. *Signs, Language and Behavior*. New York 1946, 1955.

D. Morrow. *Terrorism and the Escalation of Violence*. The Palgrave Handbook of Mimetic Theory and Religion, edited by Allison, James and Palaver, Wolfgang, London, 2017.

D. C. Mutz and J. J. Mondak. The workplace as a context for cross-cutting political discourse. *The Journal of Politics*, 68(1):140–155, 2006.

P. R. Neumann. The trouble with radicalization. *International Affairs*, 89(4): 873–893, 2013.

B. O'Dea, T. W. Boonstra, M. E. Larsen, T. Nguyen, S. Venkatesh, and H. Christensen. The relationship between linguistic expression in blog content and symptoms of depression, anxiety, and suicidal thoughts: A longitudinal study. *PloS One*, 16(5):e0251787, 2021.

E. Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

A. Park and M. Conway. Longitudinal Changes in Psychological States in Online Health Community Members: Understanding the Long-Term Effects of Participating in an Online Depression Community. *Journal of Medical Internet Research*, 19(3):e71, 2017. doi: 10.2196/jmir.6826.

J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn. The Development and Psychometric Properties of LIWC2015. Technical report, The University of Texas at Austin, 2015.

J. Pfeffer, T. Zorbach, and K. M. Carley. Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. *Journal of Marketing Communications*, 20(1–2):117–128, 2013.

J. Pfeffer, D. Matter, K. Jaidka, O. Varol, A. Mashhadi, J. Lasser, D. Assenmacher, S. Wu, D. Yang, C. Brantner, D. M. Romero, J. Otterbacher, C. Schwemmer, K. Joseph, D. Garcia, and F. Morstatter. Just Another Day on Twitter: A Complete 24 Hours of Twitter Data. In *Proceedings of the 17th International AAAI Conference on Web and Social Media*, 2023a.

J. Pfeffer, A. Mooseder, J. Lasser, L. Hammer, O. Stritzel, and D. Garcia. This sample seems to be good enough! Assessing Coverage and Temporal Reliability

of Twitter's Academic API. In *Proceedings of the 17th International AAAI Conference on Web and Social Media*, 2023b.

C. A. Platt, A. N. Raile, and A. Burnett. Strategically mean: Extending the study of relational aggression in communication. *Annals of the International Communication Association*, 40(1):151–172, 2016.

D. D. Porta and G. LaFree. Guest Editorial: Processes of Radicalization and De-Radicalization. *International Journal of Conflict and Violence*, 6:4–10, 2012.

J. Radford and K. Joseph. Theory in, theory out: The uses of social theory in machine learning for social science. *Frontiers in Big Data*, 2020.

C. W. Roberts. Conceptual framework for quantitative text analysis. *Quality & Quantity 34*, page 259–274, 2000.

E. M. Rogers. *Diffusion of Innovations*. 4th ed. New York: Free Press, 1995.

P. H. Rossi, M. W. Lipsey, and H. E. Freeman. *Evaluation: A Systematic Approach*. Sage Publications, 2004.

D. Ruths and J. Pfeffer. Social media for large studies of behavior. *Science*, 346 (6213):1063–1064, 2014. Publisher: American Association for the Advancement of Science.

J. Scharloth. Hassrede und Invektivität als Gegenstand der Sprachwissenschaft und Sprachphilosophie: Bausteine zu einer Theorie des Metainvektiven. *Aptum 2*, 2017.

T. C. Schelling. Hockey helmets, concealed weapons, and daylight saving: A study of binary choices with externalities. *Journal of Conflict Resolution*, 17 (3):381–428, 1973.

A. Schmid. Radicalisation, De-Radicalisation, Counter-Radicalisation: A Conceptual Discussion and Literature Review. *The International Centre for Counter-Terrorism – The Hague*, 4(2):105, 2013.

J. R. Searle. *Speech Acts*. Cambridge University Press, 1969.

F. Shi, M. Teplitskiy, E. Duede, and J. Evans. The wisdom of polarized crowds. *Nature Human Behaviour*, 3(4):329–336, 2019.

G. Simmel. *Conflict: The Web of Group-affiliations*. Free Press, 1955.

H. Simon. Theories of bounded rationality. *Decision and Organization, edited by C. B. McGuire, and R. Radner*, page 161–176, 1972.

L. G. E. Smith, L. Wakeford, T. Cribbin, J. Barnett, and W. K. Hou. Detecting psychological change through mobilizing interactions and changes in extremist linguistic style. *Computers in Human Behavior*, 108, 2020.

T. A. Snijders, C. E. Steglich, M. Schweinberger, and M. Huisman. Manual for siena version 3.1. university of groningen. Technical report, University of Groningen: ICS / Department of Sociology; University of Oxford: Department of Statistics, 2007.

H. R. Stevens, I. Acic, and S. Rhea. Natural Language Processing Insight into LGBTQ+ Youth Mental Health During the COVID-19 Pandemic: Longitudinal Content Analysis of Anxiety-Provoking Topics and Trends in Emotion in LGBTeens Microcommunity Subreddit. *JMIR public health and surveillance*, 7(8), 2021.

W. Strathern and J. Pfeffer. Negative Dynamics on Social Media and their Ethical Challenges for AI. Technical report, Technical University of Munich, Munich Center for Technology and Society, Institute for Ethics in Artificial Intelligence, 2020.

W. Strathern and J. Pfeffer. Identifying different layers of online misogyny. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*, 2023.

W. Strathern, M. Issig, K. Mozygemba, and J. Pfeffer. QualiAnon – The Qualiservice tool for anonymizing text data. Technical report, TUM-I2087, Technical University of Munich, Department of Informatics, 2020a.

W. Strathern, M. Schönfeld, R. Ghawi, and J. Pfeffer. Against the others! detecting moral outrage in social media networks. *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 322–326, 2020b. doi: 10.1109/ASONAM49781.2020.9381415.

W. Strathern, R. Ghawi, and J. Pfeffer. Advanced Statistical Analysis of Large-Scale Web-based Data. In P. Nymand-Andersen, editor, *Data Science in Economics and Finance for Decision Makers*, pages 43–72. Risk Books, 2021.

W. Strathern, A. Mooseder, and J. Pfeffer. The polarizing impact of continuous presence on users' behavior. In *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media*, 2022a. doi: 10.36190/2022.52.

W. Strathern, M. Schönfeld, R. Ghawi, and J. Pfeffer. Identifying lexical change in negative word-of-mouth on social media. *Social Network Analysis and Mining*, 59(12), 2022b. doi: 10.1007/s13278-022-00881-0.

N. J. Stroud. Polarization and Partisan Selective Exposure. *Journal of Communication*, 60(3):556–576, 2010.

C. R. Sunstein. The law of group polarization. *Journal of political philosophy*, 10(2):175–195, 2002.

C. R. Sunstein. Republic.com 2.0. *Princeton, NJ: Princeton University*, 2007.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, 2011.

Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 2009.

M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

C. Tilly. *From Mobilization to Revolution*. Random House, 1987.

M. Trunfio and S. Rossi. Conceptualising and measuring social media engagement: A systematic literature review. *Italian Journal of Marketing*, 2021(3): 267–292, 2021.

Y. Tsfati and A. Chotiner. Testing the Selective Exposure–Polarization Hypothesis in Israel Using Three Indicators of Ideological News Exposure and Testing for Mediating Mechanisms. *International Journal of Public Opinion Research*, 28(1):1–24, 2016.

A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Proceedings*

of the International AAAI Conference on Web and Social Media, 4(1):178–185, 2010.

S. Udupa. Artificial intelligence and the cultural problem of extreme speech. *Social Science Research Council, (20 December 2020), online*, 2020.

J. van Dijck and T. Poell. Social Media and the Transformation of Public Space. *Social Media + Society*, 1(2):1–5, 2015.

A. L. Vangelisti and A. D. Hampel. Hurtful communication: Current research and future directions. In *In S. W. Smith and S. R. Wilson (Eds.), New directions in interpersonal research*, pages 221–241. Thousand Oaks, CA: Sage, 2010.

L. Vidino. Jihadist Radicalization in Switzerland. Technical report, Center for Security Studies (CSS), ETH Zürich, 2013. URL https://www.files.ethz.ch/isn/172401/CH_radicalization_report.pdf.

C. Wagner, P. Singer, F. Karimi, J. Pfeffer, and M. Strohmaier. Sampling from social networks with attributes. In *Proceedings of the 26th International Conference on World Wide Web*, page 1181–1190, 2017. doi: 10.1145/3038912.3052665.

C. Wagner, M. Strohmaier, A. Olteanu, E. Kıcıman, N. Contractor, and T. Eliassi-Rad. Measuring algorithmically infused societies. *Nature*, 595(7866): 197–204, 2021.

M. Wich, M. Breitinger, W. Strathern, M. Naimarevic, G. Groh, and J. Pfeffer. Are Your Friends Also Haters? Identification of Hater Networks on Social Media: Data Paper. In *Companion Proceedings of the Web Conference 2021*, page 481–485, 2021. doi: 10.1145/3442442.3452310.

J. P. Williams and S. X. X. Ho. "Sasaengpaen" or K-pop fan? Singapore youths, authentic identities, and Asian media fandom. *Deviant Behavior*, 37(1):81–94, 2016.

G. K. Zipf. Human behavior and the principle of least effort. *Journal of Clinical Psychology*, 6(3), 1949.

T. Zoglauer. *Can Information be naturalized?* In: Kornwachs, K., Jacoby, K. Eds.: Information. New Questions to a Multidisciplinary Concept, 1996.