Machine Learning for the Automated Content Analysis of Incivility in Online Discussions

Inaugural dissertation
to obtain a doctoral degree in philosophy (Dr. phil.)
at the Faculty of Arts and Humanities of
Heinrich Heine University Düsseldorf

submitted by

Anke Stoll

First supervisor:
Prof. Dr. Marc Ziegele
Heinrich Heine University Düsseldorf

Second supervisor:
Prof. Dr. Stefan Conrad
Heinrich Heine University Düsseldorf

Düsseldorf, July 2023

D61

Mein Dank von Herzen gilt meinem Betreuer und Unterstützer Marc Ziegele sowie Marike Bormann, Katharina Frehmann, Dominique Heinbach und Lena Wilms, meinen Kolleginnen, Weggefährtinnen und Freundinnen.

This document contains the synopsis of my cumulative dissertation, which was submitted in a slightly different version at the Faculty of Arts and Humanities of Heinrich Heine University Düsseldorf in July 2023. In addition to the synopsis, the dissertation consists of the following seven research articles:

Publication [1] Risch, J., **Stoll, A.**, Ziegele, M., & Krestel, R. (2019). hpiDEDIS at GermEval 2019: Offensive Language Identification using a German BERT model. In S. Evert (Ed.), *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)* (pp. 403-408). German Society for Computational Linguistics & Language Technology.

Publication [2] Risch, J., **Stoll, A.**, Wilms, L., & Wiegand, M. (2021). Overview of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments. In J. Risch, A. Stoll, L. Wilms, & M. Wiegand (Eds.), *Proceedings of the GermEval 2021 SharedTask on the Identification of Toxic, Engaging, and Fact-Claiming Comments* (pp. 1-12). Association for Computational Linguistics.

Publication [3] Küchler, C., **Stoll, A.**, Ziegele, M., & Naab, T. (2022). Gender-related Differences in Online Comment Sections: Findings from a Large-Scale Content Analysis of Commenting Behavior. *Social Science Computer Review, 41*(3), 728–747. https://doi.org/10.1177/0894439321105204

Publication [4] **Stoll, A.**, Ziegele, M., & Quiring, O. (2020). Detecting Impoliteness and Incivility in Online Discussions: Classification Approaches for German User Comments. *Computational Communication Research*, *2*(1), 109-134. https://doi.org/10.5117/CCR2020.1.005.KATH

Publication [5] **Stoll, A.**, Wilms, L., & Ziegele, M. (2023). Developing an Incivility-Dictionary for German Online Discussions - A Semi-Automated Approach Combining Human and Artificial Knowledge. *Communication Methods and Measures, 17*(2), 131-149. https://doi.org/10.1080/19312458.2023.2166028

Publication [6] **Stoll, A.** (2023). The Accuracy Trap or How to Build a Phony Classifier. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech analysis* (pp. 371-381). Digital Communication Research. https://doi.org/10.48541/dcr.v12.22

Publication [7] **Stoll, A.** (2020). Supervised Machine Learning mit Nutzergenerierten Inhalten: Oversampling für nicht balancierte Trainingsdaten [Supervised Machine Learning with User-Generated Content: Oversampling for imbalanced training data.]. *Publizistik*, *65*(2), 233-251. https://doi.org/10.1007/s11616-020-00573-9

Abstract

The prevalence of *incivility* in online discussions concerns platform operators, community managers, as well as scholars in communication science and *machine learning* research. Recent progress of *artificial intelligence* raises hopes that algorithm-based moderation systems can assist the often exhausting detection and moderation of uncivil content using machine learning. Also in communication science, the need for automated, *computational methods* to analyze the growing amounts of online discussion data is becoming more and more demanding. This thesis addresses the question: *How can incivility be measured using machine learning methods?* To this end, recent methodological developments in machine learning are connected with communication science research on incivility. The seven research articles of the cumulus contribute to three further subordinate aspects of this overarching question. Articles [1] and [2] address the question: *To what extent can incivility be measured using machine learning methods*? Both publications contribute insights into and an overview of current approaches to incivility classification using machine learning, including a benchmark data set and a shared task on the identification of incivility. The second subordinated question asks: *How can machine learning extend and elaborate communication science research on incivility?* Articles [3], [4], and [5] present use cases, best practice studies, and methodological work that show how machine learning can enhance communication science research on incivility at different points of the research process. The third subordinated question asks: *What methodological challenges do communication scholars face using machine learning-based incivility classification and how can they be overcome?* To address this question, articles [6] and [7] discuss specific challenges of measuring incivility with machine learning methods and provide methodological guidance and concrete approaches to address them. Overall, the research program of this dissertation contributes significant work to the interdisciplinary task of incivility detection at the interface of communication science and machine learning research. This work includes studies that apply and discuss several approaches to incivility classification in different research settings, under different conditions, and with different objectives and outcomes. The results of this dissertation provide valuable insights for computational communication scientists that aim to apply, evaluate, and further develop machine learning approaches to incivility and comparable research subjects. Finally, findings offer practical implications for the application of algorithm-based moderation systems by pointing out potentials and weaknesses of machine learning approaches to incivility detection in online discussions.

Table of Contents

1 Introduction

1.1 Relevance and Research Interest

The prevalence of *incivility* in online discussions has evolved into a serious issue for democratic societies. With the increase of participation in user comment sections and social media, incivility has become an inherent part of societal and academic discourse. In communication science, incivility is an established research subject that has been investigated from various angles, including its prevalence (e.g., Coe et al., 2014; Santana, 2014; Su et al., 2018), perception (e.g., Boberg et al., 2018; Bormann, 2022; Bormann et al., 2023; Chen & Pain, 2017; Harlow, 2015), and effects (e.g., Anderson et al., 2014; Hsueh et al., 2015; Hwang et al., 2014; Prochazka et al., 2018; Springer et al., 2015; Stroud et al., 2016; Wang, 2020; Ziegele, Jost et al., 2018). Among incivility researchers, there exist multiple views on what exactly incivility is (Bormann et al., 2022). However, definitions share the common ground that incivility is a form of norm violation that comes with a significant amount of subjectiveness, shaped by situational context and individual perception (Coe et al., 2014; Chen, 2017; Chen et al., 2019; Herbst, 2010). Moreover, empirical studies provide evidence that different forms of incivility are perceived as differently severe (e.g., Muddiman, 2017; Muddiman, 2019; Kalch & Naab, 2017) and that not all forms of incivility are equally straightforward to identify (e.g., Bormann et al., 2023; Muddiman & Stroud, 2017).

Many scholars agree that a crucial component in addressing uncivil discussions online is moderation (e.g., Friess et al., 2021; Heinbach & Wilms, 2022; Wright, 2006; Ziegele & Jost, 2020). For online platforms and news websites in Germany, handling and countering incivility is also a legal issue (BMJ, 2022; European Commission, 2023; Kümpel & Rieger, 2019). Here, moderators and community managers play an important role by reviewing and, if necessary, removing uncivil content. Recently, the challenging task of online moderation is increasingly supported by *algorithm-based moderation systems* (Gorwa et al., 2020; Vox Media, n.d; Wilms et al., 2023). Encouraged by the fast and steady progress in the research fields of *artificial intelligence* and *machine learning*, hopes are raised that powerful algorithms can assist moderators and community managers by automatically detecting and countering uncivil comments (Beuting, 2021; Wilms et al., 2023). As a result of this development, the automated detection of incivility and related concepts including *hate speech*, *toxicity,* and *offensive language* has grown into an important issue in machine learning research (e.g., Burnap & Williams, 2015;

Davidson et al., 2017; Wiegand et al., 2018; Zampieri et al., 2020). Here, incivility detection is usually approached as a *supervised learning* task. Essentially, this means that an algorithm learns to identify uncivil content based on - oftentimes human - decisions (Bishop, 2006). At this point, the question of how incivility is to be defined and identified is moving forward into the research fields of machine learning and artificial intelligence.

In recent years, communication science has also become more interested in automated, *computational methods* for analyzing the growing amounts of digitally available data (Domahidi et al., 2019; Lazer et al., 2009; Scharkow, 2012; van Atteveldt et al., 2019), including incivility in online discussions (e.g., Ksiazek et al., 2015; Muddiman & Stroud, 2017; Sadeque et al., 2019; Theocharis et al., 2020; Ziegele, Daxenberger et al., 2018). To this end, existing machine learning models and approaches usually have to be adjusted and further developed by researchers to fit the specific research subject, which demands increasingly specific and comprehensive know-how. In communication science, however, machine learning methods are establishing themselves rather slowly alongside simpler approaches to automated content analysis (Boumans & Trilling, 2018; Hase et al., 2022; van Atteveldt & Peng, 2018). Thus, their application and evaluation still come with a degree of uncertainty (Niemann-Lenz et al., 2019; Scheper & Kathirgamalingam, 2022).

Nonetheless, the rapid progress in machine learning and artificial intelligence holds great potential for several areas of online communication research, including incivility in online discussions. The application of machine learning methods, however, requires specific knowledge, not only regarding its implementation but also to accomplish a holistic evaluation regarding its usefulness and limits. The automated measurement of incivility in particular requires a nuanced and comprehensive approach as its subjective nature already challenges practitioners and manual content analysis methods. Methodological research at the interface to computer science could initiate a beneficial exchange between machine learning research and different areas of communication science. In this sense, the research program of this dissertations aims to connect current advances in machine learning with communication science research on incivility, with the goal of an interdisciplinary, methodological-focused transfer. The overarching research question of this dissertation is therefore:

How can incivility be measured using machine learning methods?

1.2 Research Aim and Contributions of this Thesis

In this cumulative dissertation, I ask the overall research question of *how incivility can be measured using machine learning methods*. To address this question, I aim to accomplish an essential step to connect the research fields of computer science-driven machine learning and artificial intelligence with communication science research on incivility. Therefore, the research program of this thesis covers the multifaceted steps of an interdisciplinary transfer with a methodological, communication science focus. From the overarching research question, I derive the three following subordinated research questions that further structure the contributions of the research program:

1) *To what extent can incivility be measured using machine learning methods?* The subjectivity of incivility due to individual and contextual factors makes it a challenging concept to measure with standardized, text-based methods. Nonetheless, the demand for automated methods to analyze growing amounts of digital data increases, both for communication science and machine learning research and for platforms of online discussions. But how close do current machine learning approaches come to detecting uncivil content? To answer this question, articles [1] and [2] provide an overview of the performances and limits of current machine learning approaches to incivility detection for German-language online discussions.

2) *How can machine learning methods extend and elaborate communication science research on incivility?* To address this question, research articles [3], [4], and [5] present how machine learning can enhance communication science research at different points of the research process, including research designs, research questions, and analysis methods. These enhancements include a) the enlargement of samples and thus, the enabling of more demanding analysis methods, b) the examination and review of theoretical concepts of incivility from a data-driven perspective, and c) the improvement of established, simpler computational methods, namely dictionaries, in terms of saving costs and improving measurement.

3) *What methodological challenges do communication scholars face with machine learning-based incivility classification and how can they be overcome?* To answer this question, articles [6] and [7] elaborate on specific pitfalls that occur at the interface of machine learning and incivility research and provide methodological insights and guidance on how to address them, taking a communication science perspective.

The compilation of the research program covers the comprehensive and diverse steps of a methodological-focused transfer from machine learning research in computer sciences to communication science research on incivility. Led by the three subordinate research questions presented, this cumulative dissertation contributes to the three following core aspects: 1) The interdisciplinary body of research on incivility detection for German-language online discussions. The ongoing adaptation and further development of state-of-the-art approaches to specific research subjects and questions are ongoing endeavors and crucial objectives of machine learning research. In addition, this research demonstrates how a communication science perspective can be beneficial to the task of machine learning-based incivility classification (articles [1] and [2]). 2) The development, application, and deployment of novel approaches and specific instruments for the automated analysis of incivility, primarily tailored to communication science scholars and research questions. This work includes distinct tools, case studies, as well as different machine learning-based approaches to the automated analysis of incivility that can also be adapted for related research questions of online communication (articles [3], [4], and [5]). 3) The methodological transfer of machine learning methods into communication science research on incivility with the goal of a holistic understanding and critical use. Respective work focuses on how pitfalls can be avoided and how in-depth evaluation can not only enable a proper application of machine learning methods, but further can provide additional insights into the research subject (articles [6] and [7]).

1.3 Structure of Synopsis

In this synopsis, the seven research articles of the cumulus are presented, summarized, and systematically connected regarding their topics and specific contributions. After an introduction and overview of incivility research and machine learning approaches to incivility detection in chapters 2 and 3, the main contributions of the articles will be discussed in chapter 4. Even though the single publications include more than one contribution, they will be discussed regarding their main contribution to one of the three subordinated research questions. Chapter 5 provides a comprehensive discussion of the contributions, limitations, and implications of the research program. Finally, the synopsis closes with a conclusion in chapter 6. In the following, the structure of the synopsis is described in detail.

- *Incivility in Online Discussions and Deliberative Discourse* (chapter 2): This chapter focuses on the theoretical and conceptual background of incivility and gives an overview of the current state of research on incivility in online discussions. The chapter will further elaborate on the challenges of measuring incivility with standardized content analysis methods.

- *Machine Learning - Expertise from Computer Science* (chapter 3): This chapter gives an overview of different machine learning methods that have also been used in this dissertation with a focus on incivility detection. Moreover, the chapter discusses certain aspects and characteristics of machine learning that can challenge the automated detection of incivility.

- *Contributions of this Thesis* (chapter 4): This chapter summarizes and systematizes the single articles of the cumulus in terms of their main contributions to this thesis regarding the following aspects:

    - *State of Research - Incivility Detection Using Machine Learning:* This chapter addresses the question *to what extent incivility can be measured using machine learning methods* and shows, how articles [1] and [2] contribute to the current state of research on the detection of incivility in German-language online discussions.

    - *Machine Learning for Incivility and Communication Science Research:* This chapter addresses the question of *how machine learning can extend and elaborate communication science research on incivility* at different points of the research process, including research designs, research questions, and research methods and shows, how articles [3], [4], and [5] contribute to this question.

    - *Change of Perspective - Methodological Considerations:* This chapter addresses the question *what methodological challenges communication scholars face with machine learning-based incivility classification and how they can be overcome.* The contributions of research articles [6] and [7] are presented in this regard.

- *Discussion* and *Conclusion* (chapters 5 and 6): These final chapters critically discuss the findings of this dissertation and offer implications for communication science and machine learning research as well as for practitioners that employ automated incivility detection to support moderation.

2 Incivility in Online Discussions and Deliberative Discourse

Since online discussions, comment sections, and social media became an integral part of everyday life, political and societal discourse has been shaped by the concern of incivility (e.g., Boatright, 2019). In communication science, incivility is a well-established research subject and a great body of empirical work has dealt with different aspects of uncivil communication. This chapter provides an overview of the state of research on incivility as a concept and discusses, how contextual and individual factors can affect the measurement of incivility using standardized, text-based methods.

2.1 What is Incivility? Concepts and Definitions

Over recent years, online discussions have evolved into an important part of political discourse (Bergström & Wadbring, 2015; Grimmelmann, 2015; Heinbach, in press; Kümpel & Rieger, 2019; Walther & Jang, 2012). Along with the increasing engagement in comment sections, online forums, or participation platforms (Coleman & Shane, 2011; Esau et al., 2021; Frieß & Porten-Chée, 2018; Kersting, 2019), scientists as well as practitioners and politicians have raised concerns about the often low quality of contributions, frequently accompanied by a rude tone and other forms of deviant communication (Anderson et al., 2014; Coe et al., 2014; Friess et al., 2021; Su et al., 2018; Ziegele & Jost, 2020). Various empirical studies have suggested that such *uncivil* communication can have several negative effects on multiple levels of the public discourse, from individual participants to entire political and societal processes (e.g., Anderson et al., 2014; Hsueh et al., 2015; Hwang et al., 2014; Mutz & Reeves, 2005; Prochazka et al., 2018). Yet, opinions differ on how uncivil communication should defined and operationalized. Scholars agree, however, that incivility is a form of norm violation that is both highly context-dependent and subjective (Coe et al., 2014; Herbst, 2010; Sydnor, 2018).

A great body of research approaches incivility in political online discussions from the perspective of theories of democracy, where incivility is defined as a violation of democratic or deliberative norms (e.g., Andersson et al., 2014; Coe et al., 2014; Kalch & Naab, 2017; Papacharissi, 2004; Rowe, 2015; Ziegele, Jost et al., 2018). The framework of deliberation suggests that in a public sphere, citizens should discuss issues of social and political relevance in a rational, reciprocal, and respectful way (Dahlberg, 2001; Fraser, 1990; Habermas, 1996; Ruiz et al., 2011; Young, 1996). In this context, incivility is often understood as disrespectful communication or behavior towards participants or third parties that undermines a free exchange of ideas and (conflicting) opinions (Brooks & Geer, 2007; Kalch & Naab, 2017; Hwang et al.,

2018; Rowe, 2015). According to this understanding, incivility includes negative stereotyping and the undermining or threat of democratic or individual rights (Papacharissi, 2004; Rowe, 2015). A further line of research conceptualizes incivility as a violation of interpersonal norms of politeness (e.g., Brown & Levinson, 1987; Chen & Lu, 2017; Mutz, 2007; Mutz & Reeves, 2005), which includes communication that threatens the face of discussion participants and that would result in negative emotions and reactions. This concept of incivility is operationalized as, for example, name calling and the use of vulgar or pejorative language (e.g., Chen & Lu, 2017; Chen & Ng, 2017). In her pioneering work, Papacharissi (2004) argues that severe incivility, which undermines democratic norms by denying "people their personal freedoms, and stereotyp[ing] social groups" (p. 267) must be distinguished from impoliteness, which she considers as rather harmless or even useful for a heated discussion (Papacharissi, 2004). In the past years, numerous scholars have built on this conceptualization of incivility in their work (e.g., Kalch & Naab, 2017; Muddiman, 2017; Naab et al., 2018; Oz et al., 2018; Rowe, 2015). However, current research also defines incivility as a violation of multiple norms (e.g., Bormann et al., 2022; Muddiman, 2017; Stryker et al., 2016). A recent framework by Bormann and colleagues (2022) takes a novel path and conceptualizes incivility as a violation of cooperative norms, meaning forms of communication that obstruct cooperation between actors. Established norms of deliberation and respect merge in this concept and are enhanced by further norms of communication, such as comprehensive and responsive communication and providing sufficient information.

While incivility constitutes a popular - albeit heterogeneously defined - concept in communication science, additional related concepts play a role in the scientific and societal discourse about deviant communication. These concepts include both specific sub-concepts of incivility and synonymous notions that are established in different discourses or research fields (Heinbach, in press). In computer science and computational linguistics, frequently applied concepts are offensive language (e.g., Kumaresan et al., 2021; Razavi et al., 2010; Risch et al., 2018; Xiang et al., 2012; Zampieri et al., 2019; Zampieri et al., 2020) and toxicity (e.g., Georgakopoulos et al., 2018; Risch, 2023; Risch & Krestel, 2020; van Aken et al., 2018). In their benchmarking shared task for offensive language classification of Tweets, Zampieri and colleagues (2019) define offensive language as insults, threats, and any form of untargeted profanity (p. 2). The concept of toxicity is used rather inconsistently among studies and applications and is mostly derived from practical moderation instead of theoretical considerations (Paasch-Colberg & Strippel, 2022; Risch, 2023). In their API *Perspective,* Google defines toxicity as

rude, disrespectful, and unreasonable commenting behavior that "is likely to make a user leave a discussion" (Hosseini et al., 2017; Perspective, n.d.). Both in societal discourse and in practice as well as in several disciplines, a further well-established concept is hate speech (e.g., Brown, 2017; Fortuna & Nunes, 2018; Paasch-Colberg & Strippel, 2022; Paasch-Colberg et al., 2021; Schmidt & Wiegand, 2017; Wilhelm et al., 2020). Hate speech can be understood as a particularly severe form and sub-concept of incivility that is mostly defined as harmful, negative statements that stereotype, demonize, or dehumanize people based on certain group-specific characteristics, including race, religion, gender, or sexuality (Sponholz, 2023; Wilhem et al., 2020, Brown, 2017; Paasch-Colberg & Strippel, 2022; Heinbach, in press). This thesis touches on different concepts of incivility as well as sub-concepts or related notions of incivility, including offensive language, toxicity, and hate speech. Unless stated otherwise, incivility will be used as an umbrella term including these related concepts and forms of uncivil communication described above.

2.2 Incivility as a Challenge for Online Platforms and Moderation

Along with the increasing participation online, academic and societal discourse on incivility has significantly shifted from offline communication settings (e.g., Cortina et al., 2001; Mutz, 2007; Mutz & Reeves, 2005) to online discussions over the last years (Esau et al., 2021; Kersting, 2019; Kümpel & Rieger, 2019). A significant part of research has investigated incivility in comment sections of media outlets as a form of audience participation in online journalism (e.g., Rowe, 2015; Springer et al., 2015; Stroud et al., 2016; Su et al., 2018; Van Duyn & Muddiman, 2022; Van Duyn et al., 2021; Ziegele & Jost, 2020, Ziegele, Weber et al., 2018). Studies that investigated the prevalence of incivility on American news media sites found that 12% to 53% of the comments posted were uncivil (Coe et al., 2014; Santana, 2014; Su et al., 2018). Content analyses of German news sites revealed that 13% to 25% of the comments included some kind of incivility (Esau et al., 2021; Ziegele, Jost et al., 2018). Further empirical research suggests several negative effects of incivility in comment sections of news sites. For example, studies found that uncivil comments can prevent other users from participating in a discussion (Springer et al., 2015; Stroud et al., 2016; Ziegele, Jost et al., 2018), that incivility can increase polarization (Anderson et al., 2014; Hwang et al., 2014) and stereotyping (Hsueh et al., 2015), can lead to negative evaluations of other commenters and the discussion content (Wang, 2020), and can decrease the perceived quality of a news outlet (Prochazka et al., 2018).

Many scholars state that a key element of preventing and countering uncivil discussion cultures online is *moderation* (e.g., Friess et al., 2021; Heinbach & Wilms, 2022; Wright, 2006; Ziegele & Jost, 2020). Professional online moderation can be defined as actions that aim to structure and guide debates for the purpose of creating a desired discourse atmosphere (Heinbach & Wilms, 2022, p. 218). Various studies suggest that professional moderation can have positive effects on the quality of online discussions. Ruiz and colleagues (2011) found that different kinds of moderation led to a drastic decrease of insults in the comment sections of European and American newspaper websites. In an online experiment, Ziegele and Jost (2020) showed that factual responses by moderators to uncivil comments positively affected the perceived discussion atmosphere and could increase users' willingness to participate. In a field experiment with an American news station, Stroud and colleagues (2015) found positive effects of the engagement of professional moderators on the civility of comments, for example, by asking and responding to legitimate questions and sharing additional information. In a semi-automated content analysis of comment sections of American news sites, Ksiazek (2018) showed that journalistic engagement was associated with lower levels of hostility in the discussion threads. Similar effects have been reported for moderation by non-professional actors, such as users (e.g., Kalch & Naab, 2017) or activist groups (e.g., Friess et al., 2021).

For online platforms and news websites in Germany, the moderation of incivility is not only an ethical question, but also a legal issue. Since 2017, in Germany the removal of illegal content is demanded by law in the *Network Enforcement Act* ("*Netzwerkdurchsetzungsgesetz"*, *NetzDG*). Currently, the NetzDG requires operators of online platforms to remove obviously illegal content within 24 hours, including hate speech, insults, incitement of the people, and depiction of violence and threats (BMJ, 2022; Kümpel & Rieger, 2019). These regulations affect large social media platforms such as Facebook and YouTube as well as providers of journalistic content (BMJ, 2022). In 2024, the NetzDG will be extended and replaced by the *Digital Service Act* (*DSA*) that regulates the handling of such forms of incivility on EU level (European Commission, 2023). Moderators and community managers play an important role in enforcing legal regulations by reviewing and, if necessary, deleting illegal comments. However, the growing number of user contributions push platform operators and moderators to the limits of capacity. Consequently, not all online news sites rely on moderation but close the comment sections to prevent uncivil comments in advance, at least for controversial topics (Nielsen, 2012; Reich, 2011; Ziegele & Jost, 2020). To support moderation, many news sites increasingly rely on software solutions that aim to facilitate and structure the moderation process. More recently,

these solutions further offer machine learning-based support to automatically detect and highlight harmful or illegal content (Beuting, 2021; Vox Media, n.d.). These forms of *algorithmic moderation systems* (Gorwa et al., 2020), or algorithm-based moderation (Wilms et al., 2023) have also arrived in scientific discourse. Different forms of algorithmic moderation are used in moderation tools for news outlets as well as on major social media platforms, such as YouTube, X (formally Twitter), and Facebook (Gorwa et al., 2020). During this development, the question of what exactly incivility defines and how it is to be measured has expanded to the research fields of machine learning and artificial intelligence.

2.3 Error by Concept - Challenges of Measuring Incivility

2.3.1 Contextual Factors and the Eye of The Beholder

The vast majority of incivility scholars agree that, on the whole, incivility is a concept difficult to define, because what is perceived as uncivil often is context-dependent and subjective (Bormann et al., 2022; Chen, 2017; Coe et al., 2014; Herbst, 2010; Massaro & Stryker, 2012). Herbst (2010) stated that what incivility is lies to a great amount "in the eye of the beholder" (p. 3). As a result, finding a general definition of the concept "or even describe discourse that is consistently viewed as uncivil" (Chen, 2017, p. 5) remains challenging. Numerous studies have investigated the effects of individual and contextual factors on the perception of incivility. Based on two surveys in the American context, Kenski and colleagues (2020) examined how individual characteristics including demographics and personality traits determined the perception of incivility. Results showed that females were more likely to perceive statements as uncivil than males. In another survey among users of news comment sections, Diakopoulos and Naaman (2011) found that personal aspects such as revenge or disagreement affected whether a person perceived a comment as worth flagging. In addition to these factors, characteristics of the message, the sender, and the discussion setting can influence what is perceived as uncivil or not. A study by Gervais and colleagues (2015) found that incivility was less likely to polarize when it came from the own party. Similarly, Muddiman (2017) showed in two online experiments that users perceived politicians from their own party as less uncivil than others. In two survey experiments, Sydnor (2018) investigated how the channel and structure of a media platform can influence the perception of incivility. She found that individuals rather noticed incivility on Twitter than on other media platforms. Furthermore, discussion guidelines, also known as netiquette rules, can play a role for the perception of incivility. Diakopoulos and

Naaman (2011) found that users flagged comments in line with discussion guidelines, which can, however, vary between mediums and platforms, for example, due to political leaning and understanding of discussion culture.

Further research investigated how different roles of communication participants can influence perceptions of incivility. In a focus group study with users, online activists, and professional community managers, Bormann (2022) examined what these actors in public online discussions perceived as norm-violating communication. She found that members of online activist groups tended to be especially sensitive to various forms of norm violations, while community managers' perception of incivility was often in line with moderation and legal guidelines. However, for journalistic actors, individual preconditions such as personal background and journalistic self-image can further influence what they perceive as uncivil or not (e.g., Chen & Pain, 2017; Boberg et al., 2018; Harlow, 2015). Moreover, the perception of incivility may vary within actors of a certain group, for example, moderators or journalists. Paasch-Colberg and Strippel (2022) conducted 20 interviews with professional moderators of German news sites to investigate, which factors play a role in moderation decisions and what forms of user comments were considered problematic. Findings revealed that only a few types of incivility were clearly recognized as such, including calls for violence, Holocaust denial, and discrimination. Interviewees reported that deciding whether a comment should be deleted or not required experience and oftentimes double-checking with colleagues.

Overall, a variety of research underlines the assumption that incivility is a greatly subjective and context-dependent concept. What is perceived as uncivil depends on several factors on different levels of the communicative context, from individual background and professional role, over message characteristics, to discussion guidelines. These factors not only influence what is perceived as uncivil by users and moderators, but can further affect the measurement of incivility.

2.3.2 Forms and Sub-Concepts of Incivility - Easy or Easy to Measure?

Besides empirical work on the context-dependent perception of incivility, research further examined if different forms of incivility are perceived as differently harmful and severe. In an experimental study, Muddiman (2017) investigated whether users perceived so-called *personal-level incivility*, including insults, obscene, or emotional language, differently from *public-level incivility* that violated democratic norms, and found that personal-level incivility was perceived as the more severe. In a later survey among 1,000 Americans, Muddiman (2019)

found that participants perceived name-calling, profanity, and personal attacks as more uncivil than partisan conflict or descriptions of uncooperative behavior of politicians. These results coincide with findings of two surveys by Kenski and colleagues (2020), which showed that name-calling and vulgarity were perceived as more uncivil than, for example, pejorative for speech or lying accusations. Results of an online experiment by Bormann and colleagues (2023) are also in line with these findings. The authors found that participants of online discussions evaluated insults and vulgarity as more severe than other norm violations, including stereotyping, false information, and topic deviation. Additionally, an experiment by Kalch and Naab (2017) showed that users rather engaged with comments, for example, by flagging, that included explicit insults, vulgarity, or abusive language. These differences in perception of the harmfulness of incivility also play a role in comment moderation by journalists or moderators. Muddiman and Stroud (2017) analyzed the moderation of the "New York Times" online forum and found that journalists rather reacted to swear words than to other forms of incivility and further, comments including swear words were more likely to be removed by them.

Empirical studies not only suggest that distinct forms of incivility are perceived as differently harmful but are also differently straightforward to identify. Based on their analysis of the "New York Times" online forum, Muddiman and Stroud (2017) assumed that journalists reacted more to comments that include swearing, not only because they were perceived as more severe, but also because swear words were easier to detect. They concluded that journalists therefore might find it easier to justify the rejection of a comment. At the same time, deleting comments based on other forms of incivility such as profanity would be more difficult to comprehend. Boberg and colleagues (2018) also conclude in their content analysis study that swearing and obscene language are easier to identify than more subtle forms of incivility, also for algorithms. Interestingly, research further supports the assumption that what people state to consider as severely uncivil may differ from what they actually perceive. In a survey among members of the online activist group *#ichbinhier* by Ziegele and colleagues (2020) participants stated that they considered both incivility such as insults and incivility that, for example, rejected democracy, as equally harmful. In her focus group study, Bormann (2022) also found that actors tended to evaluate not only incivility such as insults or name-calling as highly severe, but also incivility that violates democratic norms, such as doubting freedom of press, free speech, or questioning democracy. However, in a related online experiment where the perception of forms of incivility were tested, participants voted the letter as less severe (Bormann et al., 2023). Muddiman (2019) also found that when asked about severity of incivility, people

primarily mentioned violations of democratic norms. However, when confronted with incivility in an experimental setting, they were likely to find incivility on a personal level such as insults as more harmful. One reason for these findings might be that swear words, insults, or personal attacks are easier to identify than, for example, anti-democratic or stereotyping statements that might be expressed in a more implicit way. In sum, research suggests that different forms of incivility are perceived as differently severe and are differently straight-forward to identify, both for different actors of online discussions and for algorithms. In addition to the context-dependency of incivility, these deviations pose an additional challenge for the definition and the measurement of uncivil content.

2.3.3 Incivility as a Challenge for Standardized, Text-Based Methods

Research designs to investigate the prevalence of incivility require methods that measure incivility based on textual data. *Quantitative content analysis* is a genuine communication science method and is applied to manually analyze large samples of text documents in a standardized manner (Früh, 2015; Krippendorff, 2018; Rössler, 2017). Based on an underlying theoretical framework, categories are derived and operationalized to be then measured based on text. In the sense of standardized, quantitative measurement, much effort is done to ensure that a concept is perceived equally by several coders (or annotators). To this end, comprehensive coding instructions are developed to provide an equal understanding of, here, uncivil content, usually followed by extensive training of the coders and several reliability tests (Hayes & Krippendorff, 2007). The reliability of measurement, meaning that a concept is perceived and measured equally among coders or researchers, is a key quality criterion of quantitative content analysis. From the perspective of standardized methods, this attempt is straightforward, since samples are later coded by only one person in most cases and hence, individual deviations in the measurement could distort the meaningfulness of research results.

Yet, previous research allows the assumption that incivility is a concept challenging to detect with text-based methods (e.g., Bormann et al., 2022; Bormann & Ziegele, 2023; Chen, 2017; Chen et al., 2019). First, because what is perceived as uncivil or not is greatly influenced by the individual or situational context, which is usually not fully reflected in the text, but lies in the eye of the recipient (Herbst, 2010). Second, different forms of incivility are perceived as differently severe and furthermore, are not equally straightforward to identify. In their conceptualization of incivility, Bormann and colleagues (2022) argue that incivility as a form of norm violation is determined by the perception and disapproval of users. Nevertheless, the authors

also state that the factor of perception and the resulting variance is not considered in the method of quantitative content analysis, as it contradicts the assumptions of standardized measurement. So far, communication scholars have considered the context-dependent deviations of incivility perception rather as measurement error and hence, a methodological side issue that stands in the way of valid and reliable measurement (e.g., Bormann & Ziegele, 2023). With the growing relevance of algorithm-based moderation, however, the question of how incivility is to be defined and to be measured validly becomes a pressing research question itself that affects several disciplines and areas of research, such as computer science and artificial intelligence research.

2.4 Conclusion and Takeaways

Among incivility scholars the understanding of what exactly incivility determines varies. A great body of research conceptualized incivility as a violation of deliberative and democratic norms that hinders democratic discourse between citizens (e.g., Andersson et al., 2014; Brooks & Geer, 2007; Coe et al., 2014; Kalch & Naab, 2017; Hwang et al., 2018; Rowe, 2015). Other scholars understand incivility as a form of impoliteness that contradicts norms of respect between participants (e.g., Brown & Levinson, 1987; Chen & Lu, 2017; Mutz & Reeves, 2005). In societal discourse and in neighboring disciplines, further concepts related to incivility are established, for example, hate speech, as a specifically harmful form of incivility, as well as offensive and toxic language (e.g., van Aken et al., 2018; Zampieri et al., 2019). However, scholars agree that what constitutes incivility is always a question of context and individual perception. Research has identified multiple factors that influence what is perceived as uncivil or not, for example, individual background (e.g., Kenski et al., 2020), platform policies (e.g., Diakopoulos & Naaman, 2011), and the role a person holds in a discussion (e.g., Bormann, 2022). Moreover, studies revealed that different forms of incivility are perceived as differently harmful. In experimental settings, users tended to evaluate swearing and insults as more severe than incivility that, for example, undermines democratic norms (e.g., Bormann et al., 2023; Muddiman, 2019). One explanation might be that obvious forms of incivility including swear words or insults are easier to recognize (e.g., Boberg et al., 2018; Muddiman & Stroud, 2017). The variations on the perceptual level of incivility have major effects of its valid, reliable measurement. Meanwhile, these challenges not only arise for standardized content analysis, but also for algorithm-based moderation systems that recently have gained popularity among several platforms and providers of online discussions (Beuting, 2021). As modern machine learning-

based applications are designed to replicate human decisions, the question of how to decide what constitutes incivility increasingly concerns the research field of artificial intelligence.

3 Machine Learning - Expertise from Computer Science

Latest developments in artificial intelligence and machine learning have transformed our society remarkably and have influenced a variety of research and application areas. Among them, one major field is *natural language processing (NLP)*. Meanwhile, the automated detection of concepts related to incivility such as hate speech, toxicity, or offensive language has evolved into an established NLP task, usually approached as a *supervised learning* problem (*classification*). That means that a machine learning model is trained to identify uncivil content based on human decisions (see chapter 3.1.1). Over the past years, several approaches including *deep learning* and *feature-based learning* have successfully been applied to the task of incivility classification (chapter 3.1.2 and 3.1.3). This chapter will give an overview on different machine learning approaches and methods in NLP with a focus on incivility classification and its methodological challenges.

3.1 Natural Language and Artificial Intelligence

In the last 25 years, innovations in artificial intelligence and machine learning have had a profound impact on research and society. The simultaneous increase of computational power, digitally available data, and the development of complex, powerful algorithms have enabled significant advances in artificial intelligence across a wide range of industries and research areas (Crawford, 2021; Jordan & Mitchell, 2015). Among them, one major field is NLP, which can be described as a collection of computational tools and techniques to analyze and represent human language (Chowdhary, 2020; Nadkarni et al., 2011). However, natural language remains a demanding and challenging data type for mathematical and statistical modeling. One major reason is that natural language is *arbitrary*. That is, the meaning of, for example, words, is not logically derivable from text or speech, but is shaped through convention and context. Second, natural language is *ambiguous*. That means that oftentimes context information is needed to understand the meaning of a word or an expression. This ambiguity can appear on several levels of language, for example, in the form of synonymy of words or of the intention of the statement (Navigli, 2009).

The computational analysis of natural language as a genuine expression of human intelligence has been connected to artificial intelligence research since its early beginnings. However, it was not before the late 1980s that machine learning provoked a significant shift in the landscape of NLP (Jones, 1994). Before this turn, the vast majority of scholars in the fields of

linguistics, psychology, as well as artificial intelligence and NLP agreed on the belief that intelligent systems have to be created by hand coding knowledge, rules of decision making, or mechanisms for reasoning, into them (Manning & Schütze, 1999; Chomsky, 1965; Chomsky, 1986). Stimulated by grammatical theories developed in linguistics between the 1950s and the 1970s, this trend has dominated NLP research during this twenty-years period of time (Jones, 1994). In the late 1970, however, *predictive modeling* became more present in artificial intelligence research and in NLP (Johri et al. 2021). Even though important achievements were made in this time that have been crucial for nowadays machine learning algorithms, it needed another 20 years to take the step to the current state of research in machine learning - deep learning (Schmidhuber, 2015). In contrast to former approaches, in deep learning a programmer or researcher does not provide rules for decision making. Instead, the algorithm itself deduces the process of mapping an input to an output. While earlier attempts have failed due to limited data and low computational power, technological advances and huge amounts of digitally available data pioneered the success of deep learning, which now represents the current state of the art in artificial intelligence and NLP (LeCun et al., 2015).

3.1.1 Supervised vs. Unsupervised Learning

Currently, the vast majority of state-of-the-art machine learning applications are built on supervised learning (Schmidhuber, 2015). In supervised learning, an algorithm learns from examples, which oftentimes means human decisions. Here, the goal is to model a statistical relationship between input (features, independent variable) and output (dependent variable), for example, the relation between written words and the content or meaning of a text. A well-trained model will eventually be able to predict the outcome values for new, unseen data, where the output values are not known (Bishop, 2006). For supervised learning, a sufficient amount of *labeled* data is needed for training. That means, not only the input but also the output values must be available in the training data. Is the output value categorical, it is called class or category and the task is called classification. In recent years, the automated identification of concepts related to incivility including hate speech (e.g., Burnap & Williams, 2015; Davidson et al., 2017; Ross et al., 2017), offensive language (e.g., Zampieri et al., 2020; Wiegand et al., 2018), and toxicity (Georgakopoulos et al., 2018; Risch & Krestel, 2020; van Aken et al., 2018) has become a novel, important classification task in NLP. In this task, documents such as user comments or Tweets are labeled by human annotators who decide whether a document contains incivility or not. Based on this labeled data, a model is trained to predict the incivility of a

comment. Even though deep learning has widely been used in the past few years, simpler features-based approaches are still being applied for classification tasks (see chapter 3.1.2).

In contrast to supervised learning, the goal of *unsupervised learning* is not to correctly predict a particular outcome, but to model underlying patterns in the data that cannot be detected by the human eye. Unsupervised learning approaches usually have the goal of some kind of dimensionality reduction and the advantage of being less expensive in terms of data resources than supervised learning, since no labeled data is needed for training (Bishop, 2006). Today, several unsupervised learning approaches are established in NLP. A widely used approach is *topic modeling*, which is the process of extracting topics in a collection of documents. Popular topic modeling algorithms include *latent semantic analysis* (Dumais et al., 1988; Valdez et al., 2018), *probabilistic latent semantic analysis* (Brants et al., 2002; Hofmann, 1999), and *latent dirichlet allocation* (*LDA*, Blei et al., 2003). While these algorithms differ in detail, they build on the shared underlying assumptions that a document includes a distribution of several topics. This assumption is usually fulfilled in journalistic news articles and related document types, which is one reason why topic modeling has been adapted quickly in communication science (Maier et al., 2018). As a combination of supervised and unsupervised approaches, so called *semi-supervised learning*, makes use of both labeled and unlabeled data. Semi-supervised learning allows the processing of large sets of unlabeled data, while a smaller amount of labeled data is used for specific training. These methods are of particular interest for scenarios where only few labeled data is available due to high costs (Van Engelen & Hoos, 2020).

3.1.2 Deep Learning vs. Feature-Based Learning

In addition to the division into supervised and unsupervised learning, machine learning approaches can further be distinguished into deep learning and feature-based learning (e.g., Risch, 2020). Many recent achievements in artificial intelligence that gained wide attention were accomplished through supervised approaches with deep learning, for example, image classification to recognize faces or objects (e.g., Krizhevsky et al., 2017; Zhao et al., 2019). Deep learning algorithms are based on artificial *neural networks*, which consist of staged layers of multiple connected processing elements, also called *neurons* (LeCun et al., 2015; Schmidhuber, 2015). There exist several types and architectures of neural networks (Goldberg, 2016). Two main types are *recurrent neural networks* (*RNNs*; Rumelhart et al., 1986), including *long short-term memory networks* (*LSTM*; Gers et al., 2000; Hochreiter & Schmidhuber, 1997), and *convolutional neural networks* (*CNNs*; Fukushima, 1980; LeCun et at., 1998). Due to their specific

architecture, RNNs have been proven especially useful for NLP tasks. The circular connection of layers allows access to previously processed information during training, such as previous words in a sentence, which can be helpful where the sequence of inputs (e.g., the sequence of words) contains information for prediction. RNNs have successfully been applied to several NLP tasks, such as text classification (e.g., Jelodar et al., 2020; Pavlopoulos et al., 2017; Wang et al., 2015) or machine translation, which is the task of translating one language into another (e.g., Rivera-Trigueros, 2022). Also for incivility classification, RNNs have successfully been applied (e.g., Bisht et al., 2020; Sadeque et al., 2019; see chapter 4.2.1). CNNs, on the other hand, are preferable when local indicators, for example, single words, are expected to have great predictive value for class membership, regardless of their position in the document or sequence (Goldberg, 2016). For example, signal words such as insults in a comment can increase the probability that the whole comment will be classified as uncivil. Approaches based on CNNs have also achieved noticeable results in several NLP tasks, for example, question answering (e.g., Dong et al., 2015; Ishwari et al., 2019) and text classification (e.g., Johnson & Zhang, 2015; Nguyen & Grishman, 2015; Wang et al., 2015), including incivility classification (e.g., Taradhita et al., 2021).

In NLP, neural networks usually contain a *word embedding* layer as input. In contrast to word frequency representations (*bag-of-words*), word embeddings display words and relations between words in a vector space where words with similar meaning happen to have similar vector representations. Therefore, they are able to consider word context (neighboring words) to some extent (Mikolov et al., 2013). Popular word embedding frameworks are *word2vec* (Mikolov et al., 2013), *Glove* (Pennington et al., 2014), and *fasttext* (Bojanowski et al., 2017), which have marked an important advance for deep learning in NLP. As the input layer in a neural network, word embeddings are trained within the model on a specific task, such as incivility classification. However, word embeddings are not designed to model a relationship between input and output, but rather learn a general understanding of how words are arranged and related to each other. Therefore, word embeddings are also useful as stand-alone, unsupervised models, for example, to analyze relations and map patterns in a certain corpus (e.g., Andrich & Domahidi, 2022; Andrich et al., 2023; Kroon et al., 2021; Mikolov et al., 2013; see also chapter 3.1.1).

Before neural networks were applied for a broad spectrum of NLP applications, simpler linear, feature-based modeling dominated machine learning for many years, including algorithms such as decision trees, linear regression, or support vector machines (Boser, et al., 1992;

Spertus, 1997). Whereas deep learning models can handle millions of parameters to fit the data, feature-based algorithms include only a few parameters and (pre-engineered) features (Goldberg, 2016). Even though powerful deep learning approaches marked a remarkable shift in NLP problem solving, feature-based approaches are still commonly used because they are less demanding in terms of training data and computational power. Furthermore, theory-driven, pre-engineered features and rules can provide a higher degree of explainability and inference on the effects of single features. Schmidt and Wiegand (2017) presented an overview study on different features that have been used to classify different forms of hate speech in recent years. Their findings revealed that researchers applied a variety of features, usually combined with bag-of-words features, meaning (weighted) frequencies of unigrams (single expressions, most often words) or n-grams (combinations of expressions, most often words), to represent a document. While bag-of-words features are reported to be highly predictive for class membership of documents, they have been enhanced with additional syntactic or semantic information, for example, *part of speech* (POS) information, meaning the syntactic function of a word (e.g., Xu et al., 2012), or dependency relationships (e.g., Burnap & Williams, 2015; Burnap & Williams, 2016; Chen et al., 2012; Nobata et al., 2016). Other approaches included additional lexical resources to engineer features for prediction, for example, publicly available word lists of general or group-specific hate-related terms (e.g., Burnap & Williams, 2016; Schmidt & Wiegand, 2017). However, these collections are often language-specific and only few resources are available for the classification of incivility, especially for German-language data.

Both deep learning and feature-based learning have proven as reasonable choices for NLP problems depending on preconditions and intentions. While deep learning is considered more powerful and to achieve higher accuracy, statistical relations are very complex and almost impossible to comprehend and understand by humans. Furthermore, deep learning is very demanding in terms of data richness, which usually means data quantity. For smaller data sets feature-based learning can still achieve satisfactory results, while deep learning approaches might be too complex to work with a limited amount of training data. Moreover, feature-based learning can be preferable over deep learning when human knowledge about rules of decision making or data characteristics can fruitfully factor into a model and might be helpful to identify a certain concept.

3.1.3 Transformers and Pre-trained Large Language Models

More recently, so-called *large language models* (LLMs) marked a remarkable shift in NLP research. LLMs are neural networks, which are based on *transformers*, a specific neural network architecture that was presented by Vaswani and colleagues from Google research lab in 2017 (Vaswani et al., 2017). Based on transformers, two of the currently most popular LLMs have been released in 2018, namely *GPT* (*Generative Pretrained Transformer*, Radford et al., 2018) and *BERT* (*Bidirectional Encoder Representations from Transformers*, Devlin et al., 2018). Even though both models differ in specific terms of training, the general idea of the training process is comparable. In a first step, the models are *pre-trained* on up to several petabytes of unlabeled text documents, while huge numbers of parameters enable very complex representations of the data. Later, the pre-trained models can be *fine-tuned* on a smaller, labeled data set to solve a specific task, for example, incivility classification. The innovative transformer architecture is based on the so-called *attention mechanism*. This novel mechanism includes two neural networks, encoder and decoder, that allow a kind of self-supervised learning. That means that while training, the model can show itself examples to learn from in an iterative process between encoder and decoder (Devlin et al., 2018; Radford et al., 2018; Vaswani et al., 2017). While models of the GPT-family learn by guessing what is most likely to come next in a sentence, BERT learns by predicting missing (*masked*) parts of a sentence (also referred to as masked language models) (Balestriero et al., 2023; Devlin et al., 2018). In recent years, LLMs based on transformers have continued to evolve (e.g., Brown et al., 2020; Liu et al., 2019; Radford et al., 2019) and still set the current state of the art for most NLP problems (Khurana et al., 2023).

The two-step training procedure of LLMs has the main advantage that major resources regarding computation and data need to be raised only once. The exhaustive pre-training process is further conducted on unlabeled data, which usually means that more data is available at lower costs. Instead of training a whole model from scratch, a pre-trained language model can then be transferred and adjusted (e.g., fine-tuned) using a significantly smaller amount of computational power and labeled data. Despite these advantages, the downside of this two-step procedure is a loss of control and transparency. Researchers are dependent on the availability of suitable, off-the-shelf models. Further, to use a pre-trained model, it has to fit the down-stream task regarding data characteristics, such as language or text type. In June 2019, deepset released the first freely available BERT model for German language that has been trained on 12 GB of German-language online data (deepset, 2019). Meanwhile, multiple BERT models are available

for languages other than English, for example, *CamemBERT* (for French, Martin et al., 2019), *AraBERT* (for Arabic, Antoun et al., 2020), or *KR-BERT* (for Korean, Lee et al., 2020). For incivility classification as well as for many classification tasks, approaches based on BERT and its derivatives such as *RoBERTa* (Liu et al., 2019) remain the current state of the art (Satapara et al., 2022; Zampieri et al., 2023).

3.2 The Data in Machine Learning

Overall, data is considered the most significant element of machine learning that greatly affects a model's performance and outcome. Alongside the technological process in computing, the availability of great amounts of digital data first enabled powerful deep learning approaches and thus, led the ground for nowadays success of artificial intelligence (LeCun et al., 2015; Schmidhuber, 2015). Due to the crucial role of data quantity and its associated costs, machine learning research sometimes addresses the question of suitable data in a somewhat pragmatic way. Since labeled data is usually scarce and expensive, free access to a data set remains a strong argument, even though it may not be a perfect fit for the current use case or task. For example, large data sets that have been compiled once and that are available freely are continued to be used as reference or benchmark data sets to develop and evaluate novel approaches and applications (Crawford, 2021). For NLP, among these established data sets is the *Large Movie Review Dataset* (Maas et al., 2011), which contains several thousand movie reviews from the *Internet Movie Database* (*IMDb*) and remains a popular training and benchmark data set for sentiment analysis research (e.g., Qaisar, 2020; Yasen & Tedmori, 2019).

However, NLP tasks usually face the challenge of high context dependency. Characteristics of the training data regarding text type and source may not fit the specific use case or task, which can limit the usefulness of available data sets. For example, language patterns and vocabulary can differ strongly between journalistic and user-generated content. Furthermore, data sets in languages other than English are still rare for many NLP tasks. This issue concerns both available benchmark data sets and off-the-shelf language models, such as BERT. For the task of incivility classification, there are some data sets available, which, however, often address one sub-concept of incivility, such as offensive language (e.g., Davidson et al., 2017), hate speech (e.g., Albadi et al., 2018; Bohra et al., 2018; Davidson et al., 2017), or bullying (Tahmasbi & Rastegari, 2018). For the German language, the number of available resources is even more limited (exceptions are Ross et al., 2017; Wiegand et al., 2018). Even though these public data sets can be valuable for classifying incivility, they are often of limited use for specific

research questions due to limitations regarding the sample (e.g., in terms of quantity or time span) and the conceptualization of incivility (Vidgen & Derczynski, 2020).

3.3 The Human in Machine Learning or What is a Good Model?

Currently, the majority of machine learning applications are built, at least partially, on supervised learning (Bishop, 2006; Schmidhuber, 2015). Here, models are trained to reproduce decisions in the training data (see chapter 3.1.2). In supervised learning, these usually human-generated decisions are also referred to as *gold standard* or *ground truth* (Basile et al., 2021). For many applications huge amounts of labeled data are needed for training. To not exceed costs, training data is often created with the help of *crowd annotation*. Here, comments are not labeled by researchers or experts, but by several hundred or thousands of crowd workers that are accessed via a commercial provider (Haselmayer & Jenny, 2017; Lind et al., 2017). The crowd coding approach is mostly established in industry, for example, for sentiment analysis in product reviews or image classification. Meanwhile, numerous commercial platforms offer this service (Crawford, 2021). In contrast to communication science content analysis, the final gold standard labels in crowd-annotated data are not based on the decision of one coder, but are calculated by majority voting or averaging the decisions of several coders (Alm, 2011; Basile et al., 2021; Romberg, 2022).

For supervised learning, the performance of a model is measured by to what extent the predicted values correspond with these gold standard labels. There exist several measurements that indicate the agreement between predicted and *true* values. The simplest measurement is the *accuracy*, which measures the amounts of matching values between model and gold standard. In machine learning, the accuracy is an established measurement and is reported in most cases to give an impression of overall model performance. In this sense, a "good" model is defined as a model that achieves high accuracy, meaning a high agreement with the gold standard. For many research questions, however, it is important how a model performs in *one* specific class, for example, in identifying uncivil in contrast to not uncivil comments. There are several measures to report the performance within one specific class (e.g., either uncivil or not uncivil), including the *F1 score*, *recall*, and *precision* (Zhou, 2021). Based on these measures the *macro F1 score* can be calculated, which is usually defined as average mean of recall and precision over all classes and hence, provides an informative overall performance measure for a classification model.

As supervised learning models are optimized to achieve a high agreement with the (human) decisions in the training data, the question of how and by whom the data is labeled is crucial, especially for decisions that are subjective to some degree. Studies revealed that the decisions of coders are dependent on several factors, including knowledge, background, gender, and personal perceptions (e.g., Dixon et al., 2018; Vidgen et al., 2019; Wich et al., 2020). A study by Binns and colleagues (2017) showed that male and female coders labeled toxicity differently and suggest that demographic factors should be considered in annotation (Binns et al., 2017). Ross and colleagues (2017) found that coders did not share a common understanding of what incivility is, regardless of whether they had received coding instructions or not. These findings are in line with empirical and conceptual research on incivility, which suggest that the perception of incivility is context dependent and subjective (e.g., Chen et al, 2019; Herbst, 2010) and thus, questions the assumption of one common perception.

3.4 Conclusion and Takeaways

Latest developments in machine learning have marked a remarkable performance increase for numerous NLP tasks, including document classification. For incivility classification, there exists a variety of approaches, including several feature-based and deep learning techniques (Schmidt & Wiegand, 2017). Lately, the introduction of LLMs such as BERT (Devlin et al., 2018) have marked a significant shift in the NLP landscape. For many tasks, LLMs that have been pre-trained once using enormous data amounts and computing resources can be transferred and fine-tuned on a smaller labeled data set to solve a specific problem, such as incivility classification. Although this approach has proven to be powerful, it comes with the disadvantage of low transparency and control as well as dependency on third parties. In general, pre-trained models, data sets, and other NLP resources are rare for German, especially for specific research questions. For incivility classification, only few data sets are available for research that are further limited regarding sampling and the conceptualization of incivility (Ross et al., 2017; Wiegand et al., 2018). As creating sufficient training data sets is often expensive, machine learning research usually employs a great number of crowd workers instead of intensively training a handful of coders. Usually, the final crowd-annotated values are then aggregated to one gold standard label via majority voting or averaging (Basile et al., 2021). Even though working with trained coders in quantitative content analysis and the more intuitive crowd annotation with untrained crowd workers differ in the specific procedure, both approaches aim for one unique, standardized measure. For incivility, however, research suggests that perception varies

depending on several individual and contextual factors (e.g., Binns et al., 2017; Chen et al., 2019; Herbst, 2010; Wich et al., 2020). Hence, the idea of one common perception of what constitutes incivility can be questioned. In this sense, the question arises how "good" a machine learning model for incivility detection can really be, if the quality of the prediction is usually measured based on an averaged perception of crowd workers, whose views are further influenced by multiple factors. This discrepancy will be a major challenge for the classification of incivility, both in research and for machine learning-based moderation systems.

4 Contributions of this Thesis

This cumulative dissertation addresses the overall research question of *how to measure incivility using machine learning methods*. To this end, I combine conceptual and theoretical knowledge of the communication science research subject incivility with state-of-the-art machine learning methods from computer science and artificial intelligence research. In this chapter, I present and discuss the contributions of the seven research articles of the cumulus related to three further subordinate research questions. In chapter 4.1, I address the initial question: *To what extent can incivility be measured using machine learning methods?* Therefore, I present how publication [1] and [2] contribute to the current state of research on incivility classification for German-language online discussions and what assumptions about the measurability of incivility can be derived from this work. Chapter 4.2. deals with the question of *how machine learning can extend and elaborate communication science research on incivility.* Publications [3], [4], and [5] contribute to this question by showing how machine learning can be integrated at different points of the communication science research process, including research questions, research designs, and methods. Finally, in chapter 4.3 I address the question of *what methodological challenges communication scholars face with machine learning-based incivility classification and how they can be overcome.* For this purpose, I present the contributions of research articles [6] and [7], which discuss various pitfalls of incivility classification and present and evaluate approaches to conquer them.

4.1 State of Research - Incivility Detection Using Machine Learning

In this chapter I, address the question: *To what extent can incivility be measured using machine learning methods?* To this end, I present the contributions of the research articles [1] and [2] to the current state of research on the detection of incivility in German-language online discussions and derive assumptions about the measurability of incivility using current machine learning methods. In research article [1] "*hpiDEDIS at GermEval 2019: Offensive Language Identification using a German BERT model*" (Risch, Stoll, Krestel, & Ziegele, 2019), we first applied and evaluated a German BERT model on the task of incivility classification, taking an important step towards the use of LLMs for this specific research question. In publication [2] "*Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments*" (Risch, Stoll, Wilms, & Wiegand, 2021), we introduced the proceedings of the latest *GermEval* shared task on incivility classification. This task contributes both a

benchmark data set for incivility detection for German-language online discussions and a comparative format to provide an overview over current approaches to this task.

4.1.1 Current Approaches to Incivility Detection

The ongoing application, adjustment, and evaluation of new approaches to specific tasks is an important part of machine learning research. In recent years, LLMs such as BERT (Devlin et al., 2018) have achieved remarkable performances on several NLP downstream tasks. LLMs are usually pre-trained on a huge corpus of unlabeled documents and are then fine-tuned on a smaller, specific data set of labeled documents (see chapter 3.1.3). To use such LLMs for classification the data a model has been pre-trained on must fit the specific research subject. In NLP the most obvious requirement is a fit regarding language. However, BERT presented by Devlin and colleagues in 2018 was first exclusively trained on English-language data (Devlin et al., 2018). After the authors had released the model to open-source in late 2018, researchers from different nations were able to reproduce BERT on new data and apply it to tasks in languages other than English. In publication [1] "*hpiDEDIS at GermEval 2019: Offensive Language Identification using a German BERT model*" (Risch, Stoll, Krestel, & Ziegele, 2019), we have been among the first to apply and evaluate an off-the-shelf BERT model for incivility classification that had been pre-trained on German-language data, which was provided by deepset in 2019 for research purposes (deepset, 2019). The model was trained on 12 GB of German-language data, including the German portion of the *Wikipedia* dump, *Open Legal Data* dump, and German news articles. The model is case sensitive, meaning upper and lower case was taken into account for training. Publication [1] is part of the *GermEval 2019 Shared Task on the Identification of Offensive Language* (Struß et al., 2019) that invited research teams to develop approaches for the identification of different forms of offensive language in Tweets, including profanity, insults, abuse, and explicit and implicit offensive language. GermEval is a series of shared task evaluation campaigns endorsed by the *German Society for Computational Linguistics (GSCL)*, an association that focuses on NLP for the German language and provides a variety of tasks and topics. Teams from academia and industry can participate to develop, evaluate, and present approaches on a provided data set. Within this shared task, we have been among the first teams that applied a German BERT model for incivility classification. In our approach, we further developed an innovative strategy to create an ensemble of multiple model predictions using soft majority voting. This method is promising, since results usually differ between runs due to

varying initialization weights for the prediction head of the model. Using this ensemble strategy, the final predictions are less dependent on the random weights of a single run. Within the competition, our approach achieved the best result for binary implicit vs. explicit offensive language classification with a macro-average F1 score of 73.1 (accuracy= 86.8) on the provided test set. Moreover, our approach achieved 76.4 on the coarse-grained binary classification task (offensive vs. not offensive) and 51.2 macro-average F1 score on fine-grained classification (profane vs. abuse vs. insult, for further details of the task description see Struß et al., 2019). In contrast to earlier GermEval tasks (e.g., Wiegand et al., 2018), at GermEval 2019 all best performing approaches (coarse-grained accuracy= 76.4; fine-grained accuracy= 73.6) first applied a BERT-based approach (e.g., Paraschiv & Cercel, 2019). Therefore, this shared task marked a significant shift for incivility classification using machine learning methods towards LLM-based approaches (Struß et al., 2019).

4.1.2 Providing Benchmark Formats and Data Sets

The accumulating process of machine learning research is supported by the provision of benchmark data sets, tasks, and formats with a comparative aspect. Formats such as the GermEval shared tasks can not only map the current state of research for a specific problem, but further provide access to benchmark data sets for training and testing novel approaches. Especially for supervised learning, these data sets are of great value as the creation of labeled data is often very costly (see chapter 3.2). Inspired by the popular international format SemEval (Zampieri et al., 2019; Zampieri et al., 2020), the GermEval shared tasks constitute an established format for offensive language detection in the German-language NLP community (e.g., Benikova et al., 2014; Wojatzki et al., 2017). With publication [2] "*Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments*" (Risch, Stoll, Wilms, & Wiegand, 2021), we continued the previous GermEval 2018 and 2019 shared tasks on offensive language identification in 2021 (Struß et al., 2019; Wiegand et al., 2018). For the GermEval 2021 shared task, we provided an extensive, high-quality data set for incivility classification that further included two novel classes, namely *engaging* and *fact-claiming* comments. The class "engaging" served as a collective category of high-quality user contributions that is based on deliberative quality (e.g., Dahlberg, 2001; Fraser, 1990; Habermas, 1996). The classification of fact-claiming comments aimed to identify contributions that included assertions and external sources. Since this publication is also directed at practitioners and scholars from other disciplines, we used the term toxicity as a synonym for incivility as incivility is

primarily established in communication science. Incivility was operationalized based on the framework by Papacharissi (2004) and is composed of violation of politeness and deliberative norms, including screaming, vulgar language, insults, sarcasm, discrimination, discreditation, and accusation of lying. The provided data set included 4,188 labeled Facebook user comments of a German public television broadcaster and was annotated using a theory-based annotation scheme. The intercoder agreement between the coders was at a satisfactory level of 0.72 (Krippendorff's alpha) and higher in all categories. Overall, the data set included 1,472 comments labeled as uncivil (35.1 percent), 1,118 comments labeled as engaging (26.7 percent), and 1,417 comments labeled as fact-claiming (33.8 percent). With our sampling strategy for the training and testing data, we followed a real-world application scenario where the sample a classifier is trained on is preceded by the test set in time. In sum, 15 teams participated in the shared tasks and submitted 87 system runs. The best performing teams achieved a macro-average F1 score of 71.8 for incivility, 70.0 for engaging, and 76.3 for fact-claiming comment classification (binary). The majority of teams applied a BERT-based approach and fine-tuned a pre-trained model on the training data. It has been shown that among these rather similar approaches, the setting of hyper-parameters significantly affected the final performances. Hence, the access to high-performance computing appeared to have great influence on the fine-tuning process and thus, the final performance of an approach.

With the 2021 GermEval shared task, we provided a novel and high-quality data set that can serve as a benchmark for both incivility classification and the identification of the two categories engaging and fact-claiming. Such data sets are of particular value, since the number of comparable data sets, especially for the German language, is very small (see chapter 3.2). The labeled data sets have been released in anonymized form and are freely available for research purposes. With the proposed set of categories, we aim to address contemporary needs of community managers and moderators to not only react to uncivil comments but also to highlight and engage with high quality content and further, support the still manually executed task of fact-checking. In sum, this publication provides an overview of the capabilities and challenges of current approaches to incivility classification and marks an important reference for researchers in academia and industry that are working on algorithm-based content moderation for German-language online discussions.

4.2 Machine Learning for Incivility and Communication Science Research

In recent years, machine learning methods have become more visible in communication science, although they are not fully established yet. So far, there is still a lack of work that bridges the gap between machine learning as a method of computer science and communication science research. In this chapter, I address the question of *how machine learning can extend and elaborate communication science research on incivility* at different points in the communication science research process, including research designs, research questions, and research methods. The three research articles [3], [4], and [5] contribute to the consolidating line of research consisting of methodological work, case studies, and best practice studies that demonstrate how machine learning can enhance communication science research on incivility.

4.2.1 Automated Content Analysis and Computational Communication Science

As a genuine communication science method, quantitative content analysis is firmly anchored in communication science research (Früh, 2015; Loosen & Scholl, 2012; Rössler, 2017). It is therefore not surprising that the analysis of text documents with automated or computational methods is referred to as *automated content analysis* as an umbrella term for NLP approaches and methods that are applied to content analysis research questions (Boumans & Trilling, 2018; Brosius et al., 2022). Even though the work by Scharkow in 2012 can be considered an early milestone for automated content analysis with machine learning in communication science, machine learning methods are establishing slowly alongside more basic, descriptive NLP techniques (Boumans & Trilling, 2018; Scharkow, 2012). Especially *dictionaries* have become very popular for automated content analysis. In dictionary approaches, a list of predefined expressions is matched with text documents to classify them into certain categories of content. In contrast to rather complex machine learning methods, approaches based on dictionaries or word counts have the advantage of being low-threshold and easy to apply within several software solutions and without requiring advanced programming skills (Günther & Quandt, 2018; Lind et al., 2019). Meanwhile, dictionaries have been applied to a wide range of communication science research questions and subjects. These include the analysis of news articles regarding populism (Rooduijn & Pauwels, 2011), sentiment (Young & Soroka, 2012), disinformation and conspiracy narratives (Boberg et al., 2020), framing (Lind et al., 2019; Meltzer et al., 2021), or migrant women's salience (Lind & Meltzer, 2021), as well as the analysis of online discussions and user comment sections, for example, regarding hate speech (Davidson et al., 2017) and incivility (Ksiazek et al., 2015; Muddiman & Stroud, 2017). A recent study by

Hase and colleagues (2022) revealed that the majority of scholars who applied automated content analysis to analyze journalistic content used dictionaries, followed by related methods such as *co-occurrences* and similarity metrics to map and describe language patterns of documents in a corpus.

Nonetheless, over the last five years, machine learning has arrived in communication science research where it is now primarily applied for automated content analysis, especially for topic modeling (Hase et al., 2022). Also in incivility research, machine learning approaches have been sporadically applied (e.g., Davidson et al., 2020; Sadeque et al., 2019; Theocharis et al., 2020; Ziegele, Daxenberger et al., 2018). Sadeque et al. (2019) used a RNN with fasttext embeddings (see chapter 3.1.2) to classify sub-forms of incivility, which was a popular state-of-the-art approach at this time. One year later, Davidson and colleagues (2020) developed and evaluated a BERT-based classifier to identify incivility on Reddit and Twitter. In a current paper Rains and colleagues (2023) also applied a BERT model to classify incivility in Tweets. Despite the growing popularity of computational methods (Domahidi et al., 2019; Hase et al., 2022; Niemann-Lenz et al., 2019; van Atteveldt et al., 2019), machine learning is still establishing itself in communication science research. Further, as experts of content analysis, communication scholars do not only consider the saving of expenses in the usage of computational methods but also the risk of oversimplified, unreliable measurement. For incivility these doubts are not unreasonable as the analysis of uncivil content already causes challenges for *manual* content analysis (see chapter 2.3.3) During this consolidating stage, there is a need for research that contributes case studies and best practices on how to apply and evaluate machine learning methods for specific communication science research areas. In the following, I present how publications [3], [4], and [5] are meeting this demand for the research subject incivility. Article [3] is a case study that shows how incivility classification can be used to extend a classic content analysis research design for the analysis of online discussions, allowing for a larger sample and a more sophisticated analysis method. Publication [4] is a study with a methodological focus that shows how feature-based incivility classification can be applied to test theoretical concepts and assumptions of incivility. It further provides an overview of the options and opportunities of feature-based approaches to incivility detection. Publication [5] is a methodological article that describes the development of the first comprehensive German-language dictionary for incivility, which was extended by means of machine learning. In doing so, the article shows how common communication sciences research methods can be improved by machine learning.

4.2.2 Enhancing Communication Science Research Designs and Questions

One major incentive for communication science scholars to apply machine learning is the generation of greater, automatically labeled samples. Publication [3] "*Gender-related Differences in Online Comment Sections: Findings from a Large-Scale Content Analysis of Commenting Behavio*r" (Küchler, Stoll, Naab, & Ziegele, 2022) adds to the growing body of work that applies machine learning to upscale research designs and samples in order to answer communication science research questions to incivility. In this study, we investigated whether the gender of Facebook users influenced commenting behavior regarding participation rates and the use of uncivil language targeted at a respective user. Further, we examined if male and female users are confronted with the same amounts of incivility as a reaction to their commenting. In this study, we analyzed a sample of over three hundred thousand user comments of the Facebook pages of 14 news media outlets using supervised learning. To measure the incivility of a comment, we applied a support vector machine-classifier with a combination of word-frequency features and word embedding vectors. The model was trained on an external, labeled sample of 10,114 user comments posted to Facebook sites of the different German news outlets. Thus, we were able to automatically label a great number of documents whose manual annotation would have exceeded the expenses for a comparable research design. As incivility we included impoliteness (Papacharissi, 2004), for example, name-calling and vulgar speech, as well as incivility that violates democratic norms, such as negative stereotyping and the threatening individual's democratic rights.

In this study, the automated classification of incivility enabled the labeling of a great number of user comments, which opened up two major opportunities. First, the large sample allowed us to analyze a higher number of instances, which usually means more stable inference. Second, the extensive data set of classified comments allowed for the application of more appropriate yet demanding statistical approaches. As the comment sections on Facebook are structured in a hierarchical order of initial comments (top-level comments) and reply comments (sub-level comments), we modeled the incivility of a reply comment as a function of the initial comment's incivility and the comment author's gender. However, to meet the requirements of such a multilevel analysis, a high number of instances (comments, reply comments) is required, since the model becomes more complex as a consequence of the superordinate structure (Bates & Pinheiro, 1998). Here, machine learning-based incivility classification was applied to enlarge the sample of labeled comments and achieve the required number instances. Results showed that the comment sections in our sample were dominated by male users who wrote the majority

of uncivil comments. However, women did not receive more uncivil replies to their comments than men. Our analysis confirms that over 12 percent of variance of a reply's incivility is captured in the initial comment. That suggests that characteristics of the initial comment hold predictive information for the incivility of a reply comment, which confirms the usefulness of multilevel modeling in this case.

Publication [4] *"Detecting Impoliteness and Incivility in Online Discussions: Classification Approaches for German User Comments"* (Stoll, Ziegele, & Quiring, 2020) also adds to the body of work that applies machine learning approaches to incivility research in communication science. The article was among the first to be published in *Computational Communication Research*, the first international journal to establish computational methods and *computational communication science* as a distinct research field in communication science (van Atteveldt et al., 2019). In our research article, we developed and evaluated machine learning-based classification approaches to detect different forms of incivility in German-language online discussions on Facebook. We built on the theoretical framework by Papacharissi (2004) who distinguishes impoliteness, meaning offensive but not necessarily harmful content, and "true" incivility, including racism, extremism, stereotyping, discrimination, and threats to democracy. Papacharissi (2004) argues that norm violations such as name-calling, vulgar or pejorative language, violate norms of interpersonal politeness, but do not necessarily threaten democratic discourse. In contrast, "true" incivility undermines democratic values and hence, is more likely to have significant negative consequences for societies as a whole. In publication [4], we investigated the extent to which these two forms of incivility can be identified with machine learning approaches based on different features, including bag-of-words and lexical resources, such as collections of swear words. Therefore, we employed a sample of 10,000 labeled user comments that were annotated using a theory-based coding scheme. As part of an extensive evaluation, we further tested the trained models on an external data set and discussed their usefulness for further, comparable research questions.

Among the approaches tested in this study, feature-based models using the *Naive Bayes* classification algorithm achieved the best performances, ranging from 64 percent accuracy in predicting incivility and 65 percent in predicting impoliteness. Even though more complex deep learning approaches had probably yield better results for this task, results suggest that impoliteness and incivility in user comments can be measured to some extent using feature-based classification approaches. Nonetheless, results show that about 35 percent of comments are classi-

fied incorrectly by the models, even though they were trained on several thousand labeled documents. However, our findings suggest that the conceptual distinction between the two forms of incivility can also be found taking a statistical, data-driven perspective. For example, features that are based on external collections of vulgar expressions and insults mainly improve performance for detecting impoliteness. These findings further indicate that the classification of different forms of incivility may vary, which is crucial for further incivility research that aims to apply comparable machine learning methods.

Overall, this study can be referenced as a case study that provides insight into the performance of different, feature-based machine learning approaches to incivility classification. To answer the question of *how machine learning can extend and elaborate communication science research on incivility,* our study showed how machine learning can be applied to enhance research questions on incivility in communication science. Here, we investigated if the established theoretical distinction of impoliteness and "true" incivility by Papacharissi (2004) also exists on a linguistic level that can be identified by a machine learning-based approach, using a concept-driven selection of features.

4.2.3 Enhancing Communication Science Methods and Tools

In recent years, communication science scholars have become increasingly interested in the automated analysis of incivility. Yet, the majority of studies applied dictionary-based approaches (e.g., Ksiazek et al., 2015; Muddiman & Stroud, 2017). Although current machine learning-based approaches to incivility classification clearly outperform manually created dictionaries, the dictionary approach has the main advantages of a low-threshold application and a comprehensive measurement (Dobbrick et al., 2021; Grimmer et al., 2022; Lind et al., 2019). However, existing incivility dictionaries and comparable resources are not sufficient to map out the multifaceted concept of incivility. Self-created instruments often suffer from a low recall, because manually compiled dictionaries can contain only a limited number of terms. So far, there have been no off-the-shelf dictionaries available for incivility detection for the German language. In research article [5] *"Developing an Incivility-Dictionary for German Online Discussions - A Semi-Automated Approach Combining Human and Artificial Knowledge"* (Stoll, Wilms, & Ziegele, 2023), we offer the first comprehensive German-language dictionary for incivility as well as a novel approach to support the elaborate dictionary creation process by means of machine learning. In our two-step, semi-automated approach, we combined standardized, theory-based content analysis with unsupervised learning to support the costly manual

creation of an incivility dictionary. Based on an extensive data set of labeled user comments and Tweets, we manually annotated an initial list of expressions that later was enhanced with automatically retrieved expressions from a pre-trained word embedding model (see chapter 3.1.1). This way, the collection of uncivil expressions could be extended to over 7,000 unigrams (e.g., single words, emojis). During an extensive evaluation, we found that the large number of collected entries led to a high recall of 76% for incivility retrieval on the test set, which was significantly higher than comparable approaches and several machine learning approaches to incivility classification, including bag-of-words and BERT-based approaches. Although the state-of-the-art BERT classifier overall outperformed the dictionary approach (macro-average F1 score= 0.65, accuracy= 0.65), the performance of our incivility dictionary was significantly higher than comparable dictionaries (macro-average F1 score between 0.55 and 0.61, accuracy between 0.55 and 0.62). We further assumed that due to the semi-automated process, the resulting dictionary will be less dependent on the perception of the individual coders. With the help of the unsupervised model the annotations of the coders are supplemented by associated words, which neither have to be part of the sample, nor have to reflect the perception of the coders. This "artificial perspective" is a further benefit of our approach, since the subjectiveness of incivility is a major issue for its measurement (Herbst, 2010; Chen et al., 2019, see chapter 2.3.3 and 3.3).

In sum, publication [5] gives an example on how machine learning can enrich established communication science research methods. In our approach, we showed how pre-trained word embeddings can enhance the sophisticated manual process of creating a dictionary. Here, machine learning had not only the benefit of saving costs, but could also tackle a major challenge of incivility measurement, namely its subjective and hence, the risk of a one-sided perception. Besides the developed approach, we present a novel instrument that is, to the best of our knowledge, the first sufficient dictionary to analyze incivility in German-language online discussions. The instruments as well as all source-code is publicly available for further research. In addition to the dictionary, we are also providing a free web application for drag-and-drop data analysis. Finally, due to its comparative structure, our study provides an overview of machine learning methods compared to dictionaries for incivility detection. Thus, findings may help researchers to weigh the costs and benefits of different computational approaches for their individual use case.

4.3 Change of Perspective - Methodological Considerations

The distance between content analysis in communication science and the labeling of training data for machine learning in computer science is small. Based on the data sets and data structure, it seems reasonable that a sample of user comments, manually coded in the scope of a quantitative content analysis, is also suitable as training data for supervised learning. Yet, natural language is a particularly heterogeneous data type and patterns often emerge only in huge samples, which go beyond the usual scale in communication science research designs. However, the biggest challenge for detecting incivility with automated approaches is the conceptualization of incivility itself. In the following sections, I will address the question of *what methodological challenges communication scholars face with machine learning-based incivility classification and how they can be overcome* and present the contributions of the research articles [6] and [7] in this regard.

4.3.1 Methodological Challenges of Incivility Classification

Incivility is a challenging concept to measure, not only with manual but also with machine learning-based methods. Publication [6] "*The Accuracy Trap or How to Build a Phony Classifier*" (Stoll, 2023) discusses how incivility as a theoretical concept affects the results of machine learning-based classifiers and explain, what pitfalls arise from the specific characteristics of this concept. In contrast to, for example, topics, incivility cannot be measured exclusively based on the text of a comment, since significant information to determine what is uncivil lies within the situational context or the personal perception of the recipient (Herbst, 2010; Ross et al., 2017; see also chapter 2.3). Nevertheless, machine learning approaches to incivility classification usually exclusively consider text-based features. However, if the meaning of words varies in different contexts, a classification model will fail to learn a robust relationship between the features and the meaning of a document and thus, results can become inaccurate. In addition to the lack of contextual information, another characteristic of incivility contributes to this uncertainty, which is the distribution of uncivil in relation to non-uncivil content. In a random sample of comments to online discussions, incivility rarely occurs compared to not uncivil comments (e.g., Coe et al., 2014; Davidson et al., 2017; Friess et al., 2021; Papacharissi, 2004; Zampieri et al., 2019). This *imbalanced* distribution is challenging because of two major reasons. First, a small number of instances makes it difficult to map a stable, statistical relation between text (e.g., words) and class affiliation (e.g., incivility). As a consequence, random samples of online discussions often lack relevant information for the detection of incivility. Second,

the resulting imbalanced distribution of classes can lead to an overestimation of the major class (here: not uncivil). This is primarily a problem for classification functions such as logistic regression, support vector machines, or decision trees, which tend to predict the major class if information is missing (Denil & Trappenberg, 2010; Haixiang et al., 2017). In combination, these two conditions can result not only in inaccurate predictions, but also in *biased* results. In the case of incivility classification, such bias is that incivility is significantly underestimated by a classification model. Furthermore, such bias can easily remain undetected, if performance is not reported by category (e.g., by reporting recall and precision) but only with accuracy (see chapter 3.3). Since the accuracy does not distinguish agreement by category, a classifier can achieve high performance by predicting the frequent class only (not uncivil) and without detecting the relevant class (incivility) at all. In other words, classifiers will not only systematically underestimate the prevalence of incivility in the data, but the overall accuracy will suggest a high performing model.

Although there is much overlap between the approaches of content analysis and document classification, article [6] draws attention to pitfalls that arise at the interface of communication science incivility research and machine learning. So far, communication science has not really been concerned with the data type text for predictive modeling and usually, several hundreds to thousands of instances have sufficient to model statistical relationships and draw inference from a sample. However, due to the very heterogeneous data type text, huge samples are needed. This is particularly true for ambiguous concepts, such as incivility. Further, the distribution of incivility is likely to distort the evaluation of performance, if detailed metrics are not applied. This can happen easily, since the respective metrics are not commonly used in communication science research.

4.3.2 Aligning Incivility Distributions Using Oversampling

Publication [7] "*Supervised Machine Learning mit Nutzergenerierten Inhalten: Oversampling für nicht balancierte Trainingsdaten [Supervised machine learning with user generated content: oversampling for imbalanced training data]*" (Stoll, 2020) responds to the challenges of incivility classification discussed in article [6] and presents one possible method to address them. The article examined whether and, if so, to what extent different *oversampling* techniques are suitable to address the issues of rare instances of incivility and resulting imbalanced class distributions. To this end, I applied the two different oversampling techniques *Random Oversampling* (*ROS*) and *Synthetic Minority Over-sampling Technique* (*SMOTE*).

These techniques were used to align class distributions in three different data sets of Tweets and Facebook user comments (n= 55,400) for three different binary classification tasks, including offensive language, incivility, and sentiment classification. While ROS weights random cases of the underrepresented class higher, SMOTE generates synthetic vectors similar to vectors of uncivil comments in the training data, using the k-nearest neighbors' algorithm (Chawla et al., 2002). In machine learning, oversampling has been successfully applied in several research areas, for example, in biotechnology and finance for the diagnosis of diseases, the identification of genes (e.g., Dubey et al., 2014; Herndon & Caragea, 2016), or to detect credit card fraud (e.g., Zakaryazad & Duman, 2016). For text classification, especially in the use case of incivility, the potentials of oversampling have not been investigated yet. To predict the outcoming classes, I applied a bag-of-words classifier, as a common baseline approach. Results show that both oversampling approaches lead to an overall improvement in prediction. Due to the alignment of class distributions, the prediction function was less prone to predict the majority class. Furthermore, oversampling gives higher weight to vocabulary that might be crucial for class membership and would otherwise have been neglected. However, because oversampling does not add new vocabulary to the training data, a classifier will not make more informed decisions. Moreover, as the weighting is random, also ambiguous vocabulary is oversampled, which can lead to a decrease of overall performance. This means that overall, oversampling is most of all useful to tackle one major problem in incivility classification, which is the biased prediction towards the overrepresented class.

In summary, article [6] presents and discusses major challenges of incivility classification using machine learning, which primarily occur because of the distribution of incivility and the ambiguous nature of the concept. In article [7], I have shown that these challenges cannot be eliminated but reduced by applying oversampling.

4.4 Conclusion and Takeaways

In line with the general research question of this thesis, *how to measure incivility with machine learning methods,* the seven articles of the cumulus contribute to different aspects of this question. Articles [1] and [2] present significant benchmarks for the development of state-of-the-art approaches to classify incivility in German-language online discussions. Article [1] was among the first to apply and evaluate a German BERT model to the downstream task of incivility classification and hence, marked a significant step towards the use of LLMs for inci-

vility detection and related tasks. With GermEval2021, Article [2] offered both a public benchmark data set and a comparative format for classifying incivility in German. The shared task further introduced the two novel categories "engaging" and "fact-claiming", which, besides incivility, play a significant role in current demands on online moderation. Articles [3], [4], and [5] contributed to the current demand for use cases, best practices, and methodological work in communication science, and incivility research in particular, that embed machine learning at different points of the research process. The articles showed how research designs, questions, and methods can be enhanced by means of machine learning. These enhancements include the enlargement of samples and the enabling of more demanding analysis methods, the examination and review of theoretical concepts of incivility from a data-driven perspective, and the improvement of established analysis methods, here dictionaries, in terms of saving costs and improving measurement. Finally, articles [6] and [7] address important methodological challenges that occur on the interface of quantitative content analysis and machine learning-based document classification of incivility. While these approaches overlap, the characteristics of incivility pose specific challenges for machine learning approaches to incivility detection, which go beyond issues of manual content analysis of incivility. Publications [6] and [7] discussed these challenges, addressed resulting pitfalls, and proposed strategies to avoid and manage them. The provided insight may help communication science scholars preventing prediction bias in incivility classification and overall, contribute to a comprehensive understanding of capabilities and limits of machine learning to automated incivility detection.

5 Discussion

5.1 Summary of Contributions

This cumulative dissertation addresses the overall research question of *how incivility can be measured with machine learning methods* (chapter 1.1 and 1.2). The seven research articles of the cumulus contribute to different aspects of this question, which can be structured into the three subordinated questions below. In addressing these questions, this dissertation provides essential steps of an interdisciplinary, methodological transfer from computer science-driven machine learning and artificial intelligence research into communication science research on incivility.

1) *To what extent can incivility be measured with machine learning methods?* To answer this research question, this dissertation contributes to the ongoing development of state-of-the-art machine learning approaches to the specific task of incivility classification in German-language online discussions. To this end, article [1] and [2] provide several benchmarks. In article [1], we were among the first to apply and evaluate a pre-trained German BERT model to incivility classification. To this day, incivility classification in both research and industry is dominated by BERT-based approaches. Article [2] presents the shared task GermEval 2021 on the identification of toxic, engaging, and fact-claiming comments. Within this format, we provide a freely available data set for incivility classification and two additional categories, which play important roles for online moderation. In sum, both articles provide valuable insights into the opportunities and current capabilities of machine learning methods for incivility classification. Findings suggests that although current LLM demonstrate a significant increase of performance for incivility classification, the subjectivity of incivility as well as the question of resources will continue to pose challenges for this area of research.
2) To answer the second subordinate research question of *how machine learning can extend and elaborate communication science research on incivility,* this dissertation presents methodological work, use cases, and best practice studies that show how machine learning methods can be integrated fruitfully on different points of the communication science research process. Article [3] shows how supervised learning can enhance communication science research designs by enabling larger samples and hence,

more elaborate analysis methods. Article [4] examines how machine learning can enable conceptual research on incivility by deriving inference from feature-based approaches about linguistic characteristics of different forms of incivility. Finally, article [5] presents a novel approach that uses unsupervised learning to extend the costly and restricted manual creation of dictionaries.

3) To answer the third subordinate research question of *what methodological challenges communication scholars face with machine learning-based incivility classification and how they can be overcome,* articles [6] and [7] contribute methodological guidance on how to apply and evaluate machine learning for incivility classification. The articles discuss specific issues that come along with the subjective nature and prevalence of incivility, which can hinder the development of functional and unbiased machine learning models, and further, provide approaches to address them.

5.2 Automated Incivility Detection - Does it Work?

The continuous, rapid progress of artificial intelligence noticeably impacts many areas of research and application. Among them, one major field is NLP. Meanwhile, the identification of concepts related to incivility including hate speech, toxicity, and offensive language, states an important NLP task, usually approached as document classification and thus, as a supervised learning problem. To measure the performance of supervised learning approaches, the predicted values of a model are aligned with manually assigned labels on a given data set. The higher the agreement between predicted values and the manually labeled gold standard is, the higher the model performance is evaluated. In recent years, milestones in artificial intelligence research have led to a constant performance increase for many NLP problems. A recent major step was the introduction of the neural network architecture *transformer*, presented by Vaswani and colleagues in 2017, and the resulting establishment of LLMs such as BERT (Devlin, et al., 2018) and GPT (Redford, et al., 2018). For text classification tasks, BERT-based approaches have been proven to be particularly suitable. Over the last five years, several standalone BERT models have been released for several languages (e.g., Antoun et al., 2020; Lee et al., 2020; Martin et al., 2019) and even specific use cases (e.g., *AngryBERT* by Awal et al., 2021). In the scope of the GermEval 2019 shared task on offensive language classification (Struß et al., 2019), article [1] of this dissertation has been among the first to present a BERT-based approach for incivility classification in German, thus marking a significant step towards the use of LLMs for automated incivility detection. Over the last years, several international studies referred to the

approach presented in article [1] when applying BERT models in German or other languages to related tasks (e.g., Bosco et al., 2023; Garcia-Diaz et al., 2022; Rajalakshmi et al., 2023). Up to now, the task of incivility classification is still dominated by BERT-based approaches (Satapara et al., 2022; Zampieri et al., 2023). However, performances of different models and approaches can vary. This is because the individual models are pre-trained on different data sets and with varying parameter settings, and are then fine-tuned and evaluated on different data sets. Article [1] fine-tuned a freely available German BERT model, which was pre-trained on 12 GB of openly available data, including Wikipedia articles and a share of the Open Legal database, which includes transcripts of judgments and laws (deepset, 2019). As this data basis does not include online discussion data exclusively, it is likely that the performance for incivility classification is limited and would increase if the training data of the model fitted the research subject completely. Besides the compilation of the data set a model is trained on, the parameter setting both while training and fine-tuning will affect the performance on the test set. The winning team of the shared task GermEval 2021 presented in article [2] of this thesis achieved a 0.72 macro-average F1 score with an ensemble approach of several pre-trained language models. They further found that with increasing size of the ensemble, performance increased significantly (Bornheim et al., 2021). Even though basic setups for fine-tuning LLMs already require high performing GPUs, an increase of computing resources enables even more demanding parameter setting, which is likely to increase overall performances. These results suggest that the performance of current and future state-of-the-art machine learning will not only depend on the availability of data, but also on resources for high-performance computing.

As artificial intelligence is a constantly evolving research area, the answer to the question *to what extent incivility can be measured with machine learning methods* can only be a snapshot. Even though model performances vary depending on the specific data set and the compilation of the individual models, benchmarking tasks and data sets are able to create a coherent impression of the capabilities of current approaches. At this moment, the state of research on the task of incivility classification is dominated by BERT-based approaches, including its derivatives such as RoBERTa (Liu et al., 2019), and lies currently between 0.70 and 0.80 macro F1 score for different languages, including German (Kumaresan et al., 2021; Zampieri et al., 2023). This cumulated picture allows the conclusion that incivility is measurable to a certain extent with machine learning-based methods. Yet, in the sense of supervised learning, this measurability is defined through the level of agreement between predicted and manually

created gold standard values on a specific data set (see chapter 3.3). Both empirical and conceptual work on incivility as well as latest research in machine learning suggests, however, that the idea of one common gold standard measure contradicts the subjective nature of concepts such as incivility (e.g., Alm, 2011; Basile et al., 2021; Chen et al., 2019; Romberg, 2022; Ross et al., 2017). The variations in the perception of incivility are not considered in the idea of standardized measurement that constitutes the basis for both content analysis and supervised machine learning (Bormann et al., 2022; Chen at al., 2019; Früh, 2015). In quantitative content analysis, coders are trained to have a shared understanding of meaning, based on theoretically grounded categorizations and extensive coding instructions and training (Früh, 2015; Rössler, 2017). In machine learning, different measurements of annotators are converted into one common value by averaging or majority voting (Basile et al., 2021; Romberg, 2022). Although these two approaches handle deviating perception differently, they both aim for one standardized, final measurement. Recent ideas from the field of machine learning argue that multi-perceptual concepts such as incivility are not compatible with the idea of one unique gold standard and hence, with established coding procedures (e.g., Akhtar et al., 2019; Akhtar et al., 2020; Basile et al., 2021; Cabitza et al., 2019; Romberg, 2022). Researchers that support the idea of *data perspectivism* demand not to disregard opinions, as they belong to a minority in many cases. Instead, they propose to take advantage of information that is usually considered as noise, for example, by modeling human error patterns or the uncertainty of prediction (e.g., Basile et al., 2021). In this sense, questions such as who a model learns from and what validity defines increasingly influence the discourse on artificial intelligence (Gunning et al., 2019). Future research on incivility should take these rather novel considerations into account when it comes to the question of valid measurement, both for computational and manual content analysis.

5.3 Machine Learning-Based Moderation - Improvement of Discourse?

From a normative perspective, incivility is often considered a threat to democratic discussions as it is assumed to hinder deliberative discourse or cooperation between participants (e.g., Andersson et al., 2014; Bormann et al., 2022; Brooks & Geer, 2007; Coe et al., 2014; Papacharissi, 2004; Rowe, 2015; Ziegele, Jost et al., 2018). In practice, the growing number of user contributions challenge moderators and platform operators whose daily business it is to encounter incivility in comment sections, which is also required by law (BMJ, 2022; European Commission, 2023; Heinbach & Wilms, 2022; Kümpel & Rieger, 2019; Wright, 2006; Ziegele & Jost, 2020). It seems plausible that algorithmic moderation systems could be a significant

relief for moderators and community managers by supporting the detection of uncivil content and thus, foster a more civil discussion environment online (Gorwa et al., 2020; Wilms et al., 2023). Nonetheless, this idea faces several challenges. Studies showes that various personal and contextual factors can affect what moderators of online discussions, journalists, and users perceive as uncivil or not (e.g., Boberg et al., 2018; Bormann et al., 2022; Bormann et al., 2023; Chen & Pain, 2017; Diakopoulos & Naaman, 2011; Paasch-Colberg & Strippel, 2022). Moreover, not all forms of incivility are equally easy to identify. Studies showed that swearing and insults are easier to detect, while implicit forms of incivility often remain unrecognized (e.g., Boberg et al., 2018; Muddiman & Stroud, 2017). Article [5] of this thesis developed an incivility dictionary using pre-trained word embeddings that includes words, emojis, and hashtags to detect uncivil content in German-language online discussions. We could show that our method is promising to retrieve a variety of uncivil expressions in German-language online discussions. Further, the employment of an "artificial perspective" has the potential to address bias due to individual perceptions of coders and researchers. Yet, we found that implicit forms of incivility remain undiscovered by the dictionary, as many comments do not include unambiguously uncivil expressions. Machine learning-based approaches are also confronted with this challenge. Article [6] argued that unambiguous text features (e.g., word distributions), meaning words that occur both in uncivil and not uncivil comments, hinder a classifier from learning a robust relationship between features and class membership. At the same time, unambiguous words such as insults are highly predictive features for incivility classification. Muddiman and Stroud (2017) argue that implicit uncivil comments can complicate the justification for flagging or deleting a comment. Against the backdrop of freedom of speech, this justification can become important in the context of online discussions. Also for crowd workers more obvious forms of incivility might be easier to identify (Binns et al., 2017; Ross et al, 2017). Yet, research suggests that violations of democratic norms, such as racism or stereotyping, lies, or defamation can occur in a rather implicit way, but are considered equally or even more harmful (e.g., Bormann, 2022; Bormann et al., 2023; Muddiman, 2019; Papacharissi, 2004). As a consequence, algorithm-based moderation systems are likely to be biased towards explicit, unambiguous forms of incivility. Yet, Papacharissi (2004) already considered impoliteness such as swearing and vulgarity as rather harmless or even fruitful for heated, but democratic discussions. This view is shared by Chen and colleagues (2019) who argue that a *sanitized* discussion space would fail to acknowledge the value of imperfect speech.

The potential bias in detecting incivility resulting from individual perception, situational context, and ambiguity also touches the issue of fairness of artificial intelligence (Gunning et al., 2019). Meanwhile, numerous studies showed that several machine learning-based systems suffer from bias on several levels that may result in discrimination (Ferrer et al., 2021). Currently, automated systems for online moderation supported by artificial intelligence learn on previous decisions of moderators or reactions of users, such as flagging (Beuting, 2021; Vox Media, n.d.). If these data sets include, for example, prejudices or a political leaning of the moderator, an algorithm is likely to reproduce these bias in the future (Danks & London, 2017; Ferrer et al., 2021; Hovy & Prabhumoye, 2021). Bias not only plays a role in this field-generated data, but also in pre-trained LLMs (see chapter 3.1.3). Research suggests that available models including BERT, RoBERTa, and GPT suffer from stereotypical bias (Nadeem et al., 2020). In a fine-tuning pipeline, this information will factor into the downstream classification task and hence, influence the decisions of the final model. As these models are much more complex than linear, feature-based approaches, such bias is even more challenging to identify (Gunning et al., 2019). However, in the context of online participation, it is crucial that moderators are able to understand and justify decisions about deleting comments or blocking users to maintain free speech without censorship (Wilms et al., 2023; Wright, 2006).

Overall, at this point, algorithm-based moderation has both the potential to improve and to weaken online discussions in terms of democratic discourse. In general, a relief for moderators and community managers is conceivable as algorithms could support the selection of incoming user comments. However, these forms of incivility may include mostly explicit and obvious forms of incivility, such as unambiguous insults and vulgarity, while implicit and often even more harmful forms of incivility could remain undiscovered. Further, artificial intelligence that is trained by moderators or crowd workers can include a limited and hence, biased understanding of incivility due to individual factors (Herbst, 2010). If, for example, minorities or certain target groups of incivility are not part of the moderation team or are underrepresented on crowd working platforms, machine learning models would reproduce a one-sided view of incivility. As a consequence, incivility against minor groups or incivility that is less straightforward to detect could remain in the discourse and might hinder an equal participation (Springer et al., 2015; Stroud et al., 2016; Ziegele, Jost et al., 2018). As long as the reference for performance for supervised approaches is the approximation to one target, gold standard value, approaches have the potential to even aggravate the risk of discrimination in online discussions as they tend to amplify biases in human decisions. To foster democratic discourse, the

current issue for automated incivility detection is the validity of measurement and the associated question, how validity should to be defined when it comes to incivility.

5.4 Machine Learning Research - An Interdisciplinary One-Way Street?

With the establishment of computational communication science as a genuine research field, machine learning is becoming more visible in communication science (Domahidi et al., 2019; Lazer et al., 2009; van Atteveldt et al., 2019; van Atteveldt & Peng, 2018). However, so far, the majority of research approaches machine learning from a pragmatic perspective. Most of all for content analysis, machine learning-based methods are promised to save costs and therefore, enable the use of larger samples and broader research designs. In article [3], we classified a sample of more than three hundred thousand user comments using supervised learning to analyze incivility in hierarchical comment sections on Facebook. Here, machine learning enabled us to consider more comments and further, to apply multilevel modeling for data analysis, which is more adequate but demanding regarding sample size. In this role, functionality of machine learning methods should ideally be presupposed. Yet, increasingly advanced and powerful machine learning methods are often challenging to adapt, and their application and evaluation still comes with a degree of uncertainty. To this day, computational methods and machine learning specifically are represented only sporadically in the curriculum of (German) communication scholars (Niemann-Lenz et al., 2019; Scheper & Kathirgamalingam, 2022). However, the broad establishment of machine learning in communication science would require the amplification of the communication science skill set. Van Atteveldt and Peng (2018) argue that as computational methods will become increasingly important in the future, communication science needs to invest in respective skills and infrastructure in their own discipline now. These skills do not necessarily include a complete spectrum of computer science methods but need to maintain a connection point to machine learning and artificial intelligence research. Traditionally, these research fields are dominated by computer science and related disciplines, and hence, research traditions and interests naturally vary from social sciences. Consequently, also an understanding of research questions and approaches in computer science is a key requirement for interdisciplinary research teams as well as for reasonable, successful research endeavors in the field of machine learning. Article [6] adds to the body of research that can bridge the gap between methods and research traditions of the different disciplines for the re-

search subject of incivility. It can be shown that besides pitfalls, a common ground exists between machine learning and quantitative content analysis, which can provide a potential interface for interdisciplinary work.

Recent developments in artificial intelligence have revealed that data and computational resources have become more crucial than ever, and further, that these resources are not equally distributed (Crawford, 2021). Meanwhile, only a small number of institutions can compete in the ongoing progress of high-performing artificial intelligence, which mainly include large tech-companies such as Google, Amazon, Meta, and IBM. In this competition, academic research faces the challenge of providing a counterweight to these commercial actors (Hilbert et al., 2019). How can communication science take a substantive role in this interdisciplinary field of artificial intelligence research? Lately, progress in artificial intelligence is not restricted to issues of optimization and mathematical understanding but increasingly raises ethical questions (Crawford, 2021), including data quality and transparency, data privacy issues, and fairness of machine learning models and applications (Mehrabi et al., 2021). Different disciplines address these questions from several perspectives. One research area is *algorithm auditing,* which is concerned with the assessment of algorithms from an ethical perspective by collecting behavioral data about artificial intelligence systems in a certain context (Brown et al., 2021). Also in communication science, a growing number of studies have dealt with biases and discrimination of artificial intelligence systems, for example, regarding news recommendation systems (Bandy & Diakopoulos, 2020), political bias (Puschmann, 2019), and digital divide in search engines (Scherr et al., 2022). In contrast to this rather observative view, the field of *Explainable AI* (*XAI)*, mainly driven by computer sciences, deals with the question of how to create artificial intelligence that are more understandable for humans from a technical perspective (Gunning et al., 2019). Against this backdrop, transparency of data and decision making of machine learning models are relevant issues. Here, communication scholars can contribute a methodological perspective on data quality as well as knowledge about genuine research subjects, including journalistic and social media communication. Article [2] built on communication science research and expertise on online discussions and online moderation to conceptualize and introduce meaningful concepts for machine learning applications. Additionally, experience from content analysis were incorporated into data quality assurance and supported the provision of a high-quality data set. Further, conceptual and empirical knowledge of a certain research subject can foster a holistic evaluation of machine learning performance. Articles [6] and [7] of this thesis

showed that mathematical optimization can be accompanied with biased results that are, however, easy to overlook within the process of increasing performance. To this end, interpreting and critically evaluating results is an important counterpart and reassurance of optimization to identify possible prediction bias of a model. Also in article [5], we supplemented the quantitative performance evaluation of the developed incivility dictionary with a qualitative, in-depth error analysis. This way, we could recognize a potential bias toward explicit forms of incivility and topic-specific incivility due to the restricted time span of the underlying data set. This pattern would not have been detected by the quantitative evaluation on the test set, even though this information is crucial for the future use of the instrument.

When it comes to machine learning and artificial intelligence research, communication science currently is in the phase of adoption and learning. However, future work must define a new role for communication science that goes beyond asking meaningful communication science research questions for machine learning methods, both to keep interdisciplinary work a two-way-street and to establish communication science in the interdisciplinary field of artificial intelligence. Here, potential lies in the assurance of data quality and in research on fairness and transparency. Besides a strong background in content analysis, expertise further lies in knowledge about certain areas of applications, such as online discussions or journalism. This perspective can help to make sense of machine learning results (article [6]), to indicate bias of existing systems in a certain application area (e.g., news recommendation systems; Bandy & Diakopoulos, 2020), or to integrate different perspectives, for example, those of moderators and users of online discussions, to support useful and proper applications (Wilms et al., 2023).

6 Conclusion

The automated identification of uncivil comments is a current and widely relevant issue that concerns communication science, machine learning and artificial intelligence research, as well as platforms and practitioners in the field of online discussions and moderation. With the growing participation online, incivility has been regarded as a significant challenge for democratic discourse, both from a normative and legal point of view. However, incivility is a highly subjective and ambiguous concept. In communication science research, the measurement of incivility has rather been considered a methodological and conceptual issue that primarily complicates a consistent operationalization and hence, the comparison of empirical studies. This is no surprise, since standardized approaches to content analysis are not designed to consider variance between measurements, such as those due to individual perceptions. Instead, the reliability of measurement constitutes a key quality criterion, which is achieved when a construct is perceived and measured equally among coders and researchers. However, research on incivility suggests that the perception of what incivility is varies significantly along individual background or the professional role a person holds. Further, empirical research on incivility points towards the fact that primarily unambiguous forms of incivility, such as insults or vulgarity, are straightforward to detect. This predicament between conceptualization and analysis method also exists in machine learning, where a model is designed to learn and reproduce human decisions based on training data. The performance of a model is then calculated by the number of correct reproductions of these decisions. Following this approach, machine learning runs the risk of reproducing bias and discrimination that is already present in the manually coded training data. This risk is particularly relevant for practical applications, where machine learning-based moderation systems aim to support human moderators in deciding which contributions violate platform or legal guidelines. Here, the reproduction of incivility that only reflects limited views could even lead to an amplification of incivility in online discussions. As long as this challenge of bias in machine learning has not been solved, a broad application of respective systems remains questionable.

The increasingly negotiated questions of fairness and meaningfulness of machine learning approaches and applications can benefit from the social science perspective. In the development and conception of fair artificial intelligence, however, communication science is holding back. So far, the growing and popular field of computational communication science mainly deals with machine learning regarding its application and further development for communica-

tion science research questions and subjects. Against this backdrop, machine learning is primarily seen as a resource to save costs in data analysis, for example, by enhancing traditional content analysis research designs. However, discrimination risks already arise during the development and conceptualization of machine learning-based systems. Through the research field of online communication, communication science could bring important expertise to the development of machine learning-based systems in the context of online discussions and also in online journalism, e.g., for news recommendation systems. This expertise can include the analysis of biases in training data of journalistic and user-generated content, as well as a holistic evaluation of machine learning-based systems by recipients, journalists, or community managers. To take this critical view, the mode mere adaptation is not sustainable. Instead, communication science must establish a role that contributes expertise and perspective that is able to enhance genuine machine learning research. To ask meaningful questions, communication science scholars need a proper understanding of current approaches to specific research subjects such as incivility, that allows the linkage of methodological, empirical, and conceptual knowledge. This dissertation aims to make a contribution to this linkage of machine learning and communication science research on incivility.

References

Akhtar, S., Basile, V., Patti, V. (2019). A New Measure of Polarization in the Annotation of Hate Speech. In M. Alviano, G. Greco, & F. Scarcello (Eds.), *AI*IA 2019 – Advances in Artificial Intelligence: Vol 11946* (pp. 588–603). Springer. https://doi.org/10.1007/978-3-030-35166-3_4

Akhtar, S., Basile, V., & Patti, V. (2020). Modeling annotator perspective and polarized opinions to improve hate speech detection. In L. Aroyo & E. Simperl (Eds.), *Proceedings of the Tenth AAAI Conference on Human Computation and Crowdsourcing: Volume 8* (pp. 151-154). The AAAI Press. https://doi.org/10.1609/hcomp.v8i1.7473

Albadi, N., Kurdi, M., & Mishra, S. (2018). Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In U. Brandes, C. Reddy, & A. Tagarelli (Eds.), *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 69-76). IEEE Press. https://doi.org/10.1109/ASONAM.2018.8508247

Alm, C. O. (2011). Subjective natural language problems: Motivations, applications, characterizations, and implications. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*: Human Language Technologies (pp. 107-112). Association for Computational Linguistics.

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The "nasty effect:" Online incivility and risk perceptions of emerging technologies. *Journal of computer-mediated communication*, *19*(3), 373-387. https://doi.org/10.1111/jcc4.12009

Andrich, A., & Domahidi, E. (2022). A Leader and a Lady? A Computational Approach to Detection of Political Gender Stereotypes in Facebook User Comments. *International Journal of Communication*, *17*, 236–255.

Andrich, A., Bachl, M., & Domahidi, E. (2023). Goodbye, Gender Stereotypes? Trait Attributions to Politicians in 11 Years of News Coverage. *Journalism & Mass Communication Quarterly*, *100*(3). https://doi.org/10.1177/10776990221142248

Antoun, W., Baly, F., & Hajj, H. (2020). *Arabert: Transformer-based model for arabic language understanding.* arXiv. https://doi.org/10.48550/arXiv.2003.00104

Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine, 36*(1), 15-24. https://doi.org/10.1609/aimag.v36i1.2564

Awal, M.R., Cao, R., Lee, R.KW., Mitrović, S. (2021). AngryBERT: Joint Learning Target and Emotion for Hate Speech Detection. In K. Karlapalem et al. (Eds.). *Advances in Knowledge Discovery and Data Mining. Proceedings of the 25th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD 2021). Lecture Notes in Computer Science: Volume 12712* (pp. 701–713). Springer. https://doi.org/10.1007/978-3-030-75762-5_55

Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., ... & Goldblum, M. (2023). *A cookbook of self-supervised learning*. arXiv. https://doi.org/10.48550/arXiv.2304.12210

Bandy, J., & Diakopoulos, N. (2020, May). Auditing news curation systems: A case study examining algorithmic and editorial logic in Apple News. In M. D. Choudhury (Ed.), *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020): Volume 14* (pp. 36-47). Association for the Advancement of Artificial Intelligence.

Basile, V., Cabitza, F., Campagner, A., & Fell, M. (2021). *Toward a perspectivist turn in ground truthing for predictive computing*. arXiv. https://doi.org/10.48550/arXiv.2109.04270

Bates, D. M., & Pinheiro, J. C. (1998). *Computational Methods for Multilevel Modelling*. ResearchGate. https://www.researchgate.net/profile/Douglas-Bates/publication/2753537_Computational_Methods_for_Multilevel_Modelling/links/00b4953b4108d73427000000/Computational-Methods-for-Multilevel-Modelling.pdf

Benikova, D., Biemann, C., Kisselew, M., & Pado, S. (2014). Germeval 2014 named entity recognition shared task: companion paper. In G. Faaß & J. Ruppenhofer (Eds.), *Workshop Proceedings of the 12th edition of the KONVENS conference* (pp. 104-112). Universitätsverlag Hildesheim.

Bergström, A., & Wadbring, I. (2015). Beneficial yet crappy: Journalists and audiences on obstacles and opportunities in reader comments. *European Journal of Communication*, *30*(2), 137-151. https://doi.org/10.1177/0267323114559378

Beuting, S. (2021, September 9). *Wenn Künstliche Intelligenz das Forum moderiert [When Artificial Intelligence Moderates the Forum]*. Deutschlandfunk. https://www.deutschlandfunk.de/sehr-wahrscheinlich-hass-wenn-kuenstliche-intelligenzdas.

2907.de.html?dram:article_id=502849

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In G. L. Ciampaglia, A. Mashhadi, & T. Yasseri (Eds.). *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II 9* (pp. 405-415). Springer. https://doi.org/10.1007/978-3-319-67256-4_32

Bishop, C. M. (2006). *Pattern recognition and machine learning* (1st ed.). Springer.

Bisht, A., Singh, A., Bhadauria, H.S., Virmani, J., Kriti (2020). Detection of Hate Speech and Offensive Language in Twitter Data Using LSTM Model. In S. Jain & S. Paul (Eds.). *Recent Trends in Image and Signal Processing in Computer Vision. Advances in Intelligent Systems and Computing* (pp. 243–264). Springer Singapore. https://doi.org/10.1007/978-981-15-2740-1_17

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(1), 993-1022.

BMJ. (2022, December 21). *Regeln gegen Hass im Netz – das Netzwerkdurchsetzungsgesetz [Rules against hate on the net - the Network Enforcement Act].* https://www.bmj.de/DE/Themen/FokusThemen/NetzDG/NetzDG_node.html

Boatright, R. G. (2019). A crisis of civility? In R. G. Boatright, T. Shaffer, S. Sobieraj, & D. Goldthwaite Young (Eds.), *A crisis of civility? Political discourse and its discontents* (pp. 1–6). Routledge.

Boberg, S., Quandt, T., Schatto-Eckrodt, T., & Frischlich, L. (2020). *Pandemic populism: Facebook pages of alternative news media and the corona crisis--A computational content analysis.* ArXiv. https://doi.org/10.48550/arXiv.2004.02566

Boberg, S., Schatto-Eckrodt, T., Frischlich, L., & Quandt, T. (2018). The moral gatekeeper? Moderation and deletion of user-generated content in a leading news forum. *Media and Communication, 6*(4), 58-69. https://doi.org/10.17645/mac.v6i4.1493

Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018). A dataset of Hindi-English code-mixed social media text for hate speech detection. In M. Nissim, V. Patti, B. Plank, & C. Wagner (Eds.), *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media* (pp. 36-41). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-1105

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. In L. Lee, M. Johnson, & K. Toutanova (Eds.). *Transactions of*

the association for computational linguistics (pp. 135-146). Association for Computational Linguistics. https://doi.org/10.1162/tacl_a_00051

Bormann, M. (2022). Perceptions and evaluations of incivility in public online discussions–Insights from focus groups with different online actors. *Frontiers in Political Science, 4*. https://doi.org/10.3389/fpos.2022.812145

Bormann, M., Heinbach, D., Kluck, J., & Ziegele, M. (2023, May 25-29). *Perceptions of and reactions to different types of incivility in public online discussions: Results of an online experiment* [Conference presentation abstract]. 73th Annual Conference of the International Communication Association (ICA), Toronto, Canada.

Bormann, M., Tranow, U., Vowe, G., & Ziegele, M. (2022). Incivility as a violation of communication norms—A typology based on normative expectations toward political communication. *Communication Theory, 32*(3), 332-362. https://doi.org/10.1093/ct/qtab018

Bormann, M. & Ziegele, M. (2023). Incivility. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 199-217). Digital Communication Research. https://doi.org/10.48541/dcr.v12.12

Bornheim, T., Grieger, N., & Bialonski, S. (2021). FHAC at GermEval 2021: Identifying German toxic, engaging, and fact-claiming comments with ensemble learning. In J. Risch, A. Stoll, L. Wilms, & M. Wiegand (Eds.), *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments* (pp. 105–111). Association for Computational Linguistics.

Bosco, C., Patti, V., Frenda, S., Cignarella, A. T., Paciello, M., & D'Errico, F. (2023). Detecting racial stereotypes: An Italian social media corpus where psychology meets NLP. *Information Processing & Management*, *60*(1), 103118. https://doi.org/10.1016/j.ipm.2022.103118

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory, USA,* 144-152. https://doi.org/10.1145/130385.130401

Boumans, J. W., & Trilling, D. (2018). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. In J. W. Boumans & D. Trilling (Eds.), *Rethinking Research Methods in an Age of Digital Journalism* (pp. 8-23). Routledge.

Brants, T., Chen, F., & Tsochantaridis, I. (2002). Topic-based document segmentation with probabilistic latent semantic analysis. In J. Xu, M. Gen, A. Hajiyev, & F. Lee Cooke (Eds.).

Proceedings of the eleventh international conference on Information and knowledge management (pp. 211-218). Springer. https://doi.org/10.1145/584792.584829

Brooks, D. J., & Geer, J. G. (2007). Beyond negativity: The effects of incivility on the electorate. *American Journal of Political Science, 51*(1), 1–16. https://doi.org/10.1111/j.1540-5907.2007.00233.x

Brosius, H. B., Haas, A., & Unkel, J. (2022). *Inhaltsanalyse III: Automatisierte Inhaltsanalyse. In Methoden der empirischen Kommunikationsforschung: Eine Einführung [Content analysis III: Automated content analysis. In Methods of Empirical Communication Research: An Introduction]*. In H.-B. Brosius, A. Haas, & J. Unkel (Eds.), Methoden der empirischen Kommunikationsforschung. Eine Einführung (pp. 179-194). Springer VS Wiesbaden. https://doi.org/10.1007/978-3-658-34195-4_10

Brown, A. (2017). What is hate speech? Part 1: The myth of hate. *Law and Philosophy, 36*(4), 419-468. https://doi.org/10.1007/s10982-017-9297-1

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. *Big Data & Society, 8*(1), 2053951720983865. https://doi.org/10.1177/2053951720983865

Brown, P., & Levinson, S. C. (1987). *Politeness. Some universals in language usage.* Cambridge University Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, & H. Lin (Eds.). *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)* (1877-1901). NeurIPS.

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, *7*(2), 223-242. https://doi.org/10.1002/poi3.85

Burnap, P., & Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data science, 5*(11), 1-15. https://doi.org/10.1140/epjds/s13688-016-0072-6

Cabitza, F., Locoro, A., Alderighi, C., Rasoini, R., Compagnone, D., & Berjano, P. (2019).The elephant in the record: On the multiplicity of data recording work. *Health Informatics Journal*, *25*(3), 475-490. https://doi.org/10.1177/1460458218824705

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication, 64*(4), 658-679. https://doi.org/10.1111/jcom.12104

Coleman, S., & Shane, P. M. (2011). *Connecting democracy: Online consultation and the flow of political communication* (1st ed.). MIT Press.

Cortina, L. M., Magley, V. J., Williams, J. H., & Langhout, R. D. (2001). Incivility in the workplace: incidence and impact. *Journal of occupational health psychology*, *6*(1), 64. https://doi.org/10.1037/1076-8998.6.1.64

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321-357. https://doi.org/10.1613/jair.953

Chen, G. M. (2017). *Online incivility and public debate: Nasty talk.* Palgrave Macmillan. https://doi.org/10.1007/978-3-319-56273-5

Chen, G. M., Muddiman, A., Wilner, T., Pariser, E., & Stroud, N. J. (2019). We should not get rid of incivility online. *Social Media + Society, 5*(3), 2056305119862641. https://doi.org/10.1177/2056305119862641

Chen, G. M., & Lu, S. (2017). Online political discourse: Exploring differences in effects of civil and uncivil disagreement in news website comments. *Journal of Broadcasting & Electronic Media, 61*(1), 108–125. https://doi.org/10.1080/08838151.2016.1273922

Chen, G. M., & Ng, Y. M. M. (2017). Nasty online comments anger you more than me, but nice ones make me as happy as you. *Computers in Human Behavior, 71*, 181-188**.** https://doi.org/10.1016/j.chb.2017.02.010

Chen, G. M., and P. Pain (2017). Normalizing Online Comments. *Journalism Practice, 11*(7), 876–892. https://doi.org/10.1080/17512786.2016.1205954

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing* (pp. 71-80). IEEE. https://doi.org/10.1109/SocialCom-PASSAT.2012.55

Chomsky, N. (1965). *Aspects of the Theory of Syntax* (1st ed.). MIT Press.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use* (1st ed.). Greenwood Publishing Group.

Chowdhary, K.R. (2020). Natural Language Processing. In K. Chowdhary (Eds.) *Fundamentals of artificial intelligence* (pp. 603-649). Springer. https://doi.org/10.1007/978-81-322-3972-7_19

Crawford, K. (2021). *Atlas of AI - Power, Politics, and the Planetary Costs of Artificial Intelligence* (1st ed.). Yale University Press.

Dahlberg, L. (2001). The Internet and democratic discourse. Exploring the prospects of online deliberative forums extending the public sphere. *Information, Communication & Society, 4*(4), 615–633. International Joint Conferences on Artificial Intelligence. https://doi.org/10.1080/13691180110097030

Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. In C. Sierra (Ed.), *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence: AI and autonomy track* (pp. 4691-4697). https://doi.org/10.24963/ijcai.2017/654

Davidson, S., Sun, Q., & Wojcieszak, M. (2020). Developing a new classifier for automated identification of incivility in social media. In S. Akiwowo, B. Vidgen, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the fourth workshop on online abuse and harms* (pp. 95-101). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.alw-1.12

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media: Vol. 11* (pp. 512-515). AAAI Press. https://doi.org/10.1609/icwsm.v11i1.14955

deepset. (2019, June 14). *German BERT Model*. https://www.deepset.ai/german-bert

Denil, M., & Trappenberg, T. (2010). Overlap versus imbalance. In A. Farzindar & V. Kešelj (Eds.), *Advances in Artificial Intelligence. Canadian AI 2010. Lecture Notes in Computer Science: Volume 6085* (pp. 220-231). Springer. https://doi.org/10.1007/978-3-642-13059-5_22

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv. https://doi.org/10.48550/arXiv.1810.04805

Diakopoulos, N., & Naaman, M. (2011). Towards quality discourse in online news comments. In P. Hinds, J. C. Tang, J. Wang, J. Bardram, & N. Ducheneaut (Eds.), *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (pp. 133-142). ACM. https://doi.org/10.1145/1958824.1958844

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In J. Furman, G. Marchant, H. Price, & F. Rossi (Eds.), *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 67-73). ACM. https://doi.org/10.1145/3278721.3278729

Dobbrick, T., Jakob, J., Chan, C. H., & Wessler, H. (2021). Enhancing theory-informed dictionary approaches with "glass-box" machine learning: The case of integrative complexity in social media comments. *Communication Methods and Measures*, *16*(4), 1–18. https://doi.org/10.1080/19312458.2021.1999913

Domahidi, E., Yang, J., Niemann-Lenz, J., & Reinecke, L. (2019). Outlining the Way Ahead in Computational Communication Science: An Introduction to the IJoC Special Section on "Computational Methods for Communication Science: Toward a Strategic Roadmap". *International Journal of Communication*, *13*, 3876–3884.

Dong, L., Wei, F., Zhou, M., & Xu, K. (2015). Question answering over freebase with multi-column convolutional neural networks. In C. Zong & M. Strube (Eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing: Volume 1* (pp. 260-269). Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-1026

Dubey, R., Zhou, J., Wang, Y., Thompson, P. M., Ye, J., & Alzheimer's Disease Neuroimaging Initiative. (2014). Analysis of sampling techniques for imbalanced data: An n= 648 ADNI study. *NeuroImage, 87*, 220-241. https://doi.org/10.1016/j.neuroimage.2013.10.005

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In J. J. O'Hare (Ed.), *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 281-285). ACM. https://doi.org/10.1145/57167.57214

Esau, K., Fleuß, D., & Nienhaus, S.-M. (2021). Different Arenas, Different Deliberative Quality? Using a Systemic Framework to Evaluate Online Deliberation on Immigration Policy in Germany. *Policy & Internet*, *13*(1), 86–112. https://doi.org/10.1002/poi3.232

European Commission. (2023, June). *The Digital Services Act package.*

https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

Ferrer, X., van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2021). Bias and Discrimination in AI: a cross-disciplinary perspective. *IEEE Technology and Society Magazine, 40*(2), 72-80. https://doi.org/10.1109/MTS.2021.3056293

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, *51*(4), 1-30. https://doi.org/10.1145/3232676

Fraser, N. (1990). Rethinking the public sphere: A contribution to the critique of actually existing democracy. *Social Text, 25*/*26*, 56–80. https://doi.org/10.2307/466240

Frieß, D., & Porten-Cheé, P. (2018). What Do Participants Take Away from Local eParticipation? *Analyse & Kritik, 40*(1), 1–29. https://doi.org/10.1515/auk-2018-0001

Friess, D., Ziegele, M., & Heinbach, D. (2021). Collective civic moderation for deliberation? Exploring the links between citizens' organized engagement in comment sections and the deliberative quality of online discussions. *Political Communication, 38*(5), 624–646. https://doi.org/10.1080/10584609.2020.1830322

Früh, W. (2015). *Inhaltsanalyse: Theorie und Praxis [Content analysis: theory and practice]* (8th ed.) utb.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, *36*(4), 193-202. https://doi.org/10.1007/BF00344251

García-Díaz, J., Caparros-Laiz, C., & Valencia-García, R. (2022). UMUTeam at SemEval-2022 Task 5: Combining image and textual embeddings for multi-modal automatic misogyny identification. In G. Emerson et al. (Eds.). *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 742-747). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.semeval-1.103

Georgakopoulos, S. V., Tasoulis, S. K., Vrahatis, A. G., & Plagianakos, V. P. (2018). Convolutional neural networks for toxic comment classification. In N. Fakotakis & V. Megalooikonomou (Eds.), *Proceedings of the 10th hellenic conference on artificial intelligence* (pp. 1-6). ACM. https://doi.org/10.1145/3200947.3208069

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural computation*, *12*(10), 2451-2471. https://doi.org/10.1162/089976600300015015

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, *57*, 345-420. https://doi.org/10.1613/jair.4992

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 7*(1), 2053951719897945. https://doi.org/10.1177/2053951719897945

Grimmelmann, J. (2015). The virtues of moderation. Yale Journal of Law and Technology, 17(1), 42–109.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences* (1st ed.). Princeton University Press.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science robotics, 4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Günther, E., & Quandt, T. (2018). Word counts and topic models: Automated text analysis methods for digital journalism research. In M. Karlsson & H. Sjøvaag (Eds.), *Rethinking Research Methods in an Age of Digital Journalism* (pp. 75-88). Routledge.

Habermas, J. (1996). *Between facts and norms: Contributions to a discourse theory of law and democracy.* MIT Press. https://doi.org/10.7551/mitpress/1564.001.0001

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications, 73*(1), 220-239. https://doi.org/10.1016/j.eswa.2016.12.035

Harlow, S. (2015). Story-Chatterers Stirring Up Hate: Racist Discourse in Reader Comments on U.S. Newspaper Websites. *Howard Journal of Communications 26*(1), 21–42. https://doi.org/10.1080/10646175.2014.984795

Hase, V., Mahl, D., & Schäfer, M. S. (2022). Der „Computational Turn“: ein „interdisziplinärer Turn“? Ein systematischer Überblick zur Nutzung der automatisierten Inhaltsanalyse in der Journalismusforschung [The "computational turn": an "interdisciplinary turn"? A Systematic Overview of the Use of Automated Content Analysis in Journalism Research.]. *Medien & Kommunikationswissenschaft*, *70*(1-2), 60-78. https://doi.org/10.5771/1615-634x-2022-1-2-60

Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & Quantity, 51*(6), 2623–2646. https://doi.org/10.1007/s11135-016-0412-4

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77-89. https://doi.org/10.1080/19312450709336664

Heinbach, D. (in press). Qualität und Inzivilität von Online-Diskussionen [Quality and incivility of online diskussionen]. In N. Kersting, J. Radtke, & S. Baringhorst (Eds.), *Handbuch Digitalisierung und politische Beteiligung*. Springer VS.

Heinbach, D. & Wilms, L. K. (2022): Der Einsatz von Moderation bei #meinfernsehen2021 [Use of moderation at #meinfernsehen2021]. In F. Gerlach & C. Eilders (Eds.), *#meinfernsehen2021. Bürgerbeteiligung: Wahrnehmungen, Erwartungen und Vorschläge zur Zukunft öffentlich-rechtlicher Medienangebote* (pp. 217-236). Nomos.

Herbst, S. (2010). *Rude democracy: Civility and incivility in American politics*. Temple University Press.

Herndon, N., & Caragea, D. (2016). A study of domain adaptation classifiers derived from logistic regression for the task of splice site prediction. *IEEE transactions on nanobioscience, 15*(2), 75-83. https://doi.org/10.1109/TNB.2016.2522400

Hilbert, M., Barnett, G., Blumenstock, J., Contractor, N., Diesner, J., Frey, S., ... & Zhu, J. J. (2019). Computational communication science: A methodological catalyzer for a maturing discipline. *International Journal of Communication 13*, 3912–3934.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hofmann, T. (1999). Probabilistic latent semantic indexing. In F.Gey,M. Hearst, & R. Tong (Eds.), *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50-57). ACM. https://doi.org/10.1145/312624.312649

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). *Deceiving google's perspective api built for detecting toxic comments.* ArXiv. https://doi.org/10.48550/arXiv.1702.08138

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass, 1*5(8), e12432. https://doi.org/10.1111/lnc3.12432

Hsueh, M., Yogeeswaran, K., & Malinen, S. (2015). "Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors? *Human Communication Research, 41*(4), 557–576. https://doi.org/10.1111/hcre.12059

Hwang, H., Kim, Y., & Huh, C. U. (2014). Seeing is believing: Effects of uncivil online debate on political polarization and expectations of deliberation. *Journal of Broadcasting & Electronic Media*, *58*(4), 621-633. https://doi.org/10.1080/08838151.2014.966365

Hwang, H., Kim, Y., & Kim, Y. (2018). Influence of discussion incivility on deliberation: An examination of the mediating role of moral indignation. *Communication Research, 45*(2), 213–240. https://doi.org/10.1177/0093650215616861

Ishwari, K. S. D., Aneeze, A. K. R. R., Sudheesan, S., Karunaratne, H. J. D. A., Nugaliyadde, A., & Mallawarrachchi, Y. (2019). *Advances in natural language question answering: A review.* ArXiv. https://doi.org/10.48550/arXiv.1904.05276

Jelodar, H., Wang, Y., Orji, R., & Huang, S. (2020). Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics, 24*(10), 2733-2742. https://doi.org/10.1109/JBHI.2020.3001216

Johnson, R., & Zhang, T. (2015). Semi-supervised convolutional neural networks for text categorization via region embedding. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.) *Advances in neural information processing systems: Volume 1* (919–927). MIT Press.

Johri, P., Khatri, S. K., Al-Taani, A. T., Sabharwal, M., Suvanov, S., & Kumar, A. (2021). Natural language processing: History, evolution, application, and future work. In A. Abraham, O. Castillo, & D. Virmani (Eds.), *Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020* (pp. 365-375). Springer Singapore.

Jones, K. S. (1994). Natural language processing: A Historical Review. In A. Zampolli, N. Calzolari, & M. Palmer (Eds.), *Current issues in computational linguistics: In honour of Don Walke*r (pp. 3-16). Springer. https://doi.org/10.1007/978-0-585-35958-8_1

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260. https://doi.org/10.1126/science.aaa8415

Kalch, A., & Naab, T. K. (2017). Replying, disliking, flagging: How users engage with uncivil and impolite comments on news sites. *Studies in Communication and Media 6*(4)*,* 395-419. https://doi.org/10.5771/2192-4007-2017-4-395

Kenski, K., Coe, K., & Rains, S. A. (2020). Perceptions of uncivil discourse online: An examination of types and predictors. *Communication Research*, *47*(6), 795-814.

Kersting, N. (2019). Online Partizipation: Evaluation und Entwicklung – Status Quo und Zukunft [Online Participation: Evaluation and Development - Status Quo and Future]. In J. Hofmann, N. Kersting, C. Ritzi, & W. J. Schünemann (Eds.), *Politik in der digitalen Gesellschaft* (S. 105–122). Transcript Verlag. https://doi.org/10.14361/9783839448649-006

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications, 82*(3), 3713-3744. https://doi.org/10.1007/s11042-022-13428-4

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). Sage publications.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84-90. https://doi.org/10.1145/3065386

Kroon, A. C., Trilling, D., & Raats, T. (2021). Guilty by association: Using word embeddings to measure ethnic stereotypes in news coverage. *Journalism & Mass Communication Quarterly, 98*(2), 451-477. https://doi.org/10.1177/1077699020932304

Ksiazek, T. B. (2018). Commenting on the news: Explaining the degree and quality of user comments on news websites. *Journalism Studies, 19*(5), 650–673. https://doi.org/10.1080/1461670X.2016.1209977

Ksiazek, T. B., Peer, L., & Zivic, A. (2015). Discussing the news: Civility and hostility in user comments. *Digital journalism, 3*(6), 850-870. https://doi.org/10.1080/21670811.2014.972079

Kumaresan, P. K., Premjith, Sakuntharaj, R., Thavareesan, S., Navaneethakrishnan, S., Madasamy, A. K., ... & McCrae, J. P. (2021). Findings of shared task on offensive language identification in Tamil and Malayalam. In D. Ganguly, S. Gangopadhyay, M. Mitra, & P. Majumder (Eds.), *FIRE '21: Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation* (pp. 16-18). ACM. https://doi.org/10.1145/3503162.3503179

Kümpel, A. S. & Rieger, D. (2019). *Wandel der Sprach- und Debattenkultur in sozialen Online-Medien [Change of speech and debate culture in social media]*. Konrad-Adenauer-Stiftung. https://doi.org/10.5282/ubm/epub.68880

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., ... & Van Alstyne, M. (2009). Computational Social Science. *Science, 323*(5915), 721-723. https://doi.org/10.1126/science.1167742

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature, 521*, 436-444. https://doi.org/10.1038/nature14539

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324. https://doi.org/10.1109/5.726791

Lee, S., Jang, H., Baik, Y., Park, S., & Shin, H. (2020). *KR-BERT: A Small-Scale Korean-Specific Language Model.* ArXiv. https://doi.org/10.48550/arXiv.2008.03979

Lind, F., Eberl, J. M., Heidenreich, T., & Boomgaarden, H. G. (2019). Computational communication science| when the journey is as important as the goal: A roadmap to multilingual dictionary construction. *International Journal of Communication*, *13*, 4000–4020

Lind, F., Gruber, M., & Boomgaarden, H. G. (2017). Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. *Communication Methods and Measures, 11*(3), 191–209. https://doi.org/10.1080/19312458.2017.1317338

Lind, F., & Meltzer, C. E. (2021). Now you see me, now you don't: Applying automated content analysis to track migrant women's salience in German news. *Feminist Media Studies, 21*(6), 923-940. https://doi.org/10.1080/14680777.2020.1713840

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. ArXiv. https://doi.org/10.48550/arXiv.1907.11692

Loosen, W., & Scholl, A. (2012). *Methodenkombinationen in der Kommunikationswissenschaft. Methodologische Herausforderungen und empirische Praxis [Combining methods in communication science. Methodological challenges and empirical practice]* (1st ed.). Herbert von Halem.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., ... & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures, 12*(2-3), 93-118. https://doi.org/10.1080/19312458.2018.1430754

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing* (1st ed.). MIT Press.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de La Clergerie, É. V., ... & Sagot, B. (2019). *CamemBERT: a tasty French language model*. ArXiv. https://doi.org/10.48550/arXiv.1911.03894

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Volume 1* (pp. 142-150).

Massaro, T. M., & Stryker, R. (2012). Freedom of speech, liberal democracy, and emerging evidence on civility and effective democratic engagement. *Arizona Law Review*, *54*(2), 375-442.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR), 54*(6), 1-35. https://doi.org/10.1145/3457607

Meltzer, C. E., Eberl, J. M., Theorin, N., Heidenreich, T., Strömbäck, J., Boomgaarden, H. G., & Schemer, C. (2021). Media effects on policy preferences toward free movement: evidence from five EU member states. *Journal of Ethnic and Migration Studies, 47*(15), 3390-3408. https://doi.org/10.1080/1369183X.2020.1778454

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* ArXiv. https://doi.org/10.48550/arXiv.1301.37

Muddiman, A. (2017). Personal and public levels of political incivility. *International Journal of Communication*, *11*, 3182–3202.

Muddiman, A. (2019). How people perceive political incivility. In R. G. Boatright, T. J. Shaffer, S. Sobieraj, & D. G. Young (Eds.), *A Crisis of Civility? Political Discourse and Its Discontents*. (pp. 31-44). Routledge.

Muddiman, A., & Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of communication, 67*(4), 586-609. https://doi.org/10.1111/jcom.12312

Mutz, D. C. (2007). Effects of "In-your-face" television discourse on perceptions of a legitimate opposition. *American Political Science Review, 101*(4), 621–635. https://doi.org/10.1017/S000305540707044X

Mutz, D. C., & Reeves, B. (2005). The new videomalaise: Effects of televised incivility on political trust. *American Political Science Review, 99*(1), 1–15. https://doi.org/10.1017/S0003055405051452

Naab, T. K., Kalch, A., & Meitz, T. G. (2018). Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society*, *20*(2), 777-795. https://doi.org/10.1177/1461444816670923

Naab, T. K., Naab, T., & Brandmeier, J. (2021). Uncivil user comments increase users' intention to engage in corrective actions and their support for authoritative restrictive actions. *Journalism & Mass Communication Quarterly*, *98*(2), 566-588. https://doi.org/10.1177/1077699019886586

Nadeem, M., Bethke, A., & Reddy, S. (2020). *StereoSet: Measuring stereotypical bias in pretrained language models*. ArXiv. https://doi.org/10.48550/arXiv.2004.09456

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An Introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544-551. https://doi.org/10.1136/amiajnl-2011-000464

Navigli, Roberto (2009). Word sense disambiguation. *ACM Computing Surveys, 41*(2), 1–69. https://doi.org/10.1145/1459352.1459355

Nguyen, T. H., & Grishman, R. (2015). Relation extraction: Perspective from convolutional neural networks. In P. Blunsom, S. Cohen, P. Dhillon, & P. Liang (Eds.), *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing* (pp. 39-48). Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-1506

Nielsen, C. (2012). Newspaper journalists support online comments. *Newspaper Research Journal, 33*(1), 86–100. https://doi.org/10.1177/073953291203300107

Niemann-Lenz, J., Bruns, S., Hefner, D., Knop-Hülß, K., Possler, D., Reich, S., ... & Klimmt, C. (2019). Computational communication science| crafting a strategic roadmap for computational methods in communication science: Learnings from the CCS 2018 Conference in Hanover–Commentary. *International Journal of Communication*, *13*, 9.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In J. Bourdeau, J. A. Hendler, R. Nkambou (Eds.), *Proceedings of the 25th international conference on world wide web* (pp. 145-153). ACM. https://doi.org/10.1145/2872427.2883062

Oz, M., Zheng, P., & Chen, G. M. (2018). Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. *New media & society, 20*(9), 3400-3419. https://doi.org/10.1177/1461444817749516

Paasch-Colberg, S., & Strippel, C. (2022). "The Boundaries are Blurry…": How Comment Moderators in Germany See and Respond to Hate Comments. *Journalism Studies*, *23*(2), 224-244. https://doi.org/10.1080/1461670X.2021.2017793

Paasch-Colberg, S., Strippel, C., Trebbe, J., & Emmer, M. (2021). From insult to hate speech: Mapping offensive language in German user comments on immigration. *Media and Communication*, *9*(1), 171-180. https://doi.org/10.17645/mac.v9i1.3399

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New media & society*, *6*(2), 259-283. https://doi.org/10.1177/1461444804041444

Paraschiv, A., & Cercel, D. C. (2019). UPB at GermEval-2019 Task 2: BERT-Based Offensive Language Classification of German Tweets. In S. Evert (Ed.), *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)* (pp. 396–402). German Society for Computational Linguistics & Language Technology.

Pavlopoulos, J., Malakasiotis, P., & Androutsopoulos, I. (2017). Deeper attention to abusive user content moderation. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 1125-1135). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1117

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Perspective. (n.d.). *About the API*. Retrieved June 18, 2023, from https://developers.perspectiveapi.com/s/about-the-api?language=en_US

Prochazka, F., Weber, P., & Schweiger, W. (2018). Effects of civility and reasoning in user comments on perceived journalistic quality. *Journalism studies*, *19*(1), 62-78. https://doi.org/10.1080/1461670X.2016.1161497

Puschmann, C. (2019). Beyond the bubble: Assessing the diversity of political search results. *Digital Journalism, 7*(6), 824-843. https://doi.org/10.1080/21670811.2018.1539626

Qaisar, S. M. (2020). Sentiment analysis of IMDb movie reviews using long short-term memory. *2020 2nd International Conference on Computer and Information Sciences (ICCIS), Saudi Arabia,* (pp. 1-4). IEEE. https://doi.org/10.1109/ICCIS49240.2020.9257657

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018, June 11). *Improving language understanding by generative pre-training.* https://openai.com/research/language-unsupervised

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019, February 14). *Better language models and their implications.* https://openai.com/research/better-language-models

Rains, S. A., Kenski, K., Dajches, L., Duncan, K., Yan, K., Shin, Y., ... & Shmargad, Y. (2023). Engagement with incivility in tweets from and directed at local elected officials. *Communication and Democracy, 57*(1), 143-152. https://doi.org/10.1080/27671127.2023.2195467

Rajalakshmi, R., Selvaraj, S., & Vasudevan, P. (2023). Hottest: Hate and offensive content identification in Tamil using transformers and enhanced stemming. *Computer Speech & Language*, *78*, 101464. https://doi.org/10.1016/j.csl.2022.101464

Razavi, A. H., Inkpen, D., Uritsky, S., & Matwin, S. (2010). Offensive language detection using multi-level classification. In A. Farzindar & V. Kešelj (Eds.), *Advances in Artificial Intelligence. Canadian AI 2010. Lecture Notes in Computer Science: Volume 6085* (pp. 16-27). Springer. https://doi.org/10.1007/978-3-642-13059-5_5

Reich, Z. (2011). User comments: The transformation of participatory space. In J.B. Singer, A. Hermida, D. Domingo, A. Heinonen, S. Paulussen, T. Quandt, & M.Vujnovic (Eds.), *Participatory journalism: Guarding open gates at online newspapers* (pp. 96-117). Wiley-Blackwell.

Risch, J. (2020). *Reader comment analysis on online news platforms* [Doctoral dissertation, University of Potsdam]. publish.UP. University of Potsdam. https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/deliver/index/docId/48922/file/risch_diss.pdf

Risch, J. (2023). Toxicity. In C. Strippel, S. Paasch-Colberg, M. Emmer & J. Trebbe (Eds.). *Challenges and perspectives of hate speech research* (pp. 219-230). Digital Communication Research. https://doi.org/10.48541/dcr.v12.13

Risch, J., Krebs, E., Löser, A., Riese, A., & Krestel, R. (2018). Fine-grained classification of offensive language. In J. Ruppenhofer, M. Siegel, & M. Wiegand (Eds.), *Proceedings*

of the GermEval 2018 Workshop: 14th Conference on Natural Language Processing KONVENS 2018 (KONVENS 2018) (pp. 38-44). Austrian Academy of Sciences.

Risch, J., & Krestel, R. (2020). Toxic comment detection in online discussions. In B. Agarwal, R. Nayak, N. Mittal, & S. Patnaik (Eds.), *Deep Learning-Based Approaches for Sentiment Analysis* (1st ed., pp. 85–109). Springer Singapore. https://doi.org/10.1007/978-981-15-1216-2_4

Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation, 56*(2), 593-619. https://doi.org/10.1007/s10579-021-09537-5

Rooduijn, M., & Pauwels, T. (2011). Measuring populism: Comparing two methods of content analysis. *West European Politics*, *34*(6), 1272-1283. https://doi.org/10.1080/01402382.2011.616665

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). *Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis.* ArXiv. https://doi.org/10.17185/duepublico/42132

Rössler, P. (2017). *Inhaltsanalyse [Content analysis]* (3rd ed.). utb.

Rowe, I. (2015). Civility 2.0: A comparative analysis of incivility in online political discussion. *Information, communication & society*, *18*(2), 121-138. https://doi.org/10.1080/1369118X.2014.940365

Ruiz, C., Domingo, D., Micó, J. L., Díaz-Noci, J., Meso, C., & Masip, P. (2011). Public sphere 2.0? The democratic qualities of citizen debates in online newspapers. *The International Journal of Press/Politics, 16*(4), 463–487. https://doi.org/10.1177/1940161211415849

Romberg, J. (2022). Is Your Perspective Also My Perspective? Enriching Prediction with Subjectivity. In G. Lapesa, J. Schneider, Y. Jo, & S. Saha (Eds.), *Proceedings of the 9th Workshop on Argument Mining* (pp. 115-125). International Conference on Computational Linguistics.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature, 323*(6088), 533-536. https://doi.org/10.1038/323533a0

Sadeque, F., Rains, S., Shmargad, Y., Kenski, K., Coe, K., & Bethard, S. (2019). Incivility detection in online comments. In R. Mihalcea, E. Shutova, L.-W. Ku, K. Evang, & S. Poria (Eds.), *Proceedings of the eighth joint conference on lexical and computational*

semantics (SEM 2019)* (pp. 283-291). Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-1031

Santana, A. D. (2014). Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism practice*, *8*(1), 18-33. https://doi.org/10.1080/17512786.2013.813194

Satapara, S., Majumder, P., Mandl, T., Modha, S., Madhu, H., Ranasinghe, T., ... & Premasiri, D. (2022). Overview of the HASOC Subtrack at FIRE 2022: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages. In D. Ganguly, S. Gangopadhyay, M. Mitra, & P. Majumder (Eds.), *FIRE '22: Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation* (pp. 4-7). ACM. https://doi.org/10.1145/3574318.3574326

Scharkow, M. (2012). *Automatische Inhaltsanalyse und maschinelles Lernen [Automatic content analysis and machine learning]* [Doctoral dissertation, Berlin University of the Arts]. Kooperativer Bibliotheksverbund Berlin-Brandenburg. https://opus4.kobv.de/opus4-udk/frontdoor/deliver/index/docId/28/file/dissertation_scharkow_final_udk.pdf

Scheper, J., & Kathirgamalingam, A. (2022). Fünf Thesen zur Integration von CCS in der Lehre [Five theses on the integration of CCS in teaching]. *aviso*, *22*(2), 6-7.

Scherr, S., Arendt, F., & Haim, M. (2022). Algorithms without frontiers? How language-based algorithmic information disparities for suicide crisis information sustain digital divides over time in 17 countries. *Information, Communication & Society*, 1-17. https://doi.org/10.1080/1369118X.2022.2097017

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks, 61*, 85-117. https://doi.org/10.1016/j.neunet.2014.09.003

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In L.-W. Ku & C.-T. Li (Eds.), *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1-10). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1101

Spertus, E. (1997). Smokey: Automatic recognition of hostile messages. In *AAAI'97/IAAI'97: Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence* (pp. 1058-1065). AAAI Press.

Sponholz, L. (2023). Hate Speech. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 4143-163). Böhland & Schremmer Verlag. https://doi.org/10.48541/dcr.v12.9

Springer, N., Engelmann, I., & Pfaffinger, C. (2015). User comments: Motives and inhibitors to write and read. *Information, Communication & Society*, *18*(7), 798-815. https://doi.org/10.1080/1369118X.2014.997268

Stroud, N. J., Scacco, J. M., Muddiman, A., & Curry, A. L. (2015). Changing deliberative norms on news organizations' Facebook sites. *Journal of Computer-mediated Communication*, *20*(2), 188–203. https://doi.org/10.1111/jcc4.12104

Stroud, N. J., Van Duyn, E., & Peacock, C. (2016). News commenters and news comment readers. https://mediaengagement.org/wp-content/uploads/2016/03/ENP-News-Commenters-and-Comment-Readers1.pdf

Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., & Klenner, M. (2019). Overview of germeval task 2, 2019 shared task on the identification of offensive language. In S. Evert (Ed.), *Proceedings of the 15th Conference on Natural Language Processing* (pp. 3541-365). German Society for Computational Linguistics & Language Technology.

Stryker, R., Conway, B. A., & Danielson, J. T. (2016). What is political incivility?. *Communication Monographs*, *83*(4), 535-556. https://doi.org/10.1080/03637751.2016.1201207

Su, L. Y. F., Xenos, M. A., Rose, K. M., Wirz, C., Scheufele, D. A., & Brossard, D. (2018). Uncivil and personal? Comparing patterns of incivility in comments on the Facebook pages of news outlets. *New Media & Society, 20*(10), 3678–3699. https://doi.org/10.1177/1461444818757205

Sydnor, E. (2018). Platforms for incivility: examining perceptions across different media formats. *Political Communication, 35*(1), 97–116. https://doi.org/10.1080/10584609.2017.1355857

Tahmasbi, N., & Rastegari, E. (2018). A socio-contextual approach in automated detection of public cyberbullying on Twitter. A*CM Transactions on Social Computing, 1*(4), 1-22. https://doi.org/10.1145/3290838

Taradhita, D. A. N., & Darma Putra, I. (2021). Hate Speech Classification in Indonesian Language Tweets by Using Convolutional Neural Network. *Journal of ICT Research & Applications, 14*(3), 225-239. https://doi.org/10.5614/itbj.ict.res.appl.2021.14.3.2

Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. A. (2020). The dynamics of political incivility on Twitter. *Sage Open, 10*(2), 2158244020919447. https://doi.org/10.1177/2158244020919447

Valdez, D., Pickett, A. C., & Goodson, P. (2018). Topic modeling: latent semantic analysis for the social sciences. *Social Science Quarterly*, *99*(5), 1665-1679. https://doi.org/10.1111/ssqu.12528

Van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. In D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, & J. Wernimont (Eds.), *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)* (pp. 33–42). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5105

Van Atteveldt, W., Margolin, D., Shen, C., Trilling, D., & Weber, R. (2019). A roadmap for computational communication research. *Computational Communication Research*, *1*(1), 1-11. https://doi.org/10.5117/CCR2019.1.001.VANA

Van Atteveldt, W., & Peng, T. Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2-3), 81-92. https://doi.org/10.1080/19312458.2018.1458084

Van Duyn, E., & Muddiman, A. (2022). Predicting perceptions of incivility across 20 news comment sections. *Journalism*, *23*(1), 134-152. https://doi.org/10.1177/1464884920907779

Van Duyn, E., Peacock, C., & Stroud, N. J. (2021). The gender gap in online news comment sections. *Social Science Computer Review*, *39*(2), 181-196. https://doi.org/10.1177/0894439319864876

Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine learning, 109*(2), 373-440. https://doi.org/10.1007/s10994-019-05855-6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000–6010). Curran Associates Inc.

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one, 15*(12), e0243300. https://doi.org/10.1371/journal.pone.0243300

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. In S. T. Roberts, J. Tetreault, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the third workshop on abusive language online* (pp. 80-93). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3509

Vox Media (n.d.). *Your community is our priority: Coral features tools and experiences for commenters, moderators, community managers, and journalists alike*. https://coralproject.net/tour/

Walther, J. B., & Jang, J. W. (2012). Communication processes in participatory websites. *Journal of Computer-Mediated Communication*, *18*(1), 2-15. https://doi.org/10.1111/j.1083-6101.2012.01592.x

Wang, S. (2020). The influence of anonymity and incivility on perceptions of user comments on news websites. *Mass Communication and Society*, *23*(6), 912-936. https://doi.org/10.1080/15205436.2020.1784950

Wang, X., Liu, Y., Sun, C. J., Wang, B., & Wang, X. (2015). Predicting polarities of tweets by composing word embeddings with long short-term memory. In C. Zong & M. Strube (Eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1343-1353). Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-1130

Wich, M., Bauer, J., & Groh, G. (2020). Impact of politically biased data on hate speech classification. In S. Akiwowo, B. Vidgen, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the fourth workshop on online abuse and harms* (pp. 54-64). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.alw-1.7

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In J. Ruppenhofer, M. Siegel, & M. Wiegand (Eds.), *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)* (pp. 1-10). Austrian Academy of Sciences.

Wilhelm, C., Joeckel, S., & Ziegler, I. (2020). Reporting hate comments: Investigating the effects of deviance characteristics, neutralization strategies, and users' moral orientation. *Communication Research*, *47*(6), 921-944. https://doi.org/10.1177/0093650219855330

Wilms, L., Gerl, K., Stoll, A., & Ziegele, M. (2023). Technology Acceptance and Transparency Demands for Automated Detection of Toxic Language – Interviews with Moderators of Public Online Discussion Fora. *Human-Computer Interaction.* https://doi.org/10.1080/07370024.2024.2307610

Wojatzki, M., Ruppert, E., Holschneider, S., Zesch, T., & Biemann, C. (2017). Germeval 2017: Shared task on aspect-based sentiment in social media customer feedback. In M. Wojatzki, E. Ruppert, T. Zesch, C. & Biemann (Eds.), *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback* (pp. 1-10). German Society for Computational Linguistics and Language Technology.

Wright, S. (2006). Government-run online discussion fora: Moderation, censorship and the shadow of Control1. *The British Journal of Politics and International Relations*, *8*(4), 550-568. https://doi.org/10.1111/j.1467-856x.2006.00247.x

Xiang, G., Fan, B., Wang, L., Hong, J., & Rose, C. (2012). Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In X. Chen (Ed.), *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 1980-1984). ACM.

Xu, J. M., Jun, K. S., Zhu, X., & Bellmore, A. (2012). Learning from bullying traces in social media. In E. Fosler-Lussier, E. Riloff, & S. Bangalore (Eds.), *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 656-666). Association for Computational Linguistics.

Yasen, M., & Tedmori, S. (2019). Movies reviews sentiment analysis and classification. In K. M. Jaber (Ed.), *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)* (pp. 860-865). JEEIT. https://doi.org/10.1109/JEEIT.2019.8717422

Young, M. (1996). *Communication and the other: Beyond deliberative democracy.* In S. Benhabib (Ed.), Democracy and difference (pp. 120–135). Princeton University Press. https://doi.org/10.1515/9780691234168-007

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, *29*(2), 205-231. https://doi.org/10.1080/10584609.2012.671234

Zakaryazad, A., & Duman, E. (2016). A profit-driven Artificial Neural Network (ANN) with applications to fraud detection and direct marketing. *Neurocomputing, 175*(A), 121–131. https://doi.org/10.1016/j.neucom.2015.10.042

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, & S. M. Mohammad (Eds.), *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 75–86). Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-2010

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., ... & Çöltekin, Ç. (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Editors), *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 1425–1447). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.188

Zampieri, M., Ranasinghe, T., Sarkar, D., & Ororbia, A. (2023). Offensive language identification with multi-task learning. *Journal of Intelligent Information Systems*. https://doi.org/10.1007/s10844-023-00787-z

Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems, 30*(11), 3212-3232. https://doi.org/10.1109/TNNLS.2018.2876865

Ziegele, M., Daxenberger, J., Quiring, O., & Gurevych, I. (2018, May 24-28). *Developing automated measures to predict incivility in public online discussions on the facebook sites of established news media* [Conference presentation abstract]. 68th Annual Conference of the International Communication Association (ICA), Prague, Czech Republic. https://public.ukp.informatik.tu-darmstadt.de/UKP_Webpage/publications/2018/2018_ICA_Ziegele_DevelopingAutomatedMeasures.pdf

Ziegele, M., Naab, T. K., & Jost, P. (2020). Lonely together? Identifying the determinants of collective corrective action against uncivil comments. *New Media & Society, 22*(5), 731-751. https://doi.org/10.1177/1461444819870130

Ziegele, M., & Jost, P. (2020). Not funny? The effects of factual versus sarcastic journalistic responses to uncivil user comments. *Communication Research, 47*(6), 891–920. https://doi.org/10.1177/0093650216671854

Ziegele, M., Jost, P., Bormann, M., & Heinbach, D. (2018). Journalistic counter-voices in comment sections: Patterns, determinants, and potential consequences of interactive moderation of uncivil user comments. *SCM Studies in Communication and Media*, *7*(4), 525-554. https://doi.org/10.5771/2192-4007-2018-4-525

Ziegele, M., Weber, M., Quiring, O., & Breiner, T. (2018). The dynamics of online news discussions: Effects of news articles and reader comments on users' involvement, willingness to participate, and the civility of their contributions. *Information, Communication & Society*, *21*(10), 1419-1435. https://doi.org/10.1080/1369118X.2017.1324505

Zhou, Z. (2021). *Machine Learning* (1st ed.). Springer Nature. https://doi.org/10.1007/978-981-15-1967-3