

MAXIMILIAN JÜRGEN WICH

ADDRESSING WEAK POINTS OF  
ABUSIVE LANGUAGE DETECTION  
ON SOCIAL MEDIA





Technische Universität München

Fakultät für Informatik

ADDRESSING WEAK POINTS OF  
ABUSIVE LANGUAGE DETECTION  
ON SOCIAL MEDIA

MAXIMILIAN JÜRGEN WICH

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

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr.-Ing. Jörg Ott

Prüfer der Dissertation: 1. Apl. Prof. Dr. Georg Groh  
2. Prof. Dr. Simon Hegelich

Die Dissertation wurde am 1. Dezember 2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 6. April 2022 angenommen.

Maximilian Jürgen Wich: *Addressing Weak Points of Abusive Language Detection on Social Media*, © November 2021

---

This document was typeset using L<sup>A</sup>T<sub>E</sub>X and the classicthesis template developed by André Miede and Ivo Pletikosić.

Hate speech is on the rise,  
threatening peace, social stability,  
and democratic values.

— *António Guterres,*  
*Secretary-General of*  
*the United Nations*



## ABSTRACT

---

Hate speech and other forms of abusive language on social media platforms have become a societal challenge. One component to cope with them is automatic abusive language detection based on machine learning (ML) and natural language processing (NLP). Despite groundbreaking advances, abusive language detection still faces challenges. Firstly, the models exhibit limited classification performance and generalizability. Secondly, they behave like black boxes, meaning it is impossible to understand a model's prediction. That is highly critical for a system that makes decisions between abusive and non-abusive speech because these decisions might warrant subsequent actions. These two problems are addressed by twelve studies. The studies contributed to this goal by increasing the quantity and quality of training data (e. g., creating new abusive language corpora) and identifying additional features relevant for abusive language classification (e. g., social network data). Besides that, they improved the generalizability of the models. Furthermore, they helped uncover unintended bias within training data and models (e. g., annotator and political bias) to avoid unfair behavior and explain the models' predictions to make them more understandable for humans (e. g., highlighting words relevant for prediction). The studies' findings contribute to better understanding the weak points of abusive language detection and building more accurate, more generalizable, and explainable classification models.

## ZUSAMMENFASSUNG

---

Hassrede und andere Formen von beleidigender Sprache auf Social-Media-Plattformen sind zu einer gesellschaftlichen Herausforderung geworden. Eine Komponente im Kampf gegen solche Hassrede ist deren automatische Erkennung basierend auf Machine Learning (ML) und Natural Language Processing (NLP). Trotz bahnbrechender Fortschritte steht die Erkennung von Hassrede immer noch vor großen Herausforderungen. Erstens weisen die Modelle eine begrenzte Klassifizierungsleistung und Verallgemeinerbarkeit auf. Zweitens verhalten sie sich wie Blackboxes, d. h. es ist fast unmöglich, die Vorhersage eines Modells nachzuvollziehen. Dies ist für ein System, das im Spannungsfeld zwischen Redefreiheit und Hassrede steht, äußerst kritisch. Diese beiden Probleme werden in zwölf Studien adressiert. Die Studien tragen zur Lösung der beiden Probleme bei, indem sie die Quantität und Qualität der Trainingsdaten erhöhen (z. B., durch die Erstellung neuer Trainingsdatensätze) und zusätzliche Merkmale identifizieren (z. B., soziale Netzwerkdaten), die für die Klassifizierung von Hassrede relevant sind. Außerdem verbessern sie die Verallgemeinerbarkeit der Modelle. Darüber hinaus leisten sie einen Beitrag dazu, unbeabsichtigten Bias (z. B., Annotator-Bias, politischer Bias) in den Trainingsdaten und Modellen aufzudecken, um unfaires Verhalten zu vermeiden, und die Vorhersagen der Modelle für Menschen verständlicher zu machen (z. B., Markieren von Wörtern, die relevant für die Vorhersage sind). Die Ergebnisse der Studien tragen dazu bei, die Schwachstellen bei der Erkennung von Hassrede besser zu verstehen und genauere, verallgemeinerbare und erklärable Klassifizierungsmodelle zu entwickeln.



## PUBLICATIONS

---

This dissertation is based on the following publications that are relevant for examination<sup>1</sup> and are marked with • in the dissertation.

- Wich, Maximilian, Jan Bauer, and Georg Groh (Nov. 2020). “Impact of Politically Biased Data on Hate Speech Classification.” In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, pp. 54–64. DOI: [10.18653/v1/2020.alw-1.7](https://doi.org/10.18653/v1/2020.alw-1.7). URL: <https://www.aclweb.org/anthology/2020.alw-1.7>.
- Wich, Maximilian, Melissa Breitingger, Wienke Strathern, Marlena Naimarevic, Georg Groh, and Jürgen Pfeffer (Apr. 2021). “Are Your Friends Also Haters? Identification of Hater Networks on Social Media: Data Paper.” In: *Companion Proceedings of the Web Conference 2021*. WWW ’21. Ljubljana, Slovenia: Association for Computing Machinery, pp. 481–485. ISBN: 9781450383134. DOI: [10.1145/3442442.3452310](https://doi.org/10.1145/3442442.3452310).
- Wich, Maximilian, Tobias Eder, Hala Al Kuwatly, and Georg Groh (July 2021). “Bias and comparison framework for abusive language datasets.” In: *AI and Ethics*. ISSN: 2730-5961. DOI: [10.1007/s43681-021-00081-0](https://doi.org/10.1007/s43681-021-00081-0).
- Wich, Maximilian, Edoardo Mosca, Adrian Gorniak, Johannes Hingerl, and Georg Groh (Sept. 2021). “Explainable Abusive Language Classification Leveraging User and Network Data.” In: *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*. Ed. by Yuxiao Dong, Nicolas Kourtellis, Barbara Hammer, and Jose A. Lozano. Cham: Springer International Publishing, pp. 481–496. ISBN: 978-3-030-86517-7. DOI: [10.1007/978-3-030-86517-7\\_30](https://doi.org/10.1007/978-3-030-86517-7_30).
- Wich, Maximilian, Svenja Räther, and Georg Groh (Sept. 2021). “German Abusive Language Dataset with Focus on COVID-19.” In: *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*. Düsseldorf, Germany: KONVENS 2021 Organizers, pp. 247–252. ISBN: 978-1-954085-83-1. URL: <https://aclanthology.org/2021.konvens-1.26>.
- Wich, Maximilian, Christian Widmer, Gerhard Hagerer, and Georg Groh (Sept. 2021). “Investigating Annotator Bias in Abusive Language Datasets.” In: *Deep Learning for Natural Language Processing Methods and Applications*. Held Online: INCOMA Ltd., pp. 1515–1525. ISBN: 978-954-452-072-4. DOI: [10.26615/978-954-452-072-4\\_170](https://doi.org/10.26615/978-954-452-072-4_170). URL: <https://aclanthology.org/2021.ranlp-1.170>.

---

<sup>1</sup> in accordance with Exhibit 6 of the regulations for the award of doctoral degree

The following publications contributed to the dissertation. They are not formally relevant for examination<sup>2</sup> and are marked with † in the dissertation.

- Al Kuwatly, Hala, Maximilian Wich, and Georg Groh (Nov. 2020). "Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics." In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, pp. 184–190. DOI: [10.18653/v1/2020.alw-1.21](https://doi.org/10.18653/v1/2020.alw-1.21). URL: <https://www.aclweb.org/anthology/2020.alw-1.21>.
- Lerch, Laurence, Maximilian Wich, Tobias Eder, and Georg Groh (2022). "Mediale Hasssprache und technologische Entscheidbarkeit: Zur ethischen Bedeutung subjektiv-perzeptiver Datenannotationen in der Hate Speech Detection." In: *Medien – Demokratie – Bildung: Normative Vermittlungsprozesse und Diversität in mediatisierten Gesellschaften*. Ed. by Gudrun Marci-Boehncke, Matthias Rath, Malte Delere, and Hanna Höfer. Wiesbaden: Springer Fachmedien Wiesbaden, pp. 295–310. ISBN: 978-3-658-36446-5. DOI: [10.1007/978-3-658-36446-5\\_17](https://doi.org/10.1007/978-3-658-36446-5_17).
- Mosca, Edoardo, Maximilian Wich, and Georg Groh (June 2021). "Understanding and Interpreting the Impact of User Context in Hate Speech Detection." In: *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. Online: Association for Computational Linguistics, pp. 91–102. DOI: [10.18653/v1/2021.socialnlp-1.8](https://doi.org/10.18653/v1/2021.socialnlp-1.8). URL: <https://aclanthology.org/2021.socialnlp-1.8>.
- Rieger, Diana, Anna Sophie Kümpel, Maximilian Wich, Toni Kiening, and Georg Groh (Oct. 2021). "Assessing the Extent and Types of Hate Speech in Fringe Communities: A Case Study of Alt-Right Communities on 8chan, 4chan, and Reddit." In: *Social Media + Society* 7.4. DOI: [10.1177/20563051211052906](https://doi.org/10.1177/20563051211052906).
- Wich, Maximilian, Hala Al Kuwatly, and Georg Groh (Nov. 2020). "Investigating Annotator Bias with a Graph-Based Approach." In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, pp. 191–199. DOI: [10.18653/v1/2020.alw-1.22](https://doi.org/10.18653/v1/2020.alw-1.22). URL: <https://www.aclweb.org/anthology/2020.alw-1.22>.
- Wich, Maximilian, Adrian Gorniak, Tobias Eder, Daniel Bartmann, Burak Enes Çakici, and Georg Groh (May 2022). "Introducing an Abusive Language Classification Framework for Telegram to Investigate the German Hater Community." In: *Proceedings of the International AAAI Conference on Web and Social Media* 16.1, pp. 1133–1144. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/19364>.

---

<sup>2</sup> in accordance with Exhibit 6 of the regulations for the award of doctoral degree

## ACKNOWLEDGMENTS

---

First and foremost, I would like to thank my supervisor Georg Groh who made this work possible. He was a valuable source of advice, expertise, and inspiration. I will miss our amusing discussions during lunch.

Many thanks to my colleagues from the Social Computing Group, especially to Tobias Eder and Edoardo Mosca, who started as master thesis students and became valuable colleagues.

Furthermore, I would also thank all the students who I supervised during the last two and a half years and who contributed to my research—in particular Hala Al Kuwatly, Daniel Bartmann, Jan Bauer, Melissa Breitingner, Burak Enes Çakici, Laurence Lerch, Svenja Räther, and Christian Widmer.

I would like to pay my special regards to all my other co-authors: Tonni Kiening, Anna Sophie Kümpel, Marlina Naimarevic, Jürgen Pfeffer, Diana Rieger, and Wienke Strathern.

I gratefully acknowledge the grant from the Hanns Seidel Foundation financed by the German Federal Ministry of Education and Research. It gave me the freedom to focus on my dissertation entirely.

Thanks to all my friends who directly and indirectly supported me by providing inspiration, advice, or pleasant distraction.

I would like to express my sincere gratitude to my parents Gudrun and Reinhardt for all their encouragement and support throughout my life.

Finally, special thanks to Julia for her love and continuous support—especially for getting through the pandemic and the several lockdowns with me. I'm looking forward to our next chapter.



## CONTENTS

---

1	INTRODUCTION	1
1.1	Motivation	1
1.2	Problem Statement	1
1.3	Research Objectives	3
1.4	Structure	4
2	BACKGROUND	5
2.1	Abusive Language	5
2.1.1	Terminology	5
2.1.2	Abusive Language Detection	6
2.1.3	Concrete Challenges	7
2.1.4	Abusive Language in Humanities and Other Disciplines	8
2.2	Explainable Artificial Intelligence	10
2.2.1	Terminology and Taxonomy	10
2.2.2	XAI relevant for Text Classification	11
2.2.3	XAI in Abusive Language Detection	11
3	METHODOLOGY	13
4	STUDIES	15
4.1	Descriptive Analysis of Hate Speech from Fringe Communities	16
4.1.1	Motivation	16
4.1.2	Study I <sup>†</sup>	16
4.1.3	Discussion	17
4.2	Bias and Comparison Framework for Abusive Language Datasets	19
4.2.1	Motivation	19
4.2.2	Study II <sup>•</sup>	20
4.2.3	Discussion	21
4.3	Annotator Bias in Abusive Language Datasets	23
4.3.1	Motivation	23
4.3.2	Study III: Annotators' Demographic Characteristics <sup>†</sup>	24
4.3.3	Study IV: Graph-Based Approach <sup>†</sup>	25
4.3.4	Study V: Bias Matrix <sup>•</sup>	26
4.3.5	Discussion	28
4.4	Political Bias in Abusive Language Datasets	30
4.4.1	Motivation	30
4.4.2	Study VI <sup>•</sup>	30
4.4.3	Discussion	32
4.5	Ethical Consideration of Annotator Bias	33
4.5.1	Motivation	33
4.5.2	Study VII <sup>†</sup>	33
4.5.3	Discussion	34
4.6	Integration of User and Network Data into Abusive Language Detection	36
4.6.1	Motivation	36

4.6.2	Study VIII: Abusive Language Dataset Containing Social Network Data •	37
4.6.3	Study IX: Integration of User Context in Hate Speech Detection †	38
4.6.4	Study X: Explainable Abusive Language Classification Leveraging User and Network Data •	39
4.6.5	Discussion	41
4.7	German Abusive Language Dataset Focusing on COVID-19	44
4.7.1	Motivation	44
4.7.2	Study XI •	44
4.7.3	Discussion	45
4.8	Investigation of the German Hater Community on Telegram	47
4.8.1	Motivation	47
4.8.2	Study XII †	47
4.8.3	Discussion	50
5	DISCUSSION	53
5.1	Limited Performance and Generalizability	53
5.2	Missing Transparency and Interpretability	55
6	CONCLUSION	57
Appendix		
A	PUBLICATIONS •	61
A.1	Study II	61
A.2	Study V	86
A.3	Study VI	99
A.4	Study VIII	112
A.5	Study X	119
A.6	Study XI	137
B	PUBLICATIONS †	145
B.1	Study I	145
B.2	Study III	161
B.3	Study IV	170
B.4	Study VII	181
B.5	Study IX	194
B.6	Study XII	208
	BIBLIOGRAPHY	223

## LIST OF FIGURES

---

Figure 2.1	Taxonomy of abusive language adapted from Poletto et al. (2021, p. 482). . . . .	5
Figure 3.1	Design science research cycles adapted from Alan R Hevner (2007, p. 88). . . . .	13
Figure 4.1	Overview of conducted studies mapped to research objectives. . . . .	15
Figure 4.2	Results of Heuristic Function (figures from Wich, Al Kuwatly, and Groh, 2020, p. 196 and 197). . . . .	26
Figure 4.3	Aggregated bias matrices for the selected datasets (figures and captions from Wich, Widmer, et al., 2021, p. 5). . . . .	27
Figure 4.4	Offensive tweet misclassified by right-wing classification model (figure from Wich, Bauer, and Groh, 2020, p. 62). . . . .	31
Figure 4.5	Dataset creation methodology (figure from Wich, Breiting, et al., 2021, p. 3). . . . .	37
Figure 4.6	Feature contribution based on Shapely values for the tweet " <i>&lt;user&gt; I think Arquette is a dummy who believes it. Not a Valenti who knowingly lies.</i> " (figure and parts of the captions from Mosca, Wich, and Groh, 2021, p. 95). . . . .	39
Figure 4.7	Visualized latent space of multi-modal model trained on Waseem and Hovy (2016); each dot represents a tweet, colored by label of the tweet (figure from Mosca, Wich, and Groh, 2021, p. 96). . . . .	40
Figure 4.8	Explanations for predictions of text, history, and network submodel in the form of Shapely values; red, positive values favor a classification as hateful; blue, negative values favor a classification as non-hateful (figures and captions from Wich, Mosca, et al., 2021, p. 12). . . . .	41
Figure 4.9	Dataset collection and annotation process (figure from Wich, Räther, and Groh, 2021, p. 2). . . . .	45
Figure 4.10	Topical similarity and connectedness in the network between seed Telegram channels (figure from Wich, Gorniak, et al., 2022, p. 7). . . . .	49

## LIST OF TABLES

---

Table 4.1	Bias and comparison framework for abusive language datasets (table from Wich, Eder, et al., 2021, p. 3). . . .	20
Table 4.2	Macro F1 scores from classifiers of the different annotator subsets (table and parts of the caption from Wich, Widmer, et al., 2021, p. 7). . . . .	28

## ACRONYMS

---

BERT	Bidirectional encoder representations from transformers
CNN	Convolutional neural network
GEF	Generative explanation framework
LIME	Local interpretable model-agnostic explanations
LRP	Layer-wise relevance propagation
LSTM	Long short-term memory
ML	Machine learning
NLP	Natural language processing
SHAP	Shapley additive explanations
t-SNE	T-distributed stochastic neighbor embedding
XAI	Explainable artificial intelligence



## INTRODUCTION

---

### 1.1 MOTIVATION

In the last decade, social media platforms (e. g., Twitter, Facebook, YouTube) have shaped how we communicate. Social media, for example, played an essential role during the Arab Spring (Howard et al., 2011). Additionally, it helped to make sexual harassment a subject of public discussion during the #metoo campaign (Hosterman et al., 2018). These platforms, however, also have a dark side. One part of the dark side is hate speech and other forms of abusive language. Due to the promise of freedom of expression (Chetty and Alathur, 2018), hate speech has widely spread with the rise of platforms (Duggan, 2017; Vogels, 2021). This could not only lead to a toxic atmosphere on social media but could also have more severe consequences (Kümpel and Rieger, 2019). Studies have shown that hate speech could be a threat to democracy (Schaez et al., 2020) and could even lead to physical hate crimes (Müller and C. Schwarz, 2020; Williams et al., 2020). Online hate is such a severe issue that governments passed new laws to cope with it. Germany, for example, introduced the Network Enforcement Act<sup>1</sup> in 2017 to take effective action against online hate speech (Baldauf, Ebner, and Guhl, 2019).

Despite the legislation and the willingness of platform providers to fight against hate speech and other forms of abusive language, the sheer volume of user-generated content makes manual monitoring impossible. Therefore, machine learning (ML) technologies enabling automatic detection of abusive language have been developed (Fortuna and Nunes, 2018; Mishra, Yannakoudakis, and Shutova, 2020; A. Schmidt and Wiegand, 2017). The problem, however, is that these technologies have severe drawbacks making them essentially only partially useful for the envisioned task despite groundbreaking advances in ML and natural language processing (NLP) in recent years. (MacAvaney et al., 2019; Vidgen, Harris, et al., 2019).

Nevertheless, research in the field of abusive language detection is crucial because reliable detection is a key component in the fight against abusive language. The advances made in recent years give hope that substantial improvements can be achieved by more research.

### 1.2 PROBLEM STATEMENT

Abusive language detection is facing a wide range of problems. The dissertation focuses on two of them: (1) limited performance and generalizability

---

<sup>1</sup> Netzwerkdurchsetzungsgesetz (NetzDG)

of classification models and (2) missing transparency and interpretability of classification models.

Concerning the first problem, state-of-the-art models have issues with reliably classifying abusive language. For the German language, for example, the best performing model from the GermEval 2019 Task 2 that distinguishes between offensive and non-offensive texts has an F1 score of 76.95%, while the precision of the offensive class is only 66.26% (Struß et al., 2019). Moreover, models often perform worse when they are applied to another dataset to test their generalizability (Arango, Pérez, and Poblete, 2019; Swamy, Jamatia, and Gambäck, 2019). One of the main reasons is the limited quality, quantity, and extension of the datasets available to the research community. Platform providers, such as Facebook, Twitter, and Google, possess the required amount and quality of data, but they are not willing to make them public. Leaving abusive language detection to these private companies jeopardizes our moral and technical independence. It is totally up to them to define the moral norms that underlie their abusive language detection systems—without any public participation. Therefore, we have to expand research activities to improve the classification models and training approaches that cope with the amount of data being available for the research community.

The second problem, missing transparency and interpretability, arises from the complexity of ML models that comes with the advances in this area. ML models have improved in recent years, but the predictions produced by these models are often not interpretable for humans anymore (Barredo Arrieta et al., 2020). This is a severe issue in a world where predictions produced by algorithms become more and more relevant for critical decisions (e. g., medical diagnostic or criminal prosecution) (Rudin and Ustun, 2018). Without interpretability, these models cannot earn trust regarding their functions and might therefore lose their social acceptance (Molnar, 2019; Rudin and Ustun, 2018). Furthermore, this black box characteristic may conceal unintended bias that the model learned from biased data. Inherently, bias in a dataset is required so that an algorithm can "detect and confirm patterns" (Hildebrandt, 2019, p. 103). Unintended bias, however, is a form of bias that compromises the generalizability of models (Geva, Goldberg, and Berant, 2019; Wiegand, Ruppenhofer, and Kleinbauer, 2019) and can cause unfair behavior of models (e. g., discriminating minorities or other groups of persons) (Dixon et al., 2018; Papakyriakopoulos et al., 2020). Consequently, the missing interpretability of models and their predictions hamper ensuring fairness, privacy, reliability, and causality of ML models (Doshi-Velez and Kim, 2017)—criteria that should be satisfied by every such system. In its white paper on artificial intelligence, the EU Commission has claimed these criteria, amongst others, as key requirements for artificial intelligence (Commission, 2020). Especially in the case of abusive language detection, interpretability is a crucial requirement that should be addressed by researchers. "The reason is the value-based nature of hate speech classification, meaning that perceiving something as hate depends on individual and social values and even the social values are non-uniform across groups and societies" (Wich, Bauer, and Groh, 2020, p.1).

Hence, a hate speech or in general a abusive language classifier should either be transparent or provide understandable explanations how the model come to a specific prediction (Kiritchenko, Nejadgholi, and Fraser, 2021; Vidgen, Harris, et al., 2019). The latter can be achieved by using explainable artificial intelligence (XAI). Since this is hardly addressed in research, it is a new and valuable research objective (Mishra, Yannakoudakis, and Shutova, 2020, 2021; Niemann, 2019; C. Wang, 2018).

### 1.3 RESEARCH OBJECTIVES

The dissertation's goal is to identify factors that address these two problems—(1) limited performance and generalizability and (2) missing transparency and interpretability. As both problems are comprehensive and hard to address as individual research objectives, they are broken down into the following five objectives:

The first three objectives aim to improve the performance and generalizability of abusive language classification models. The first one (A) focuses on increasing the quantity and quality of training data—in particular for other languages than English (Aken et al., 2018; Poletto et al., 2021; Vidgen and Derczynski, 2021). Secondly, the integration of context data (e. g., user or social network data) into the classification model (B) is investigated to provide additional features in order to produce more reliable predictions. Lastly, the third objective (C) aims to improve the generalizability of classification models.

The second problem (II) is addressed by two objectives. First, methods are investigated that help to discover unintended bias in training data and classifiers (D). This kind of bias is a severe issue because it can lead to unwanted and unfair model behavior (e. g., discriminating minorities) (Dixon et al., 2018; Papakyriakopoulos et al., 2020; Sap et al., 2019). Second, different methods for visualizing single predictions or an entire model are investigated and developed to better understand the models and their decisions and gain the users' trust (E).

Overall, the research focuses primarily on the German language because the progress in German abusive language detection research is limited compared to English. Therefore, it is essential to build up these competencies in the research community to gain technological and moral sovereignty and to prevent private companies from controlling these fields (e. g., Facebook, Twitter). In addition to German, some studies use English datasets due to the limited data availability.

These research objectives serve as a foundation for the twelve studies that were conducted in the context of this dissertation. The studies approached weak points of abusive language detection, which contributes to achieving these research objectives.

#### 1.4 STRUCTURE

The rest of the thesis is structured as follows. In Chapter 2, the current state of research on abusive language detection and XAI is briefly outlined. Chapter 3 deals with the underlying research methodology for the dissertation. In Chapter 4, the twelve conducted studies (I–XII) are described and discussed. The studies marked with • are relevant for examination; those marked with † are not formally relevant but contributed to the dissertation. Chapter 5 contains an overarching discussion outlining the studies' contribution to the research objectives of the thesis, followed by the conclusion (Chapter 6).

## BACKGROUND

---

### 2.1 ABUSIVE LANGUAGE

#### 2.1.1 Terminology

So far, there is no standard definition of hate speech or abusive language (Fortuna and Nunes, 2018; Poletto et al., 2021). Moreover, we can find many synonyms and related terms in literature, which are also not agreeably defined (e. g., offensive language, toxicity, cyberbullying, insults, harassment, flames) (Niemann et al., 2020; Poletto et al., 2021; A. Schmidt and Wiegand, 2017). A potential reason can be the complexity and multifacetedness of the terms, impairing a precise definition (Vidgen, Harris, et al., 2019; Waseem, Davidson, et al., 2017). Besides that, the definitions are often country-specific because some of them describe criminal acts, making them depend on the countries' legal frameworks.

For this dissertation, the framework proposed by Poletto et al. (2021) is applied to structure the different terms and concepts, as depicted in Figure 2.1. Abusive (or toxic) language is the overarching concept combining various facets, which is in line with other research (Nobata et al., 2016; Vidgen, Nguyen, et al., 2021; Waseem, Davidson, et al., 2017). It is defined as "hurtful language and includes hate speech, derogatory language, and also profanity" (Fortuna and Nunes, 2018, p. 8). Hate speech is defined as "an expression that is abusive, insulting, intimidating, harassing, and/or incites to violence, hatred, or discrimination. It is directed against people on the basis of their race, ethnic origin, religion, gender, age, physical condition, disability, sexual orientation, political conviction, and so forth" (Erjavec and Poler, 2012, p. 900). Some studies in the dissertation use slightly different definitions due to the used datasets and their labeling schema. But they fit under the umbrella of abusive language.

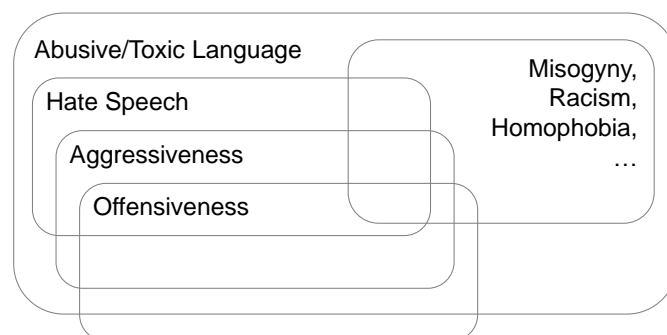


Figure 2.1: Taxonomy of abusive language adapted from Poletto et al. (2021, p. 482).

### 2.1.2 Abusive Language Detection

Abusive language detection or classification is a typical supervised learning task in ML. A function (model) is learned based on input-output pairs (training data) so that the function can infer the output from the input (Russell and Norvig, 2009). In the case of abusive language, the input is a text produced by a user on a social media platform, and the output is the label (e. g., abusive or non-abusive). The text can be, for example, a social media post (e. g., tweet), a comment, or a personal message. The challenge is to train a model so that it always infer the correct output/label.

The first abusive language classification model dates back to 1997 (A. Schmidt and Wiegand, 2017). Spertus (1997) combined manually created rules to generate features with a decision tree to classify messages as abusive or not abusive. Over time, abusive language classifiers became more sophisticated and performance increased. With the rise of neural networks, manual feature engineering approaches (e. g., lexicon-based features, bag-of-words features) became largely obsolete. Djuric et al. (2015) used the neural network-based embedding approach *Paragraph Vector*, also known as *paragraph2vec*, (Le and Mikolov, 2014) to create document embeddings. With the rise of deep learning, researchers started to adopt these methods for abusive language detection. Gambäck and Sikdar (2017) proposed an abusive language classifier based on word embeddings and convolutional neural networks (CNNs). Badjatiya et al. (2017) examined and compared various deep learning architectures using word embeddings, CNNs, and long short-term memory networks (LSTMs). A milestone in NLP that also strongly influenced abusive language detection was the language representation model BERT, which stands for *Bidirectional Encoder Representations from Transformers* (Devlin et al., 2019). BERT models are pre-trained on large corpora using an automatic, intrinsic supervision approach. The pre-trained models are then fine-tuned for the actual tasks (e. g., sentiment classification, abusive language classification), requiring only a fraction of the labeled training data compared to classical deep learning approaches (Devlin et al., 2019). Meanwhile, classification models based on BERT or other transformer models are state of the art in abusive language classification and dominate the rankings of recent shared tasks (Mandl, Modha, Kumar M, et al., 2020; Risch, P. Schmidt, and Krestel, 2021; Zampieri et al., 2019). That is the reason why nearly all classification models in this research project are based on transformer models. Beyond that, these types of models allow us to build multilingual abusive language classification models, which is helpful for languages with limited training data (Aluru et al., 2021; Ranasinghe and Zampieri, 2020; Stappen, Brunn, and Schuller, 2020). They used Transformer models that were pretrained with data in multiple languages—e. g., mBERT, which was pretrained on data from Wikipedia in 104 languages (Aluru et al., 2021). However, these cross-lingual approaches have limitations, especially in the case of language-specific hate (Nozza, 2021).

Besides focusing on the textual features, researchers sought for additional features to identify abusive language (Mishra, Yannakoudakis, and Shutova, 2020; A. Schmidt and Wiegand, 2017). These features, for example, comprise user profile data, such as gender, account age, number of followers (Fehn Unsvåg and Gambäck, 2018; A. M. Founta et al., 2019), social network data (Mishra, Del Tredici, et al., 2018, 2019; Vijayaraghavan, Larochelle, and Roy, 2019), or the post or comment history of the user (Dadvar et al., 2013; Pitsilis, Ramampiaro, and Langseth, 2018; Rangel et al., 2021). In Study IX and X, we seized this idea and developed abusive language classification models that leverage the social network data and previous posts of the users.

Instead of classifying messages or posts, some researchers focused on account level and developed models to predict whether a Twitter account is a hater (Das et al., 2021; Li et al., 2021; M. Ribeiro et al., 2018). That makes the identification of hateful actors independent from textual data, which can be helpful for subtle or implicit forms of abusive language. We applied a similar approach on Telegram, a messaging platform with a social media component, in Study XII.

Despite many research activities, abusive language classification is not yet a solved problem. In the following subsection, further challenges are elaborated, which abusive language detection is facing.

### 2.1.3 Concrete Challenges

One challenge arises from the complexity and variety of abusive language, impairing the development of an universal classifier. As mentioned, there are different forms, types, and targets (Vidgen, Nguyen, et al., 2021; Waseem, Davidson, et al., 2017). To gain a better understanding of this complexity, we conducted an in-depth analysis of hate speech from alt-right fringe communities in Study I. Furthermore, language is not a static construct; it evolves and produces new abusive terms at short intervals (Raisi and B. Huang, 2016). Current trends and events, such as the refugee crisis (Ross et al., 2016) or the COVID-19 pandemic (Vidgen, Hale, et al., 2020), influence abusive language. That is why we developed an approach to collect abusive language datasets with a topical focus and applied it to build a German COVID-19 dataset in Study XI.

A set of challenges is associated with datasets used to train abusive language classifiers. First of all, there is no standard abusive language dataset for benchmarking classifiers (A. Schmidt and Wiegand, 2017), exacerbating the comparison of different models. Secondly, available datasets are degrading because researchers sometimes only publish references to the comments or posts due to platform policies (e. g., tweet IDs instead of the tweets). Some of these are removed over time because of their violating nature (Vidgen, Harris, et al., 2019), resulting in shrinking datasets. Thirdly, the annotation quality varies or is not specified (e. g., inter-rater reliability) (Aken et al., 2018; Vidgen, Harris, et al., 2019). That is connected to the first challenge because manually annotating abusive language is not a trivial task due to its complexity.

Fourthly, available abusive language datasets differ widely in size, labeling schema, class balance, and quality (Vidgen and Derczynski, 2021; Vidgen, Harris, et al., 2019). A factor impairing the quality of a dataset is unintended bias because it can lead to unwanted and unfair model behavior (Dixon et al., 2018; Papakyriakopoulos et al., 2020). Examples already investigated are topic bias (Wiegand, Ruppenhofer, and Kleinbauer, 2019), annotator bias (Binns et al., 2017; Larimore et al., 2021; Waseem, 2016), and racial bias (Davidson, Bhattacharya, and Weber, 2019; Sap et al., 2019). Therefore, it is necessary to identify unintended bias and remove it to develop trustworthy and fair models.

Several studies in the dissertation address this set of challenges. In Study II, a bias and comparison framework for abusive language datasets was developed. It helps to compare such datasets and to uncover several forms of bias in them. Studies III, IV, and V investigated annotator bias from various perspectives because it has not received much attention from the research community yet. A form of bias in abusive language datasets that has not been examined is political bias, which Study VI dealt with.

Another challenge is the missing interpretability of most of the classification models, particularly deep learning models (Kiritchenko, Nejadgholi, and Fraser, 2021; Mishra, Yannakoudakis, and Shutova, 2020; Vidgen, Harris, et al., 2019; C. Wang, 2018). Since these models behave like black boxes, it is a huge or nearly unsolvable challenge for a human to understand how the model came to a particular decision. Nevertheless, such insights would help to improve the models, to identify bias and unfair behavior, and to build trust. Therefore, Study II, VI, IX, and X used XAI techniques to explain predictions and to make bias in models visible. Further details on these studies can be found in Section 2.2.

#### 2.1.4 *Abusive Language in Humanities and Other Disciplines*

Abusive language is a phenomenon that is not only investigated by computer science but also by many other disciplines. Research from these areas can provide valuable insights for our discipline as well as for this research endeavor. In turn, those disciplines can benefit from our results.

One that investigates abusive language from various perspectives is social science. Its researchers, for example, study the targets of abusive language, its perception, its impact, and counter-measures against it (Kümpel and Rieger, 2019). Obermaier, Hofbauer, and Reinemann (2018) showed that hate speech targeting journalists is a growing problem in Germany, which could damage society's sentiment towards them. Regarding the impact, Müller and C. Schwarz (2020) found "that social media can act as a propagation mechanism for violent crimes by enabling the spread of extreme viewpoints" (Müller and C. Schwarz, 2020, p. 1). In regards to counter-measures, a broad range is proposed starting from community management to counter speech and counter-narratives to encouraging media competence (Kümpel and Rieger, 2019). The findings of this discipline are relevant in two ways for this research



project. First, they create a better and broader understanding of abusive language's complexity and diversity (e. g., various targets and forms). Second, they underline the urgency of developing reliable classification models to support counter-measures—automatic detection is one key component in this context (e. g., in community management).

Law and criminology also deal with abusive language. One of their research objectives in this context is the effective prosecution of hate crimes. Due to the amount and global nature of abusive language as well as the anonymity of the internet, law enforcement is impaired. Therefore, Banks (2010) claimed a combination of legal and technological regulations to take action against hate speech and other forms of abusive language. In particular, the technological part is relevant for this research because it relies on capabilities, such as monitoring of websites and platforms and filters, that require abusive language classification models.

Another study from this field investigated the relationship between online and offline hate crimes, showing a direct link between them (Williams et al., 2020). It confirms the findings from Müller and C. Schwarz (2020) and emphasizes the necessity of developing such detection systems.

An open question highly relevant for all disciplines is how to deal with online abusive language. Shall it be deleted, marked, quarantined, countered, or tolerated? Ullmann and Tomalin (2020), for example, suggested a quarantining approach to retain freedom of speech and to protect users from harmful content concurrently. Yong (2011) argued that not all types of hate speech should be regulated as it may violate freedom of speech. In the case of a stricter regulation of hate speech, an approach that only focuses on one platform could lead to a shift of hate speech to less controlled platforms (Johnson et al., 2019). It is a complex question with a crucial impact on productive abusive language detection systems. Unfortunately, there is no answer to this question yet. However, the quarantining approach introduces an interesting concept. It allows multiple perspectives on abusive language. One user could have a more liberal perspective, while another user is stricter (Ullmann and Tomalin, 2020). This concept of multiple perspectives was examined by Study IV, V, and VI. Moreover, Study VII adds an ethical discussion about annotations influenced by the subjective perception of annotators.

The listed examples are not intended to be exhaustive. However, they show the interdisciplinary nature of the topic and how the conducted studies are related to research from other disciplines.

## 2.2 EXPLAINABLE ARTIFICIAL INTELLIGENCE

### 2.2.1 Terminology and Taxonomy

With the increasing adoption of ML in our daily life, concerns about the black box nature of these models and the demand for understandable predictions have been raised—especially in risky use cases (e. g., medical diagnosis). Hence, the field of XAI has emerged to address this shortcoming. Since XAI is relatively new and covers a broad spectrum, it lacks commonly defined terms and taxonomy. Therefore, many related terms and concepts can be found in the literature (e. g., understandability, comprehensibility, interpretability, explainability, transparency) (Adadi and Berrada, 2018; Barredo Arrieta et al., 2020; Lipton, 2018). They are sometimes interchangeably used even though they have different meanings, particularly in terms of interpretability and explainability (Barredo Arrieta et al., 2020). Interpretability describes "a passive characteristic of a model referring to the level at which a given model makes sense for a human observer" (Barredo Arrieta et al., 2020, p.4). In contrast to that, explainability refers to "an active characteristic of a model, denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal functions" (Barredo Arrieta et al., 2020, p.5).

Within the scope of the dissertation, XAI is defined as proposed by Barredo Arrieta et al. (2020): "Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand" (Barredo Arrieta et al., 2020, p.6). This definition was chosen because the purpose of XAI in the context of abusive language detection is to provide the user information that facilitates subsequent actions (e. g., rephrasing or deleting a post).

According to Barredo Arrieta et al. (2020), XAI can be split into two groups: (1) transparent models and (2) post-hoc explainability. The first one refers to ML models that are intrinsically understandable (e. g., decision trees) (Barredo Arrieta et al., 2020; Molnar, 2019). The latter refers to techniques that explain non-understandable ML models (Barredo Arrieta et al., 2020; Molnar, 2019). Since the current state-of-the-art approaches for abusive language detection (e. g., deep learning) are far beyond transparent models, the dissertation focuses on post-hoc explainability.

Post-hoc explainability can be further divided according to three dimensions. The first one describes the level of explanation: Is only an explanation for a single prediction (*local*) or for the entire model (*global*) provided? The second dimension distinguishes between *model-specific* and *model-agnostic* methods. While the use of the first group depends on the ML models that are supposed to be explained, the second group works for all models. The third dimension refers to the *form of explanation* for the user (e. g., text explanation, explanation by example, visual explanation) (Barredo Arrieta et al., 2020; Molnar, 2019).

### 2.2.2 XAI relevant for Text Classification

We can find a range of different XAI approaches for text classification in literature—mostly local ones. The following list contains local post-hoc techniques that can be used to assign to each word of a document a relevance score for the classification: *LRP* (*Layer-wise Relevance Propagation*) (Bach et al., 2015), *LIME* (*Local Interpretable Model-agnostic Explanations*) (M. T. Ribeiro, Singh, and Guestrin, 2016), *anchors* (M. T. Ribeiro, Singh, and Guestrin, 2018a), *DeepLIFT* (*Deep Learning Important FeaTures*) (Shrikumar, Greenside, and Kundaje, 2017; Shrikumar, Greenside, Shcherbina, et al., 2017), *LIMSSE* (Poerner, Schütze, and Roth, 2018), and *SHAP* (*SHapley Additive exPlanations*) (Lundberg and Lee, 2017). Another approach is the *Generative Explanation Framework* (*GEF*) proposed by Liu, Q. Yin, and W. Y. Wang (2019). They built a system to generate reasonable explanations for predictions (e. g., text justifications) (Liu, Q. Yin, and W. Y. Wang, 2019). An alternative approach is to generate adversarial examples to provide insights into how a model comes to a particular prediction (Ebrahimi et al., 2018; M. T. Ribeiro, Singh, and Guestrin, 2018b). For this research project, we used the SHAP framework as local explainability framework (see Study II, VI, IX, and X). It provides well understandable explanations and is compatible to transformer-based models, which are used in several studies (Lundberg and Lee, 2017).

In regards to global approaches, the number of techniques for NLP is limited (Danilevsky et al., 2020)—presumably due to the complexity of the models and natural language. M. T. Ribeiro, Singh, and Guestrin (2016), for example, introduced SP-LIME together with LIME that provides a global view of a model by selecting appropriate and representative predictions. Another approach is to aggregate local explanations to get an understanding of the model’s reasoning (Linden, Haned, and Kanoulas, 2019).

Concerning both use cases, local and global, the methods are in the fledgling stage of development and require more research to be used in production.

### 2.2.3 XAI in Abusive Language Detection

Similar to XAI for text classification, XAI has rarely been studied by the research in abusive language detection, even though there is a demand for it (Kiritchenko, Nejadgholi, and Fraser, 2021; Mishra, Yannakoudakis, and Shutova, 2021; Niemann, 2019; Vidgen, Harris, et al., 2019). Only a few studies have focused on this topic. MacAvaney et al. (2019) developed a transparent abusive language classification model based on support vector machines. C. Wang (2018) applied techniques from computer vision to visualize predictions of abusive language classifiers based on neural networks. Švec et al. (2018) used the LRP-based approach from Arras et al. (2017) to build a Slovak abusive language classifier. Vijayaraghavan, Larochelle, and Roy (2019) applied LIME to their hate classifier, which uses network data besides textual data as input, to explain predictions. Risch, Ruff, and Krestel (2020) compared

different transparent models and models combined with explainability approaches. Mathew et al. (2021) examined various hate speech classification models combined with LIME and attention mechanism as XAI method to improve interpretability and reduce unintended bias.

The listed studies are not intended to be exhaustive. But it reflects the current state of research on this topic, underlining the demand for further research. This gap is addressed by the following four studies of the dissertation: Study II and VI applied XAI technique to uncover and visualize unintended bias, similarly to Mathew et al. (2021). Study IX and X combined multi-modal classification models using text and social network data with XAI techniques to provide insights on the prediction results.

## METHODOLOGY

The dissertation follows the design science methodology combined with empirical research (A. Hevner and Chatterjee, 2010; Alan R. Hevner et al., 2004). Design science focuses on developing "innovative artifacts to solve real-world problems" (A. Hevner and Chatterjee, 2010, p.9), as this research intends to do.

Every design science research endeavor consists of three cycles—*relevance cycle*, *rigor cycle*, and *design cycle*—that engage with each other and are conducted at least once (A. Hevner and Chatterjee, 2010; Alan R Hevner, 2007), as depicted in Figure 3.1. The *relevance cycle* connects the real-world problem and its environment with the design activities. The *rigor cycle* establishes a connection between the design activities and the knowledge base (e. g., providing theories and methods) for the design process and as a target for generated results. Within the *design cycle*, the artifact (e. g., theory, software, classification model) is iteratively built and evaluated based on the input from the *relevance* and *rigor cycles*.

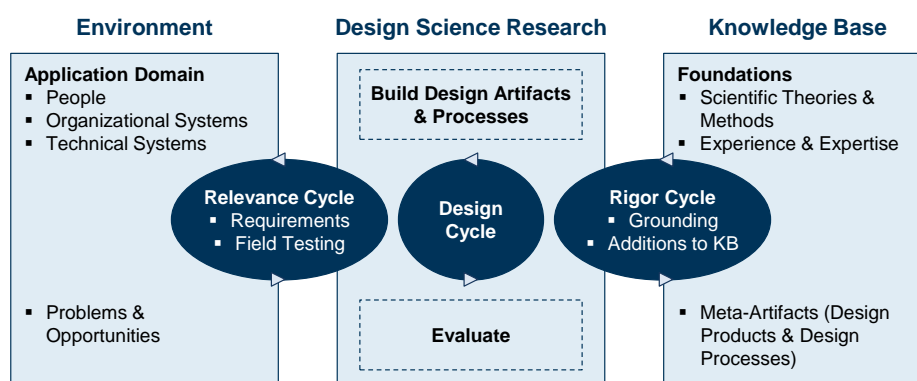


Figure 3.1: Design science research cycles adapted from Alan R Hevner (2007, p. 88).

For the *relevance cycle* of this research, the environment was abusive language on social media and its users. The problems to be solved were (1) deficient performance and generalizability of abusive language classifiers and (2) missing interpretability of these classifiers. Concerning the *rigor cycle*, the knowledge base comprised the state of research from different disciplines including but not limited to computer science as researchers from various fields have investigated abusive language (e. g., computer linguistics, communication science, political science, philosophy). Based on these inputs from the environment and the knowledge base, abusive language classification models were built and evaluated in the *design cycle*. The results of the design process contribute to the knowledge base through adding the generated knowledge

(*rigor cycle*) and to the application domain through IT system components that solve the problem (*relevance cycle*).

For this research project, a set of design science research cycles was conducted as the goal was to identify factors solving the two stated problems. These different sets of cycles arose from various conducted studies, which are described in the following section. Depending on the study, empirical methods were integrated in the cycles (e. g., annotating text corpus, developing classification models).

## STUDIES

Twelve studies were conducted to achieve the research objectives and address the weak points of abusive language classification. Figure 4.1 provides an overview of the studies and their contribution to the research objectives.

		1 Limited performance and generalizability			2 Missing transparency and interpretability	
		A Increasing quantity and quality of training data	B Integrating context data into classification	C Improving generalizability of classification models	D Uncovering unintended bias within training data and models	E Making predictions of models more understandable for humans
		<ul style="list-style-type: none"> <li>● Fully addressed</li> <li>◐ Partially addressed</li> <li>○ Not addressed</li> </ul>				
Study						
I	Descriptive analysis of hate speech from fringe communities	●	○	○	○	○
II	Bias and comparison framework for abusive language datasets	◐	○	○	●	◐
III	Identification of annotator bias based on annotators' demographic characteristics	○	○	○	●	○
IV	Investigation of annotator bias with a graph-based approach	○	○	○	●	○
V	Investigation of annotator bias with bias matrix	○	○	○	●	○
VI	Political bias in abusive language datasets	○	○	○	●	◐
VII	Ethical consideration of annotator bias	○	○	○	●	○
VIII	Abusive language dataset containing social network data	●	◐	○	○	○
IX	Integration of user context in hate speech detection	○	●	○	○	●
X	Explainable abusive language classification leveraging user and network data	○	●	○	○	●
XI	German abusive language dataset focusing on COVID-19	●	○	○	○	○
XII	Investigation of the German hater community on Telegram	●	●	●	○	○

Figure 4.1: Overview of conducted studies mapped to research objectives.

#### 4.1 DESCRIPTIVE ANALYSIS OF HATE SPEECH FROM FRINGE COMMUNITIES

##### 4.1.1 *Motivation*

Abusive language is a complex and manifold construct that can have various forms (Vidgen, Harris, et al., 2019; Waseem, Davidson, et al., 2017). It can target an individual, identity (e. g., Republicans, Muslim), or concept (e. g., capitalism) (Vidgen, Harris, et al., 2019). Furthermore, the sender can use various kinds to articulate abuse: "hatefulness, aggression, insults, derogation, untruths, stereotypes, accusations, and undermining comments" (Vidgen, Harris, et al., 2019, p. 82). Based on this variety, it becomes clear why detecting abusive language is such a challenging task.

Therefore, it was necessary to conduct an in-depth examination of abusive language in the first study to better understand this phenomenon's complexity and provide a basis for the other studies.

Instead of using one of the traditional social media platforms as data source for the study (e. g., Facebook, Twitter), we decided to collect data from alt-right communities on Reddit, 4chan, and 8chan. There were multiple reasons for this decision. First, the content on these platforms is hardly or not moderated in contrast to the traditional platforms (Arthur, 2019; Horta Ribeiro et al., 2021; Knuttila, 2011), leading to a more considerable amount and variety of abusive content. Secondly, the alt-right movement and especially the communities on 4chan and 8chan are or were known for hateful content (Hine et al., 2017; Tuters and Hagen, 2020; Wong, 2019). The shooters from Christchurch, Newzealand, and from Poway, California, in 2019, for example, used 8chan to publish their racist manifestos. A further reason for selecting these platforms was that hardly any abusive language dataset contains comments from these platforms, making the annotated dataset a valuable contribution to the research community.

##### 4.1.2 *Study I<sup>†</sup>*

In the study "Assessing the Extent and Types of Hate Speech in Fringe Communities: A Case Study of Alt-Right Communities on 8chan, 4chan, and Reddit" (Rieger et al., 2021), we conducted an empirical analysis of hate speech collected from three different alt-right fringe communities—"Reddit (r/The\_Donald), 4chan (4chan/pol/), and 8chan (8chan/pol/)" (Rieger et al., 2021, p. 2). The goal was to understand better hate speech and its forms that users of these communities spread. For that reason, the following research questions were addressed by the study:

RQ1 "What percentage of user comments in the three fringe communities contains explicit or implicit hate speech?" (Rieger et al., 2021, p. 4)

RQ2 "a) In which ways is hate speech expressed, and b) against which persons/groups is it directed?" (Rieger et al., 2021, p. 4)



RQ3 "What is the topical structure of the coded user comments?" (Rieger et al., 2021, p. 4)

To answer the research questions, we collected 149,504 comments from the three discussion boards in April 2019 and annotated 6,000 of them based on a fine-grained, multidimensional schema. The schema includes, among others, the following dimensions: form of hate speech (direct and indirect), type of hate speech (e. g., threat of violence, negative stereotyping, inhuman ideology), and hate speech targets (e. g., Jews, Muslims, African-Americans). Based on the granularly annotated comments, we addressed RQ1 and RQ2 with "a manual quantitative content analysis" (Rieger et al., 2021, p. 2). To answer RQ3, we applied topic modeling on the annotated data—an unsupervised ML approach to identify topics within a dataset. The created topic model was then analyzed together with the hate speech dimensions.

A finding concerning RQ1 was that 24% of the annotated comments contained implicit or explicit hate speech. The prevalence of hate speech varies between the three studied communities. The most hateful community was 8chan/pol/ with 34.4%, followed by 4chan/pol/ with 24.0% and r/The\_Donald/ with 13.8%. Regarding RQ2a, the two most dominant forms of hate speech were general insults (39.6% within hateful comments) and disinformation and conspiracy theories (31.8%). Concerning the targets of hate speech (RQ2b), we observed that hate speech most often targets Jews (31.3%), black people (18.1%), and political opponents (15.6%). The results from topic modeling (RQ3) were in line with the ones from RQ2. Notably but not surprisingly, political topics were more prevalent on r/The\_Donald/ than on the other two boards. More radical topics exhibited a higher prevalence on 8chan/pol (religious topic, such as anti-Semitic) and 4chan/pol/ (topic related to slur words).

The study provided a systematic, in-depth analysis of hate speech in alt-right communities. It showed that hate speech is a comprehensive term with numerous facets, aggravating its automatic detection. A further contribution was the granularly annotated hate speech dataset that can serve other researchers as a research object.

#### 4.1.3 Discussion

The goal of the empirical analysis was to provide a better understanding of hate speech's complexity. Even though the analyzed communities are or were part of the alt-right movement, meaning that the expressed hate should be similar, we observed a wide variety concerning type, target, and topic. This diversity exacerbated the detection of hate speech or other forms of abusive language because a ML model would require a vast amount of training data to learn the complexity. Additionally, it needs to be said that the analyzed comments were only a small sample from one radical group, meaning that the actual variety can be assumed to be much larger.

Another interesting finding is the observed hate speech prevalence of 24% in the analyzed communities. This is a very high value, considering that the overall share of abusive content on Twitter, for example, is estimated to be up to only 3% (A. Founta et al., 2018). The explanation is that these communities are a melting pot for members of the alt-right movement who cannot arbitrarily spread their hateful messages on traditional social media platforms (e. g., Facebook, Twitter) due to violating the platform policies. An observation that confirms this thesis is the following: the prevalence of hate speech on Reddit, a platform with content moderation to some extent, is lower than the one from the unmoderated communities on 4chan and 8chan.

Since the data collection in April 2019, the subreddit */r/The\_Donald* and the platform 8chan have been closed—*/r/The\_Donald* because of the hateful content (Roose, 2020) and 8chan due to its contribution to several mass shootings (Prince, 2019). Only few months later, 8chan was relaunched as 8kun in autumn 2019—this time without the controversial and toxic board */pol/* (Glaser, 2019). However, many users migrated to other platforms (e. g., Telegram, Discord) that are less moderated (Glaser, 2019). Similar behavior was observed when Reddit closed */r/The\_Donald* (Horta Ribeiro et al., 2021). A part of the community migrated to its own platform—first *TheDonald.win*, then *Patriots.win* (Horta Ribeiro et al., 2021; Timberg and Harwell, 2021). Horta Ribeiro et al. (2021) found out that not all members migrated, but the ones who migrated became more radicalized. The phenomenon of banning users or communities from social media platforms due to violation of platform policies is called deplatforming (Fielitz and K. Schwarz, 2020; Rogers, 2020). The consequences are often the same as in the cases of 8chan and */r/The\_Donald*: the communities move to alternative platforms with less or no moderation activities (Fielitz and K. Schwarz, 2020; Rogers, 2020). Therefore, it is arguable whether banning users or communities is the right way to counter hate speech.

A limitation of the empirical analysis is that we analyzed only the textual parts of the comments. However, photos, especially memes, and videos are used to transport messages in these communities (Tuters and Hagen, 2020). Therefore, it could be interesting to integrate also media files in such an analysis as part of future work.

Nevertheless, the empirical analysis helps to understand the complexity of abusive language better. In addition, the annotated hate speech dataset is a valuable contribution to the community because it has a granular, multidimensional labeling schema and contains comments from platforms that are hardly covered by abusive language datasets (Madukwe, X. Gao, and Xue, 2020; Poletto et al., 2021; Vidgen and Derczynski, 2021).

## 4.2 BIAS AND COMPARISON FRAMEWORK FOR ABUSIVE LANGUAGE DATASETS

### 4.2.1 *Motivation*

In recent years, researchers have released many abusive language datasets to advance research activities in this area (Madukwe, X. Gao, and Xue, 2020; Poletto et al., 2021; Vidgen and Derczynski, 2021). To keep track of the dataset and make the data more accessible to researchers, Vidgen and Derczynski (2021) published an online catalog for hate speech datasets<sup>1</sup>, which currently lists more than 60 datasets from different languages.

The variety of datasets was the reason why researchers published survey papers to compare the datasets and to provide fellow researchers an overview (Madukwe, X. Gao, and Xue, 2020; Poletto et al., 2021; Vidgen and Derczynski, 2021). These surveys are helpful because the datasets strongly differ in various aspects. First, they used different tasks to label the data, leading to different labeling schema. The tasks vary in "the nature of abuse" (Vidgen and Derczynski, 2021, p. 5) and "the granularity of taxonomies" (Vidgen and Derczynski, 2021, p. 7). The first one refers to the question of what abusive content targets or attacks (e. g., racism, misogyny, Anti-Semitism, cyberbullying, offensive speech) (Vidgen and Derczynski, 2021). The second one refers to the structure of the labeling schema (e. g., binary, multi-class, multi-label, hierarchical structure) (Vidgen and Derczynski, 2021). Second, data is collected from different sources (e. g., comment sections from news websites, Facebook, Twitter, YouTube, Reddit), which strongly influences the structure of the text documents. Third, the approaches differ how the data is gathered from the data sources (e. g., keyword-based querying, random sampling).

All surveys are linked by the common fact that they do not look into the data. They only compare the datasets on a meta-level (e. g., source, size, language, labeling schema, annotation process)—with a small exception: Poletto et al. (2021) conducted a small and limited lexical analysis on the data. But why is it necessary to look under the surface of these datasets? A classification model trained on a biased dataset might learn this bias, which can lead to unfair and, in the worst-case, discriminatory decisions (Dixon et al., 2018; Papakyriakopoulos et al., 2020). Arango, Pérez, and Poblete (2019), for example, showed that state-of-the-art classifiers can learn to identify users based on the writing style instead of offensive language. Consequently, the classifier performs well on one dataset, but poorly on another that contains tweets from other users. Such an effect is caused by the fact that a considerable portion of tweets in the training set is posted by a small amount of users (Arango, Pérez, and Poblete, 2019; Wiegand, Ruppenhofer, and Kleinbauer, 2019). Another issue can be topic bias, causing a classifier to recognize topics instead of offensive language (Wiegand, Ruppenhofer, and Kleinbauer, 2019). Therefore, it is necessary to have a framework for abusive language datasets

---

<sup>1</sup> <https://hatespeechdata.com/>

Table 4.1: Bias and comparison framework for abusive language datasets (table from Wich, Eder, et al., 2021, p. 3).

Perspective	Method	Problem
1. Meta	a) Class distribution and availability	Degradation
	b) Time distribution	Temporal bias
	c) Pareto analysis of authors	Author bias
2. Semantic	a) LSI-based intra-dataset class similarity	Similarity/dissimilarity of classes
	b) Word embedding based intra- and inter-dataset class similarity	Similarity/dissimilarity of classes
	c) Cross-dataset topic model	Topic bias
	d) PMI-Based word ranking for class	Topic bias
3. Annotation	a) Distribution of inter-rater reliability	Annotator bias
4. Classification	a) Cross-dataset performance	Generalizability
	b) Explainable classification models	Generalizability

to systematically examine datasets regarding different forms of bias and to compare them with each other.

#### 4.2.2 Study II •

In the study “Bias and comparison framework for abusive language datasets” (Wich, Eder, et al., 2021), the need for a framework to systematically analyze abusive language datasets and to compare them was addressed by developing such a framework and applying it on eleven datasets. To demonstrate its language independence, five of the eleven datasets were in English, the other six were in Arabic.

The framework was developed based on findings from other studies. They helped to identify relevant forms of bias (e. g., topic, author, and temporal bias) and additional factors that can influence classification models (e. g., limited availability of training data due to dataset degradation). The resulting framework consists of four perspectives comprising ten methods to examine abusive language datasets. Table 4.1 provides an overview including a mapping describing which problem is addressed by which method.

Applying the framework to the eleven datasets revealed interesting insights. For example, the degradation of the datasets that do only contain an document identifier referring to the online resource (e. g., the tweet ID and not the text of the tweet) is on average around 41% at the time of inquiry. That means that researchers can only work with a portion of the complete dataset if the dataset’s author published only the reference. But why do researchers only publish the IDs and not the actual text? It is a result of the platforms’ policies. Twitter, for instance, allows only publishing the tweet IDs and not the actual text of a tweet. A further finding concerns the topical focus of

the datasets. While political and religious topics are dominant in the Arabic datasets, the topics in the English datasets are more diverse with specific focuses of some datasets (e. g., misogyny, xenophobia, insults, American politics, and COVID-19). Overall, no dataset stands out in particular. Every dataset has its advantages and disadvantages. However, the English dataset released by A. Founta et al. (2018) is a diverse and comprehensive dataset showing decent performance with respect to training generalizable classifiers. Regarding the Arabic datasets, the datasets from Mubarak et al. (2020) and Alsafari, Sadaoui, and Mouhoub (2020) made a positive impression due to their size, compatible labeling schema, and good generalization results.

The framework is helpful for data scientists and researchers that either build new abusive language datasets or use existing ones. Regarding the first group, the awareness of the forms of bias and additional influencing factors can help them avoid or mitigate these pitfalls and increase the quality of their datasets. Regarding the second group, the framework supports selecting appropriate datasets and the analysis of their findings in case of using several abusive language datasets.

### 4.2.3 Discussion

The study showed that the abusive language datasets exhibit considerable differences that are hidden under the surface. Therefore, such a framework as proposed by the study is a useful tool to reveal such peculiarities and ensure comparability.

The proposed framework is not the first one addressing the missing comparability and the need for better and consistent documentation of a dataset's characteristics. Bender and Friedman (2018) and Gebru et al. (2020) suggested datasheets to approach these problems. However, our proposed framework is not supposed to be seen as an alternative. It is rather an extension to these datasheets that deal with the special characteristics of abusive language datasets because the datasheets are designed to be as generic as possible.

Even if the framework provides a broad spectrum of various analysis methods, there is room for improvement as part of future work. A form of bias that could receive more attention in the framework is annotator bias. Annotators introduce this bias due to their subjective perception of what abusive language is (Wich, Bauer, and Groh, 2020). The framework contains one method covering this bias. But it has limited expressiveness and comparability. In addition, only one of eleven datasets provided the raw annotations—the mandatory requirement for analyzing this form of bias. A useful extension would be the bias matrix that quantifies the annotator bias and is proposed by Wich, Widmer, et al. (2021). The method is described and discussed in Section 4.3.4. Another improvement could be to integrate a benchmark dataset, such as proposed by Mathew et al. (2021) and Röttger et al. (2021), into the cross-dataset performance method of the framework. These benchmark datasets can be used as test sets for the models trained on the abusive language datasets that are examined. By doing so, the datasets

can be evaluated and compared on a fine-grained level. HateCheck, for example, reveals how well a model detects hate against particular target groups (e. g., women, Muslims, or immigrants) (Röttger et al., 2021). The downside, however, would be that integrating one of the datasets impairs the multilingual capabilities of the framework because they are only available in English at the moment. Beyond adding new methods to the framework, the framework and the usefulness of the individual methods should be evaluated by potential users of the framework. Such a study could help to assess the ease of use and to identify gaps and further potential for improvement.

## 4.3 ANNOTATOR BIAS IN ABUSIVE LANGUAGE DATASETS

### 4.3.1 *Motivation*

In the discussion part of the previous section, it was mentioned that the framework could benefit from methods facilitating the analysis of annotator bias. This type of bias has not received much attention in the framework because the analyzed datasets did not provide the necessary raw data to explore it. Furthermore, the number of studies investigating this form of bias and its impact is limited. This gap is addressed by the following three studies that are presented and discussed in this section.

Before diving into the studies, annotator bias is defined and explained. It is a form of bias "caused by the subjective perception and different knowledge levels of annotators regarding the annotation task" (Wich, Al Kuwatly, and Groh, 2020, p. 191). An annotator, for example, could have a more liberal perspective on abusive language because freedom of speech is more relevant in his or her opinion. Or it is the opposite: he or she has a more authoritarian or stricter attitude.

There is a limited number of studies that examined annotator bias in the context of abusive language detection. According to Waseem (2016), models trained on annotations provided by experts perform better than by crowd workers. Ross et al. (2016) underlined the relevance of detailed annotation guidelines to improve the consistency of annotations and consequently the models' performance. Binns et al. (2017) observed that models trained on data labeled by men perform differently than ones trained on data annotated by women. Sap et al. (2019) found that models trained on two commonly used abusive language datasets tend to label posts in African American dialect as offensive. They traced their observation to the annotators' missing sensitivity towards the African American dialect. Larimore et al. (2021) also investigated racial annotator bias. They observed "that White and non-White annotators exhibit significant differences in ratings when reading tweets with high prevalence of certain racially-charged topics" (Larimore et al., 2021, p. 81). Akhtar, Basile, and Patti (2020) split annotators into groups based on the polarization of the annotations. Kanclerz et al. (2021), Davani, Díaz, and Prabhakaran (2021), and Kocoń et al. (2021) proposed to train classification models representing an individual or group-specific perspective. Such approaches can be used as filters customized for every user (Kanclerz et al., 2021).

We conducted three studies to examine this phenomenon. Study III utilized demographic features of the annotators to identify annotator bias. The difficulty of such an approach is that it requires personal data from the annotator. Therefore, Study IV followed an unsupervised approach. We used graphs representing the similarity between the annotators based on their annotation behavior to recognize bias. In Study V, we also employed an unsupervised approach. We developed a method that measures the deviation between an annotator and the gold standard of a dataset and produces a bias matrix for

each annotator. In addition, a set of methods that use these bias matrices to analyze annotator bias was proposed in the study.

#### 4.3.2 *Study III: Annotators' Demographic Characteristics*<sup>†</sup>

In the study "Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics" (Al Kuwatly, Wich, and Groh, 2020), we examined whether annotators from various demographic groups annotate abusive language differently. The resulting research question was the following:

RQ "How do annotators' demographic features such as gender, age, education, and first language impact their annotations of hateful content?" (Al Kuwatly, Wich, and Groh, 2020, p. 184)

We addressed this question by training abusive language classification models for each demographic group that used only the annotations from the respective group and comparing the classification performances of the models. The basis of the experiment was the personal attack dataset of Wikipedia's Detox project (Wulczyn, Thain, and Dixon, 2017) containing each annotator's raw annotations and demographic data. The demographic data comprised the features gender (female, male), age group (under 18, 18-30, 30-45, 45-60, over 60), English first language (yes, no), and education (none, some, high school, bachelors, masters, doctorate, professional). For the experiment, we grouped some of the attributes so that all features became binary: male/female, English first language yes/no, under 30/over 30, below and equal to high school/above high school. The conversion was necessary to have demographic groups with a reasonable number of annotators and annotated data. For each feature, we split the annotators into two groups based on the attributes. In the next step, we selected the text documents that both groups annotated and trained two classification models on these text documents and the group-specific labels derived from the annotations. The two classifiers were then evaluated on three test sets (two test sets with the annotations from each group and one test set with all annotations). We repeated this step 20 times to get a distribution of the F1 scores for each group. The two distributions allowed us to use the Kolmogorov-Smirnov test to check whether there is a significant difference between the two demographic groups regarding the annotation behavior ( $p < 0.05$ ). We conducted this analysis for each demographic feature.

The findings revealed that the classification models for the following demographic groups perform significantly differently, indicating that the two groups differ in their annotation behavior:

- Annotators whose first language is English or not.
- Annotators with an educational background below (and equal to) or above high school.
- Annotators under or over 30 years



For the demographic feature gender, no significant difference was discovered.

The results demonstrated that the approach is suited to identify differences between annotators based on their demographic characteristics. However, the challenge is to get this additional data because it is often not collected or published to ensure the anonymity of the annotators.

#### 4.3.3 *Study IV: Graph-Based Approach*<sup>†</sup>

The study “Investigating Annotator Bias with a Graph-Based Approach” (Wich, Al Kuwatly, and Groh, 2020) examined annotator bias in abusive language datasets but it went beyond pre-defined features to detect annotator bias (e. g., demographic data). The approach purely relied on the annotators’ annotation behavior to identify groups that annotate similarly. The corresponding research question was the following:

RQ “Is it possible to identify annotator bias purely on the annotation behavior using graphs and classification models?” (Wich, Al Kuwatly, and Groh, 2020, p. 192)

The underlying idea was to use a graph to model the annotators of an abusive language dataset and their annotation behavior. By applying community detection on the graph, we aimed to find groups of annotators with similar annotation behavior and outliers. The graph was modeled as follows: “An edge between two nodes exists if both annotators annotate at least one same data record” (Wich, Al Kuwatly, and Groh, 2020, p. 192). Each edge also has a weight. The weight is a measure for the similarity of the annotation behavior of both annotators. Within the study, four different weight functions were evaluated: Agreement Rate, Cohen’s kappa, Krippendorff’s alpha, Heuristic Function. We proposed the first and latter functions; the other two are statistical measures for inter-rater reliability from the literature (McHugh, 2012). In the next steps, groups in the graphs were detected by a community detection algorithm, and the inter-rater reliability scores within and between the groups were calculated to compare the results. For the two best distance functions, classification models were trained on the annotations of the different groups and evaluated on all groups. The resulting matrix helped to identify annotator groups with a diverging annotation behavior. For the experiment, we employed the personal attack dataset of Wikipedia’s Detox project (Wulczyn, Thain, and Dixon, 2017) containing 4,053 annotators and 1,365,217 annotations for 115,864 documents.

The study’s findings were that the graph-based approach helped identify outlier groups of annotators and that Krippendorff’s alpha and our proposed Heuristic Function worked best as distance functions. As an example, the results for the Heuristic Function as weight function are outlined in the following. Figure 4.2a displays the inter-rater reliability scores for the different groups. Figure 4.2b visualizes the macro F1 scores of the classification models trained on a group and evaluated on all groups. The scores are reported

relative to the macro F1 score of group 0 on the test set 0 to compare them easily. We observed that group 3 had the lowest inter-rater reliability of all groups (39.8%) and its classifier performed weakly on all test sets (-0.30pp) including the baseline, which was group 0 and contained all annotators. In contrast, group 1 was the group with the highest inter-rater reliability (49.2%) and its classifier performed better than the baseline (+0.07pp). That means that group 1 exhibits a coherent annotation behavior in contrast to group 3. Based on this insight, the recommendation is that either the data annotated by group 3 should be labeled by other annotators with a more coherent annotation behavior or the annotators from group 3 should receive additional training to align their annotation behavior.

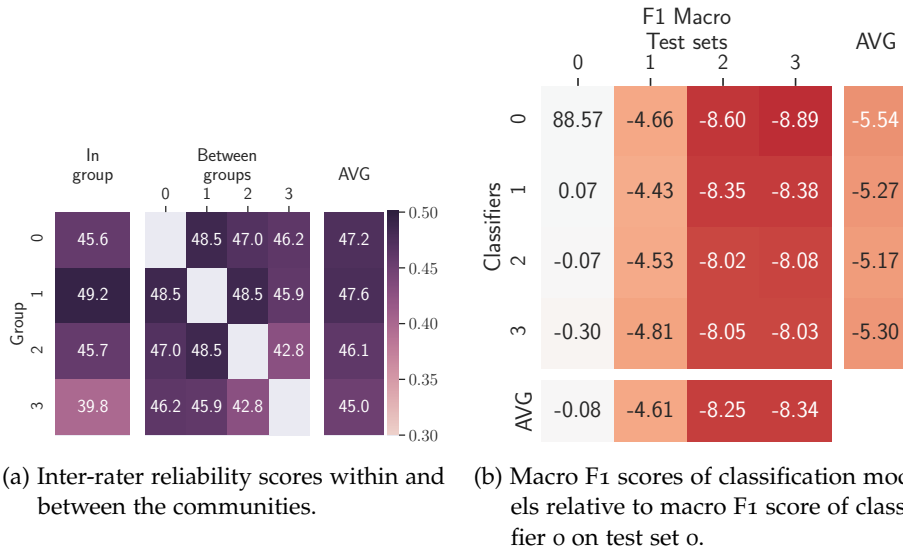


Figure 4.2: Results of Heuristic Function (figures from Wich, Al Kuwatly, and Groh, 2020, p. 196 and 197).

#### 4.3.4 Study V: Bias Matrix •

The study “Investigating Annotator Bias in Abusive Language Datasets” (Wich, Widmer, et al., 2021) dealt with the following two research questions:

RQ1 “How can we measure and visualize annotator bias in abusive language datasets?” (Wich, Widmer, et al., 2021, p. 1)

RQ2 “How can we identify and visualize different perspectives on abusive language of the annotators?” (Wich, Widmer, et al., 2021, p. 1)

To answer both research questions, we developed a method, called bias matrix, to characterize an annotator based on how strictly or liberally he or she labels the data. This method was the basis for the experiments in the study.

In the first experiment, the annotators were split into groups based on the characterization in order to identify outliers. This experiment assumed that there is only one perspective, whether a text contains abusive language or not. The second experiment, however, assumed that there are multiple perspectives. It used the bias matrix to identify annotators groups that are coherent by themselves but distinguish from other groups.

In the first experiment, four English abusive datasets were used—*Vidgen* (Vidgen, Hale, et al., 2020), *Guest* (Guest et al., 2021), *Kurrek* (Kurrek, Saleem, and Ruths, 2020), and *Wulzcyn* (Wulzcyn, Thain, and Dixon, 2017). They differed in size and in number of annotators (from 6 to 4,053). Figure 4.3 shows the bias matrices of four datasets. They are the aggregations of the annotators’ bias matrices. The most interesting cells of such a matrices are the top right (pessimistic score) and the bottom left (optimistic score). If the pessimistic score is larger than the optimistic one, the annotators annotated more strictly. If it is the other way, the annotators were more liberal. In the case of the selected datasets, we observed that the annotators of *Kurrek* were stricter, while the ones from the other three were more liberal.

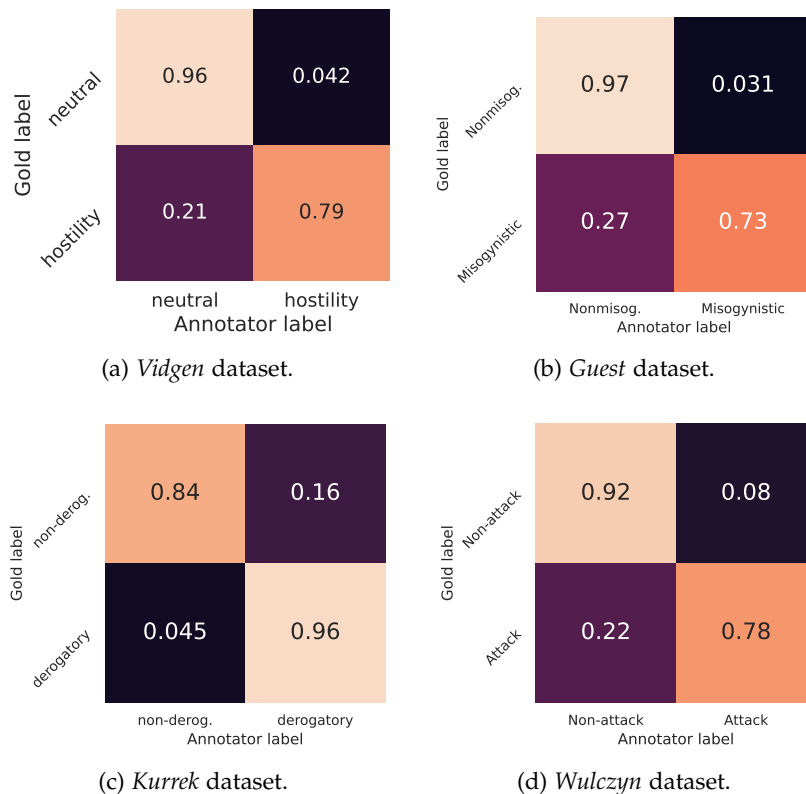


Figure 4.3: Aggregated bias matrices for the selected datasets (figures and captions from Wich, Widmer, et al., 2021, p. 5).

In the second experiment, the annotators of the *Wulzcyn* dataset were split into three groups depending on their pessimistic and optimistic scores (optimistic, neutral, and pessimistic). The goal was to identify varying perspectives on abusive language. For each group, a classification model was

trained based on the labels provided by its group members. Each model was evaluated on all three test sets (optimistic, neutral, and pessimistic). The evaluation results (macro F1) can be found in Table 4.2. We observed that the pessimistic and optimistic models performed decently on the corresponding test sets but poorly on the others. This implies that each group had a coherent perspective on abusive language but the perspectives differed between the annotator groups.

Both experiments demonstrated that our proposed bias matrix is a useful method to describe the annotation behavior. It is easy to apply because it is calculated directly from the raw annotations. However, it works only for binary labels so far.

	Pessimistic	Medium	Optimistic
Pessimistic	80.2	80.6	71.0
Medium	73.5	81.9	83.1
Optimistic	64.3	74.4	87.5

Table 4.2: Macro F1 scores from classifiers of the different annotator subsets (table and parts of the caption from Wich, Widmer, et al., 2021, p. 7).

#### 4.3.5 Discussion

The studies showed that all of the analyzed abusive language datasets face annotator bias to a certain degree. The explanation is the annotation task’s complexity combined with the subjective influence of the annotators.

In Study III and V, we observed that annotators from different demographics groups partially exhibited a varying annotation behavior. Examples are the age and whether the Annotators are native speakers or not. However, the studies do not agree on all results. Study III did not discover a difference between female and male annotators in contrast to Study V, which confirmed the results reported by Binns et al. (2017). The differences can be ascribed to the various methods to measure the bias. Study III used classification models trained on data with group-specific labels, while Study V measured the bias on annotation level without using any classification model as proxy. The advantage of the classifier-based approach is that it focuses on the actual impact of the annotator bias on the classification. But it comes with the downside that the data used for training is reduced to achieve comparable results, which is not the case in Study IV’s approach. Furthermore, the second approach does not rely on any pre-trained classification model or word embeddings that can already contain some form of bias. Consequently, it requires less computational resources and data. All in all, both approaches have their advantages and disadvantages. The results imply that a diverse group of annotators is required to annotate abusive language datasets, especially in crowd setups with non-expert annotators.

Using demographic features to identify annotator bias comes with two challenges. First, the data has to be gathered from the annotators, which is exacerbated due to privacy concerns—only one abusive language provides additional data about the annotators. Second, if demographic differences do not cause the annotator bias, the approach is unsuccessful. Therefore, Study IV and V applied unsupervised methods that do not need any additional data about the annotators but only the raw annotations. The way proposed by Study IV is quite unusual because it uses a graph representing the annotators' annotation behavior to identify groups. The downside of the method is that it needs datasets labeled by a large number of annotators. Otherwise, the graph is not big enough to discover groups with a similar annotation behavior. Study V addresses this problem by proposing the bias matrix as a measurement for annotator bias. It can be easily calculated independently from the number of annotators, as shown in the study. In addition, there is a range of different methods to analyze and compare the annotators' bias matrices within and between abusive language datasets. Therefore, it is an ideal extension of the bias and comparison framework presented in Section 4.2 and would close the need of a more sophisticated method to examine annotator bias.

The primary use case of the unsupervised methods is to detect outliers within the annotators and measure annotator bias. The assumption here is that it exists only one perspective on abusive language. But in both studies, the idea of multiple perspectives was examined and discussed. The results showed that more than one perspective on abusive language exists in the analyzed dataset and that the proposed methods help identify them. Such an assumption is in line with studies from other researchers that claim a paradigm shift. They do no longer adhere to the paradigm of a single gold standard for subjective tasks, such as abusive language detection, and suggest a multi-perspective approach (Akhtar, Basile, and Patti, 2020; Basile, 2020; Davani, Díaz, and Prabhakaran, 2021; Kanclerz et al., 2021). Kocoń et al. (2021), for example, proposed approaches for abusive classification models personalized on an individual or a group level.

The advantages of such an approach are versatile. First, the classification performance can be improved in multi-perspective models because the inconsistencies in annotations caused by annotator disagreement are reduced (Akhtar, Basile, and Patti, 2020; Davani, Díaz, and Prabhakaran, 2021). Second, classification models that represent the social norms of an individual or a group can be developed. Such models would enable perspective-aware handling of abusive language (e. g., personalized filters for social media) (Kanclerz et al., 2021; Kocoń et al., 2021). Therefore, the research community should investigate the multi-perspective approach further.

#### 4.4 POLITICAL BIAS IN ABUSIVE LANGUAGE DATASETS

##### 4.4.1 *Motivation*

This section deals with another form of unintended bias in abusive language datasets—political bias. It overlaps with annotator bias, which was discussed in the previous section, and topic bias because a dataset can be politically biased in two ways. Firstly, the annotators can bring in a bias due to their political orientation. Secondly, the data can focus on political topics because of the data gathering and sampling approach (e.g., when explicitly posts from right-wing accounts are collected).

In ML research, political bias has already gained the attention of some fellow researchers (Aksenov et al., 2021; Chun et al., 2019; Gordon, Babaeian-jelodar, and Matthews, 2020; Hajare et al., 2021). However, studies in the field of abusive language detection are highly limited. Jiang, Robertson, and Wilson (2020), for example, examined the influence of a video’s political tendency on the content moderation decision. The following study aimed to address this gap and investigated the relevance of political bias on abusive language detection. Since there is no abusive language dataset with any information about political tendencies, the study relied on a simulation of political bias in data. The results should set the stage for further research in this area.

##### 4.4.2 *Study VI*•

In the study “Impact of Politically Biased Data on Hate Speech Classification” (Wich, Bauer, and Groh, 2020), the influence of political bias in datasets on hate speech detection was investigated. To achieve this goal, the following research questions were addressed:

- RQ1 "What is the effect of politically biased datasets on the performance of hate speech classifiers?" (Wich, Bauer, and Groh, 2020, p. 54)
- RQ2 "Can explainable hate speech classification models be used to visualize a potential unintended bias within a model?" (Wich, Bauer, and Groh, 2020, p. 55)

To answer these questions, three datasets with simulated political biases were created and used to train classifiers, whose performances (F1 scores) were compared. The foundation for the three datasets was a combination of the German Twitter corpora from the Shared Task on the Identification of Offensive Language 2018 (Wiegand, Siegel, and Ruppenhofer, 2018) and 2019 (Struß et al., 2019)—both can be merged due to the same labeling schema. The offensive labeled tweets were extracted and combined with tweets from political subnetworks (politically left-wing, politically right-wing, politically neutral) that topically matched with the non-offensive tweets from GermEval 2018 and 2019. The tweets from the political subnetworks

were implicitly labeled as non-offensive. The underlying assumption was that the existence of the tweets demonstrated a general acceptance by the subnetwork and accordance with the norms’ of the subnetwork. Otherwise the tweets would have been deleted after some time, if members of the subnetwork had felt offended by these tweets. Based on the created datasets, three classification models were trained and their F1 scores were compared, addressing RQ1. Concerning RQ2, the explainability framework SHAP is applied to one classifier for each political orientation and the explanations of randomly selected tweets were analyzed. The aim was to investigate whether the technique can help make such a bias visible.

The results showed that classifiers trained with politically biased datasets exhibited a significantly<sup>2</sup> different performance. The best performing model was the politically neutral one (84.8%), followed by the left-wing one (83.3%) and the right-wing one (78.7%). That means that hate speech classifiers trained on a dataset with a political bias performed worse than the one trained on a neutral dataset, which answers RQ1. The in-depth analysis of single predictions with SHAP confirmed the results and also provided an explanation. Figure 4.4 shows the output of SHAP for each politically biased classifier for the offensive tweet “@<user>@<user> Natürlich sagen alle Gutmenschen ‘Ja’, weil sie wissen, dass es dazu nicht kommen wird..” (@<user>@<user> Of course, all do-gooders say ‘yes’, because they know that it won’t happen.)” (Wich, Bauer, and Groh, 2020, p.60). The red bars indicate to favor the classification as offensive, the blue ones as non-offensive. The longer the bar, the more relevant the word is for the tendency. It appears to the reader that the word *Gutmensch* is not uniformly relevant for the models. While the word strongly favors a classification as offensive in the case of the left-wing and neutral classifiers, it plays a less relevant role for the right-wing model. Because *Gutmensch* is an abusive term often used by right-wing scene (Hanisch and Jäger, 2011), we should not be supervised by the low relevance for the right-wing classifier. That shows that explainable models can help to visualize unintended bias.

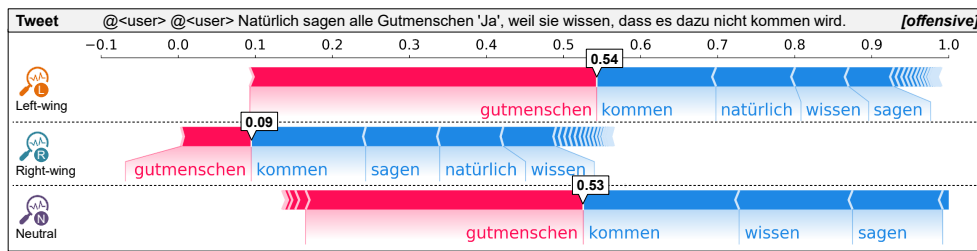


Figure 4.4: Offensive tweet misclassified by right-wing classification model (figure from Wich, Bauer, and Groh, 2020, p. 62).

<sup>2</sup>  $p < 0.01$ ; each classifier was trained 30 times to get meaningful results

#### 4.4.3 Discussion

The study showed that a politically biased dataset could impair the classification performance of an abusive language classifier. The reported  $F_1$  scores suggest that a right-wing bias negatively impacts the performance more than a left-wing bias. But the experiment does not allow such a statement because there are no details of the dataset's original composition. If the original dataset contained more right-wing tweets, for instance, the performance of the right-wing classifiers would be worse than the left-wing. The reason is that the added right-wing bias would confuse the classifier more than a left-wing bias in this case. Therefore, it is not possible to compare the impact of right-wing and left-wing bias. The results only show that political bias impairs classification performance in abusive language detection.

The XAI part of the study demonstrates that explainable models work and provide insightful results, allowing the user to understand a model's prediction. It is interesting to see how the relevance scores of the words for a prediction vary between the different classifiers. The visualizations uncover which and how relevant words are for the classifiers. Explainable models that provide insights on predictions are crucial notably for abusive language detection because classifying something as abusive or not is a trade-off between "censorship, free speech, and privacy" (Vidgen, Hale, et al., 2020, p. 88) and consequently a value-based decision. Therefore, it is necessary to make these values transparent. Explainable capabilities will increase the trust of the users in such systems (Kiritchenko, Nejadgholi, and Fraser, 2021)—a mandatory requirement for their success. Further use cases of explainable models are identifying model bias and debugging models during development phase (Kiritchenko, Nejadgholi, and Fraser, 2021).

An explicit limitation of the study is that the political bias was only simulated as no appropriate dataset was available. This can be addressed in future work by creating an abusive language dataset annotated by three groups of different political orientations (left-wing, right-wing, political center). The data to be annotated can be collected according to the data gathering process proposed by the study. Finding appropriate and enough annotators is more challenging due to the various political orientations. A possible solution would be to use a Twitter bot that automatically contacts Twitter users and kindly asks them to annotate a small number of tweets (Alperin et al., 2017). For selecting the appropriate users, the methods proposed by Shahrezaye, Papakyriakopoulos, et al. (2019) or Shahrezaye, Meckel, and Hegelich (2020) can be applied to predict the political orientation of a Twitter account. Such an abusive language dataset with politically biased annotations would be ideal for further investigating the phenomenon and producing more reliable results.



## 4.5 ETHICAL CONSIDERATION OF ANNOTATOR BIAS

### 4.5.1 *Motivation*

The two previous sections dealt with the impact of annotators' subjective perception on data annotation from a technical perspective: How can we identify such a bias? How can we measure it? What is its impact on classification performance?

The ethical perspective, however, has not been addressed yet, which is done by this section. It is crucial because ethical challenges accompany the fight against hate speech and other forms of abusive language. Examples are: Is it hate speech or freedom of expression? Should hate speech be deleted or countered? Should users that spread hate speech be banned from platforms?

The study presented and discussed in this section examined the ethical component of data annotation influenced by the annotators' subjective perception.

### 4.5.2 *Study VII<sup>†</sup>*

In the transdisciplinary study "Mediale Hasssprache und technologische Entscheidbarkeit: Zur ethischen Bedeutung subjektiv-perzeptiver Datenannotationen in der Hate Speech Detection" (Lerch et al., 2022), we examined the detection of hate speech and the impact of subjectively influenced annotations from an ethical perspective. The aim was to connect the technical aspects of hate speech detection and ethical criticism.

After defining hate speech and outlining the technological aspect of its detection, the concept of hate speech was examined from an ethical perspective based on the three different theories ranging from classical liberalism over 20th-century philosophers and recently published studies—i.e., Mill (2011), Rawls (1999), and Waldron (2012). Key findings were that hate speech is characterized by objectively negative consequences or by the we-they dichotomy<sup>3</sup>. Feelings, such as feeling offended, are relevant. Hate speech can be offensive, but not every offense is hate speech. However, this blurred boundary aggravates an objective annotation and fosters subjective influence.

In the next part, we added the technological perspective to the ethical discussion and examined the implications for implementing the hate speech detection model. To address the challenge of subjective annotations, we have two options: First, we can choose a top-down model, making logical, rule-based decisions, also known as symbolic artificial intelligence. Second, we can choose a bottom-up model, meaning the detector is trained on coded knowledge (annotated data) to learn the decision process. While the first option is not feasible due to the complexity of hate speech, the second one is vulnerable to the annotators' subjective perception. One may argue that subjective annotations are unproblematic if the annotators are a diverse and

---

<sup>3</sup> The perspective "we against they" aims to disrespectfully exclude the other group.

representative society sample. However, this would discriminate minorities and their perception. A possible solution is a hybrid approach combining top-down and bottom-up—an annotation process in which annotators make rule-based decisions. But it requires further research to meet the ethical requirements.

In the third part of the study, we proposed a compromise solution to address the ethical challenges and make the process feasible. As of now, the most suitable way is a rule-based annotation process. It has to satisfy the following requirements: (1) the annotators have to be experts and have diverse backgrounds. (2) The annotators need precise guidelines with instructions based on clear definitions to reduce ambiguity. (3) The annotators need room for discussion. (4) The process and materials have to be documented and made public to ensure transparency. By doing so, the influence of annotators' subjective perception can be reduced or at least made transparent.

#### 4.5.3 *Discussion*

Comparing the requirements defined by the ethical examination with best practices from the abusive language detection community (Vidgen, Hale, et al., 2020), we observed full agreement. However, this agreement does not make the ethical study worthless. Quite the contrary, it is a valuable extension because it provides an ethical perspective on the problem, while the best practices represent the practical and engineering perspective.

If we examine published abusive language datasets and especially the commonly used ones with respect to the requirements, we recognize some shortcomings. A large portion of the researchers that have released abusive language datasets have not published the annotation guidelines, making it hard to reproduce or review the annotation process (Vidgen, Harris, et al., 2019). Another shortcoming related to missing transparency is that most of the datasets contain only the final labels and not the raw annotations of the annotators. However, raw annotations would help to evaluate the quality of annotation regarding subjectivity (Vidgen and Derczynski, 2021). To annotate an abusive language dataset, researchers often used crowdsourcing platforms to outsource the annotation process, such as Figure Eight (former CrowdFlower) or Amazon Mechanical Turk (Poletto et al., 2021; Vidgen and Derczynski, 2021). This approach, however, can impair the quality of the annotations because the annotators are often no experts in annotating abusive language and the possibilities to properly instruct them and offer room for alignment discussions are highly limited. This can lead to a stronger annotator bias. Considering recently published datasets, we can observe an improvement. Vidgen, Hale, et al. (2020), for example, published the dataset together with all raw annotations provided by non-crowdsourcing workers and the codebook. The abusive language dataset of Kurrek, Saleem, and Ruths (2020) provides a similar degree of transparency. Furthermore, the authors conducted a comprehensive and interactive annotator training to instruct and align the annotators.

In the previous sections, two studies were presented and discussed that brought up the idea of multiple perspectives on abusive language, meaning two or more groups can perceive abusive language differently (Wich, Al Kuwatly, and Groh, 2020; Wich, Widmer, et al., 2021). The ethical study did not explicitly address this question of multiple perspectives. But we argue that the various perspectives are not an error due to low-quality annotations but rather a result of different perceptions. That is not in contrast to the results of the study. We could have more than one social group and each of them has its own norms. These norms determine the boundary conditions whether a text is abusive or is covered by freedom of expression. This can lead to multiple perspectives on abusive language. An interesting suggestion provided by Basile (2020) is to discard the assumption that there is only one perspective on abusive language and to build "perspective-aware models" (Basile, 2020, p. 39). Kanclerz et al. (2021), Davani, Díaz, and Prabhakaran (2021), and Kocoń et al. (2021) developed such perspective-aware models for abusive language detection. However, such approaches require a stronger interdisciplinary discourse between ethics and machine learning to define the circumstances of such systems in order to meet the ethical requirements and to be socially accepted.

## 4.6 INTEGRATION OF USER AND NETWORK DATA INTO ABUSIVE LANGUAGE DETECTION

### 4.6.1 *Motivation*

Most research in the area of abusive language detection focuses only on textual data as features for the classification (Mishra, Yannakoudakis, and Shutova, 2020; A. Schmidt and Wiegand, 2017; Vidgen, Harris, et al., 2019). However, we have learned in Section 4.1 that hate and other forms of toxicity can be expressed in various ways, impairing a classification relying only on textual data. Therefore, researchers started to integrate additional data to improve classification performance. Two auspicious data sources are the social network and the post history of the user whose post is to be classified. The reason for adding the first data source is that relatively small networks of accounts produce a high percentage of offensive and hateful content according to Kreißel et al. (2018), meaning network data could be appropriate features. Similar findings were reported by Evkoski et al. (2021). A range of studies used network data as additional features—either topological properties of the network (e. g., degree, centrality) (Chatzakou et al., 2017; Fehn Unsvåg and Gambäck, 2018; A. M. Founta et al., 2019; Papegnies et al., 2017) or network embeddings (Mishra, Del Tredici, et al., 2019). Instead of classifying text documents (e. g., tweets, comments), some studies aimed to classify the users as haters or non-haters, using graph embeddings based on the social graph and textual data (Das et al., 2021; Li et al., 2021; M. Ribeiro et al., 2018). The reason for adding the user’s post history is related to the previous one. If a user frequently posts abusive content, the previous posts are a good indicator for classifying a new post (Chaudhry and Lease, 2020; Pitsilis, Ramampiaro, and Langseth, 2018; Qian et al., 2018; Raisi and B. Huang, 2017; Rangel et al., 2021).

We have already learned that there is a demand for explainability in abusive language detection to identify bias and build trust (Mishra, Yannakoudakis, and Shutova, 2021; Vidgen, Harris, et al., 2019). Integrating additional data sources into the classification strengthens this need because it increases the potential sources of bias. However, studies that built and examined explainable multimodal abusive language classification models are very limited (Vijayaraghavan, Larochelle, and Roy, 2019).

The following three studies addressed the gap in the field of explainable abusive language classifiers leveraging user and network data. In Study VIII, an abusive language dataset was created that contains social network data of the users in addition to the annotated texts. It was necessary because nearly all abusive language datasets do not provide any social network data of the users. Study IX dealt with a classification model that uses text, user, and network data as features and methods that explain predictions of the model. While the explainable methods target expert users who build and train such models, the explainable multimodal classifiers in Study X is more end-user

friendly. Furthermore, the model from Study X used more sophisticated sub-models to provide state-of-the-art performance.

#### 4.6.2 Study VIII: Abusive Language Dataset Containing Social Network Data •

In the study “Are Your Friends Also Haters? Identification of Hater Networks on Social Media: Data Paper” (Wich, Breiting, et al., 2021), a methodology was developed to identify and collect subnetworks on Twitter that contain a large portion of haters. A further outcome was a German offensive language dataset comprising social network data from the authors of the collected tweets.

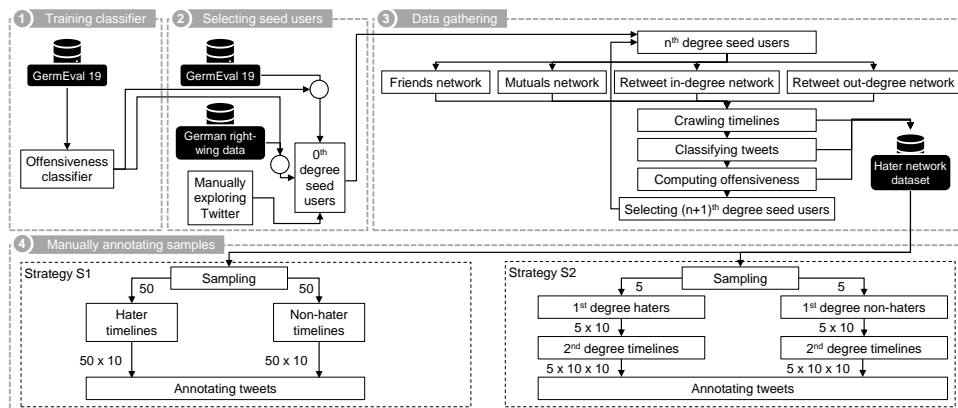


Figure 4.5: Dataset creation methodology (figure from Wich, Breiting, et al., 2021, p. 3).

Figure 4.5 visualizes the dataset creation process. In step 1, an offensive language classification model was trained on the combination of two datasets. We used the German Twitter corpora from the Shared Task on the Identification of Offensive Language 2018 (Wiegand, Siegel, and Ruppenhofer, 2018) and 2019 (Struß et al., 2019), as in Study VI. The basis of the classification model was a pre-trained BERT model (MDZ Digital Library, 2021). The model was used in step 2 to detect haters in a German right-wing dataset that served as seeds for the gathering process. Additional seeds were extracted from the GermEval 2019 dataset and from manual exploration of Twitter, resulting in a total of nine seed accounts. In the next step, the networks of the seed accounts were iteratively collected. Four different types of social relations were relevant for collecting the network. Assuming we want to collect the network of account A, the four network relations are: accounts that A follows (friends network), accounts that follow A and are followed by A (mutuals), accounts that retweet A (retweet in-degree), and accounts that are retweeted by A (retweet out-degree). After identifying these accounts, we collected the tweets posted by them and classified them with the classification model. Based on the classified messages, an offensiveness score was calculated for each account. This score together with the accounts’ interconnectedness in the collected network served as selection criteria for the seed accounts of

the next iteration. In the fourth step, two different sampling strategies were applied to sample data that was manually annotated. Two different strategies were employed to get a diverse sample with a large portion of offensive content. Due to limited resources and a large amount of collected data, only a sample was manually annotated. Each tweet of this sample was annotated by at least two annotators. In case of disagreement, a third annotator provided an additional vote. The annotated data was used to evaluate the approach. The rest of the collected tweets was only pseudo-labeled by the classification model.

Besides the methodology, the study produced an abusive language dataset containing 4,647,200 labeled German tweets, 49,353 users, and 122,053 social relations. 1,356 of 4,647,200 tweets were manually annotated. The evaluation of the used classification model based on the annotated tweets shows a macro F1 score of 72.5%, which is a decent performance. Another interesting finding was that the network type with the highest portion of offensive content was the retweet out-degree network.

#### 4.6.3 *Study IX: Integration of User Context in Hate Speech Detection*<sup>†</sup>

In the study “Understanding and Interpreting the Impact of User Context in Hate Speech Detection” (Mosca, Wich, and Groh, 2021), an explainable abusive language classification model was built that leveraged text, user, and social network data to classify text documents. The research objectives (RO) of the study were the following:

RO1 Improving the model’s classification performance

RO2 Making the model’s predictions more transparent to better understand its behavior

To combine the three different types of input data, our classification model consisted of three submodels. We kept the model’s architecture simple because we wanted to demonstrate the improvement of adding user and network data and not to achieve state-of-the-art performance. Therefore, the first submodel processed the text document, which is to be classified, in form of a simple bag-of-words vector. The second submodel dealing with the user’s history also employed a bag-of-words vector representing all text documents posted by the user. The third submodel received the user’s social network as a binary vector extracted from the adjacency matrix.

We trained and evaluated this model on two different datasets—Waseem and Hovy (2016) and Davidson, Warmesley, et al. (2017). For Waseem and Hovy (2016), the multi-modal model outperformed the pure text model by 4.3pp (F1 score). For Davidson, Warmesley, et al. (2017), we measured an improvement of 1.0pp.

To make the predictions of the multi-modal model more understandable for humans, we applied two different techniques. The first one employed Shapley values that estimate the relevance of a feature for a prediction (Lundberg

and Lee, 2017). Figure 4.6 visualizes the feature contribution of a selected sexist tweet that was correctly classified by the model. The first twelve rows represent the words used as input for the text model. The row *VOCABULARY* shows the contribution of the user’s previous tweets and *NETWORK* the contribution of the user’s social network. We can observe that the tweet history of the user has a huge impact on the classification of the tweet.

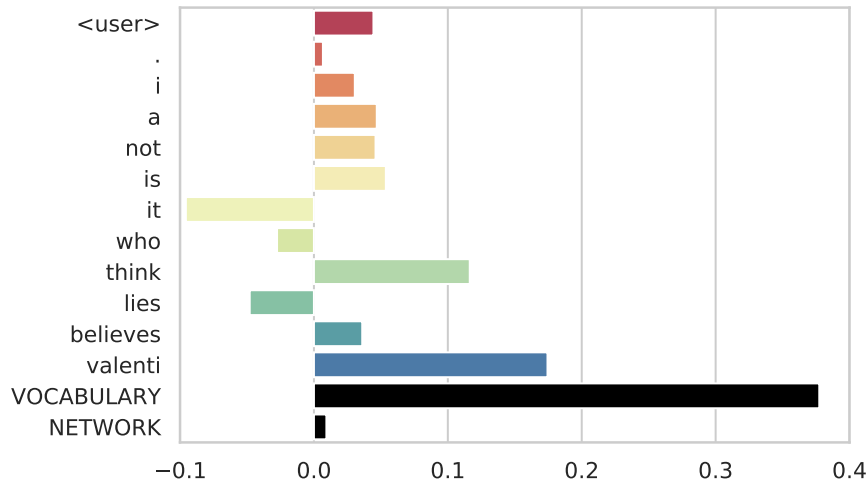


Figure 4.6: Feature contribution based on Shapely values for the tweet “<user> I think Arquette is a dummy who believes it. Not a Valenti who knowingly lies.” (figure and parts of the captions from Mosca, Wich, and Groh, 2021, p. 95).

The second technique, called *learned feature space exploration*, used the last hidden layer (latent space) of the multi-modal model and visualized it with *t-Distributed Stochastic Neighbor Embedding (t-SNE)* (Cieslak et al., 2020; Maaten and Hinton, 2008). The result is a two-dimensional visualization of the data, as depicted in Figure 4.7. The plot helps to better understand what the model learned and to examine the model. This technique addresses expert users who develop such models, while the first technique can be also used by end-users. Nevertheless, both techniques contribute to increased transparency of predictions.

#### 4.6.4 Study X: Explainable Abusive Language Classification Leveraging User and Network Data •

In the study “Explainable Abusive Language Classification Leveraging User and Network Data” (Wich, Mosca, et al., 2021), an abusive language classification model was developed that uses the textual data and network data to classify a tweet. Additionally, an XAI technique was used to make the prediction more understandable for humans. The study addressed two research questions:

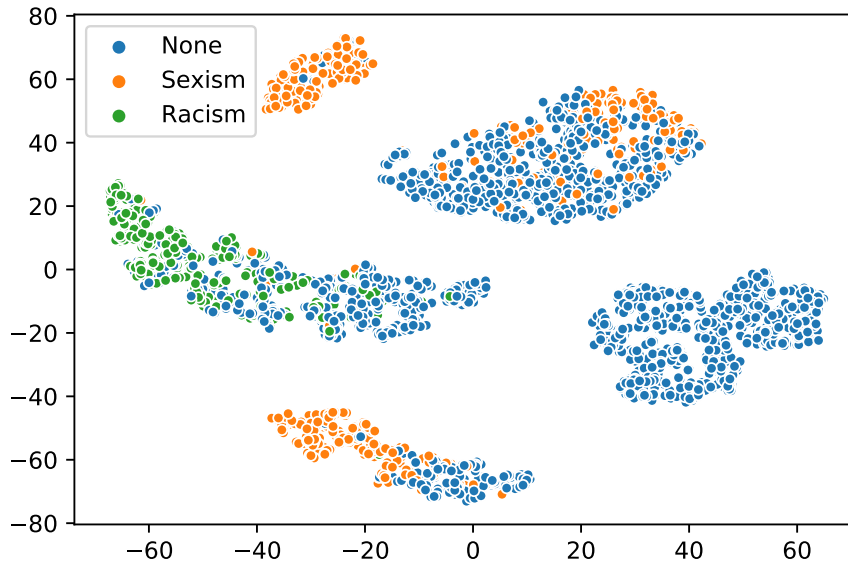


Figure 4.7: Visualized latent space of multi-modal model trained on Waseem and Hovy (2016); each dot represents a tweet, colored by label of the tweet (figure from Mosca, Wich, and Groh, 2021, p. 96).

RQ1 "Can abusive language classification be improved by leveraging users' previous posts and their social network data?" (Wich, Mosca, et al., 2021, p.1)

RQ2 "Can explainable AI be used to make predictions of a multimodal hate speech classification model more understandable?" (Wich, Mosca, et al., 2021, p.1)

To answer the RQ1, a multimodal classification model was built consisting of three submodels. The first one was a transformer-based model that processed the text data; it employed a pre-trained DistilBERT model (Sanh et al., 2020). The second one was a bag-of-words model handling the post history of a user. The third one was a GraphSAGE graph embeddings that modeled the social network of a user (Hamilton, Ying, and Leskovec, 2017). To answer RQ2, SHAP (Lundberg and Lee, 2017) was used to explain the prediction of the multimodal model. The explainable multimodal model was trained and evaluated on three different datasets—Davidson, Warmesley, et al. (2017), Waseem and Hovy (2016), and Wich, Breitinger, et al. (2021). The latter was the dataset described in Section 4.6.

One of the main findings was that user and network data can enhance the classification of abusive language. The improvements were not tremendous; the increase of the macro F1 score ranged between 0.1pp and 2.4pp depending on the dataset. There are two reasons for this: First, the network data of the datasets from Davidson and Waseem were very sparse because they were not collected from a connected subnetwork. Second, the text submodel from the Wich dataset performed so well that it outperformed the other submodels.



Consequently, the network and user submodels played a less relevant role in the classification.

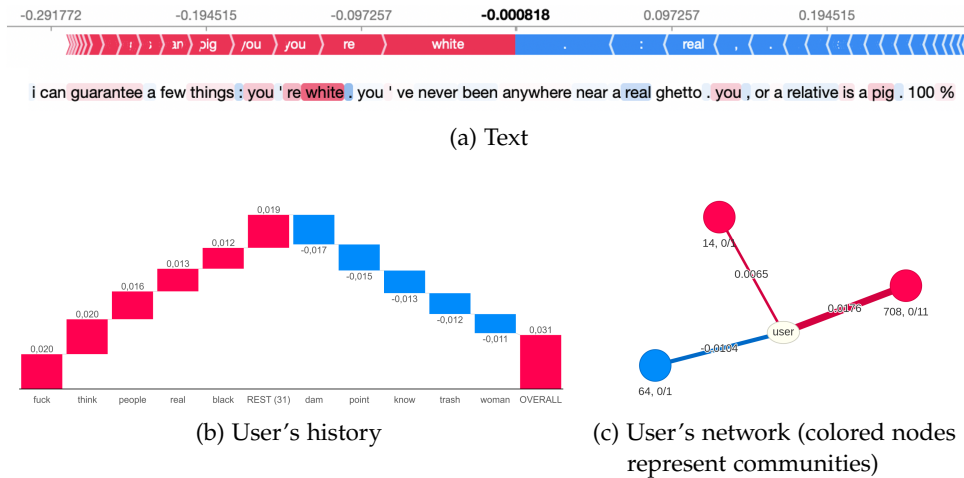


Figure 4.8: Explanations for predictions of text, history, and network submodel in the form of Shapely values; red, positive values favor a classification as hateful; blue, negative values favor a classification as non-hateful (figures and captions from Wich, Mosca, et al., 2021, p. 12).

Regarding the explainable part of the study, the explanations for the submodels provided helpful insights to make a prediction understandable for humans. They made the contributions of the submodels transparent. An example was visualized by Figure 4.8. The text submodel (cf. Figure 4.8a) misclassified the tweet as neutral because the word favoring a neutral classification (blue) overruled the other group (red), resulting in a score of  $-0.000818$ . However, the other two submodels corrected the prediction (cf. Figure 4.8b and 4.8c). The example does not only present the value of the explanations but also how multimodality can improve the predictions.

#### 4.6.5 Discussion

The three studies show that social network data and the user's previous posts are suited to improve the abusive language classification of a new post and to make predictions more understandable.

Study VIII propose a methodology to identify and collect data from hater subnetworks on Twitter. The dataset, the second contribution of the study, is one of a few abusive language datasets containing text and network data (Das et al., 2021; M. Ribeiro et al., 2018; Ziems et al., 2021) and is the only one in German. However, it comes with one limitation—most tweets in the dataset are only pseudo-labeled by a classification model. The results from Study X, which used this dataset, showed that the pseudo-labels are less suitable for training a classification model using user and network data as additional features. The text submodel of the classification model performed so well that the network and user part become nearly irrelevant because the

text submodel emulated the model used for pseudo-labeling. Nevertheless, the dataset is a valuable contribution as it can be used, for example, to train classification models for users instead of texts.

The findings regarding the classification performance from Study IX and X are that social network data and post history of a user can improve abusive language detection. They are in line with the results from many other studies integrating context data, such as social network data, previous posts, or account data (Chatzakou et al., 2017; Chaudhry and Lease, 2020; Fehn Unsvåg and Gambäck, 2018; A. M. Founta et al., 2019; Mishra, Del Tredici, et al., 2018, 2019; Papegnies et al., 2017; Qian et al., 2018; Raisi and B. Huang, 2017; Rangel et al., 2021). However, the increase of the F1 score is limited. An explanation is the lack of datasets suitable for developing such models. Two of the three used datasets were collected based on keywords (Davidson, Warmlesley, et al., 2017; Waseem and Hovy, 2016) and enriched with network data afterwards. The consequence of the keyword-based approach is that the datasets do not stem from dense and connected subnetworks, making the network data less expressive. The issue with the third dataset (Study VIII) is already outlined in the previous paragraph. Notwithstanding, that does not limit the findings of Study IX and X. On the contrary, the fact that the classification performance could be increased by integrating the additional data sources despite the limitations emphasizes the benefit of social network data and post history.

The improvement of the classification performance is as relevant as the explainability of the models. This capability allows us to understand better what the model has learned and how it comes to a particular prediction. But why is it crucial? Multimodal models using different data sources as input are more vulnerable to unintended bias because one part of the model could unreasonably overrule another part. Therefore, it is crucial to provide information about how relevant the individual data sources and submodels are. The demonstrations in Study IX and X showed how valuable these explanations could be. When we compare the proposed explainability methods from both studies, we can find some similarities and differences. The Shapley-based method from Study IX is similar to the one from Study X because it relies on the same technology. The advance of Study X is that the explanations for the social network and user's history component are more fine-grained and provide details in contrast to the aggregation of Study IX. The target group of this approach is the end-user with low or no technical experience, who could be, for example, a moderator or a user of a social media platform. In contrast, the t-SNE-based method from Study IX targets engineers and data scientists who build and train such models. It "can be considered a global explainability technique" (Mosca, Wich, and Groh, 2021, p. 95), meaning it aims to explain the complete model. However, this aim comes along with the increased complexity of the explanations. Therefore, it is less suitable for end-users. Even if both studies address the gap of missing explainability and integrating context data, as demanded by the research community (Mishra, Yannakoudakis, and Shutova, 2021; Vidgen, Harris, et al., 2019), there is still much work to do. As a next step, it would be helpful to conduct user experience evaluations to

assess and improve the different approaches. Furthermore, the explainability approaches, especially in the area of global explainability, require further research to produce more expressive explanations.

Based on the promising findings from the studies, the research community should foster studies in this direction. However, integrating context data into the abusive language classification may not happen without combining it with XAI. Otherwise, there is a risk that the models already vulnerable to bias become more prone to unfair behavior.

## 4.7 GERMAN ABUSIVE LANGUAGE DATASET FOCUSING ON COVID-19

### 4.7.1 *Motivation*

The COVID-19 pandemic had a massive impact on the world. It nearly influenced every facet of our daily lives, which caused a lot of polarization in our society. As a consequence, the pandemic also shaped online hate (Guhl and Gerster, 2020; Velásquez et al., 2020). The reason is that an outbreak causes fear, "and fear is a key ingredient for racism and xenophobia to thrive" (Devakumar et al., 2020, p. 1). China and its population, for example, were slandered and stigmatized because they were made responsible for the rise of the virus (Fan, Yu, and Z. Yin, 2020; Vidgen, Hale, et al., 2020).

The challenge associated with new trends in abusive language is that classifiers trained on datasets from the pre-COVID-19 era might not distinguish abusive texts from normal ones effectively. The datasets do not contain the topic-specific abusive terms; consequently, the classification model cannot recognize them. Therefore, there is a demand for abusive language datasets covering new trends and topics, such as COVID-19.

We can already find a small number of datasets that specifically address this need. Vidgen, Hale, et al. (2020) collected and annotated an abusive language corpus that contains 20,000 English tweets about East Asian prejudice in the context of COVID-19. Ziems et al. (2021) published an English Twitter dataset with hate speech targeting Asians and counter-speech with 2,400 annotated tweets. Furthermore, the dataset contains one million tweets labeled by the classification model trained on the annotated data (Ziems et al., 2021). Cotik et al. (2020) are currently building a hate speech corpus with Spanish tweets collected during the pandemic.

Nevertheless, there is still a demand for abusive language datasets focusing on COVID-19 due to the diversity of the topic. In Germany, for example, online far-right hate actors became very popular during the pandemic due to their anti-government attitude and due to conspiracy theories regarding the COVID-19 measures (Fielitz and K. Schwarz, 2020; Guhl and Gerster, 2020). Besides, the number of German abusive language datasets is quite limited. Therefore, the following study addressed these gaps.

### 4.7.2 *Study XI* •

In the study "German Abusive Language Dataset with Focus on COVID-19" (Wich, Räther, and Groh, 2021), we collected and annotated a German abusive language dataset with a topical focus on COVID-19. Besides releasing such a dataset, the study developed a methodology to collect abusive tweets from Twitter with a topical focus.

The data collecting and annotating process is visualized by Figure 4.9. It started with the selection of Twitter accounts that served as seeds for the data collection. For the COVID-19 dataset, we used three accounts that spread misinformation related to the pandemic according to Richter et al.

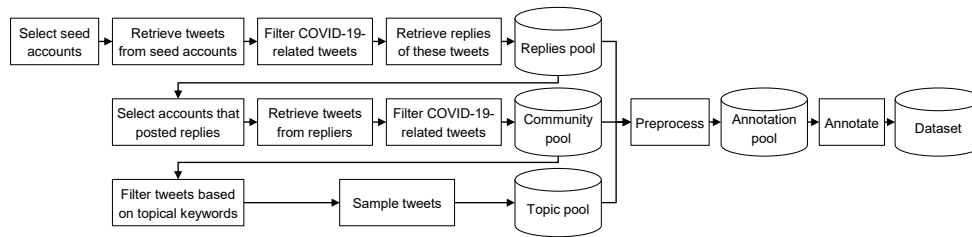


Figure 4.9: Dataset collection and annotation process (figure from Wich, Räther, and Groh, 2021, p. 2).

(2020). Next, we retrieved all tweets posted by the seed accounts and filtered COVID-19 related tweets based on keywords. In the fourth step, all replies to these COVID-19-related tweets were extracted and stored in the replies pool. Then, we collected a maximum of 500 tweets from every account that appeared in the replies pool. After filtering these tweets with the same set of COVID-19-related tweets, we stored the selection in the community pool. In the next step, we filtered the tweets in the community pool based on topical keywords. These keywords were derived from topics that appear in the context of hate speech related to COVID-19 according to sCAN (2020). A sample of the filtered tweets was stored in the topic pool. The purpose of the topic pool was to raise the prevalence of abusive tweets because the annotation pool was equally fed by samples from the three pools. The tweets stored in the annotation pool have passed through a preprocessing step to remove particular elements (e. g., URLs, user names) and duplicates. In the last step, a sample from the annotation pool was labeled by three annotators. The labeling schema had two classes—*abusive* and *neutral*. To measure the inter-rater reliability, a subset of the tweets was annotated by at least two of them. To provide a baseline for the COVID-19 dataset and compare it with other German datasets, transformer-based classification models were trained for all datasets and evaluated on all datasets.

The outcome of the dataset creation process was a dataset that contains 4,960 German tweets that mainly deal with COVID-19-related topics. 22% of the data was labeled as *abusive*, 78% as *neutral*. The classification model trained on the COVID-19 dataset achieved a macro F<sub>1</sub> score of 85.9%—a decent performance. The cross-dataset evaluation revealed that classifiers trained on already existing datasets performed worse on the COVID-19 dataset, while the COVID-19 classifier showed a comparable performance on the other datasets. This observation indicates that classifiers being not familiar with new topical domains of abusive language show a decreased classification performance, underlying the demand for datasets covering current trends in abusive language.

### 4.7.3 Discussion

The dataset collection methodology and the dataset itself are valuable contributions to the research community. The methodology describes a novel

process on how to collect abusive tweets from Twitter with a topical focus. The systematic approach aims to reduce unintended bias in the data by different sampling steps (e. g., avoid oversampling of topics or tweets from particular users). Even if it was developed and applied for the COVID-19 topic, it is not limited to this use case. The collection process can be used for other use cases by replacing the seed accounts and the keywords.

The dataset extends the portfolio of German abusive language datasets and fills the gap of a COVID-19-related dataset. Its size is comparable to other German datasets, which rang between 469 and 8,541 documents (Bretschneider and Peters, 2017; Mandl, Modha, Kumar M, et al., 2020; Mandl, Modha, Majumder, et al., 2019; Ross et al., 2016; Struß et al., 2019; Wiegand, Siegel, and Ruppenhofer, 2018). In contrast to other datasets, the collection process is well documented and transparent, and raw annotations are available—requirements resulting from Studies II, IV, V, VI, and VII. A limitation of the dataset is the binary labeling schema. Recently published German abusive language datasets have a more fine-grained and hierarchical schema to distinguish between various forms of abusive language (Mandl, Modha, Kumar M, et al., 2020; Mandl, Modha, Majumder, et al., 2019; Struß et al., 2019; Wiegand, Siegel, and Ruppenhofer, 2018). Due to resource limitations, it was not possible to apply a more granular labeling schema for the annotations process of the COVID-19 dataset. As the boundaries between abusive language and misinformation blur in the context of COVID-19 (sCAN, 2020), a labeling schema for such datasets should consider that in the future.

The challenge of creating new datasets is the resource-intensive annotations process. It costs money and time. An alternative is domain adaption, a subcategory of transfer learning. It aims to fine-tune a model trained on a domain (source) to perform well on another (target) (Sun, Shi, and Wu, 2015). Bashar et al. (2021) proposed an unsupervised approach that does not require manual annotation to improve classifiers regarding the detection of COVID-19-related hate posts. The models adapted for the target domain (COVID-19) perform better than the baseline. However, their performance is still worse than the performance of the models trained on data from the target domain. That means that domain adaption approaches are valuable for quickly preparing classification models for new topics. But the best performance can be achieved only with annotated training data so far.

## 4.8 INVESTIGATION OF THE GERMAN HATER COMMUNITY ON TELEGRAM

### 4.8.1 *Motivation*

The studies so far relied on data collected from Twitter, Reddit, Wikipedia, 4chan, and 8chan. Minor parts of some used datasets also stem from Facebook and YouTube. A platform that has not been considered yet is Telegram. But why is Telegram relevant? The platform founded in 2013 is a mixture of instant messenger and social media (Telegram, 2021). On the one side, it offers encrypted private chats and group chats, similar to WhatsApp (Urman and Katz, 2020). On the other side, it is possible to create public group chats and channels<sup>4</sup> that users can easily find via a search and join (Urman and Katz, 2020). According to its FAQs, Telegram aims to protect privacy and freedom of expression (Telegram, 2021). These principles, combined with the classical messaging feature and the social media components, attracted many hate actors that were deplatformed from traditional social media platforms (Fielitz and K. Schwarz, 2020; Rogers, 2020; Urman and Katz, 2020). Deplatforming means that users are banned from social media platforms due to violating their policies (e. g., spreading misinformation or hate speech) (Rogers, 2020). Examples are the far-right, conspiracy theorists, and COVID-19-deniers (Fielitz and K. Schwarz, 2020; Holzer, 2021; Owen, 2019; Rogers, 2020; Urman and Katz, 2020). So, the instant messenger became one of the "darker corners of the Internet [sic]" (Rogers, 2020, p. 216) and consequently an attractive source for abusive language detection research.

Telegram has not been received much attention concerning abusive language detection, though. For an empirical study on public Telegram channels, Rogers (2020) used an abusive language classifier based on keywords from hatebase.org (a lexicon of hate speech terms). Solopova, Scheffler, and Popa-Wyatt (2021) released the first labeled abusive language datasets with Telegram messages. The 26,431 messages stem from a channel of Donald Trump supporters. In general, research on Telegram is limited, but it seems to have recently gained more attention from researchers.

The following study addressed this gap. In the study, abusive language classification models for German Telegram messages and channels were developed based on datasets gathered from other social media platforms to examine the generalizability of such models. Moreover, the models were used to analyze the German hater community on Telegram.

### 4.8.2 *Study XII<sup>†</sup>*

In the study "Introducing an Abusive Language Classification Framework for Telegram to Investigate the German Hater Community" (Wich, Gorniak, et al., 2022), we developed classification models to detect abusive messages

<sup>4</sup> A channel is comparable to a news feed. Its administrators can send messages to its subscribers. But the subscribers cannot send a message in the channel.

and hater channels on Telegram. Additionally, we used the models to analyze further the collected data from the German hater community on Telegram. Since there were no abusive language datasets containing Telegram messages, we decided to use existing datasets from other social media platforms for training our message classification models. The study addressed the following four research questions:

- RQ1 "Can existing abusive language datasets from other platforms be used to develop an abusive language classification model for Telegram messages?" (Wich, Gorniak, et al., 2022, p. 2)
- RQ2 "How did the prevalence of abusive content evolve in the last years on Telegram?" (Wich, Gorniak, et al., 2022, p. 2)
- RQ3 "Can a classification model be used to predict whether a Telegram channel is hateful or not?" (Wich, Gorniak, et al., 2022, p. 2)
- RQ4 "Can we leverage the topical distribution and graph embeddings to derive meaningful clusters from channels?" (Wich, Gorniak, et al., 2022, p. 2)

Before answering the research questions, we had to collect the messages from the German hater community on Telegram. We applied a snowball sampling strategy using a list of 51 German hate actors as seeds, which was provided by Fielitz and K. Schwarz (2020).

To answer RQ1, we trained six classifiers on existing German abusive language datasets from other platforms (mainly Twitter), using pre-trained BERT models. We evaluated the models with a sample of 1,149 Telegram messages that we annotated. Furthermore, we used Google's Perspective API to have an external benchmark for our models. The annotated dataset also helped us to find the best combination of the six classifiers. The idea behind combining the models was that the models cover various aspects of abusive content. By combining them, we could increase the classification performance. The best performing combination of the models was used to answer the other three research questions. Regarding RQ2, we employed the combined models to classify all collected German Telegram messages. That classified data allowed us to examine the prevalence of abusive content in the subnetwork around popular German hate actors. RQ3 was addressed by building a classification model for channels based on graph embeddings and a topic model. The labels for the channels in the training and test sets were derived from the messages classified by the combined models. To answer RQ4, we reused the graph embeddings and the topic model from RQ3 to cluster the channels.

At the end of the data collection process, we gathered 13,822,605 messages from 4,962 Telegram channels. 5,421,845 (39.2%) of all collected messages are written in German. In 2,478 channels, German was the most or second-most used language. Regarding the message classification model (RQ1), the best performing model achieved an F1 score of 54.95% on the abusive



class and a macro F1 score of 71.91%, outperforming Google’s Perspective API. Combining all six models based on majority voting principle slightly improved the classification performance (a text is classified as abusive, if at least four of the six classifiers vote for abusive). Based on the classifications of the combined models, we observed that the overall prevalence of abusive content within the collected subnetwork rose from 2.4% to 3.4% between January 2019 and February 2021. In regard to RQ<sub>3</sub>, the classification model for hateful channels achieved an F1 score of 64.9% for the hater class and a macro F1 score of 74.2%, which is comparable to similar experiments on Twitter accounts (Li et al., 2021; M. T. Ribeiro, Singh, and Guestrin, 2016). One outcome of the analysis addressing RQ<sub>4</sub> was the similarity matrix of the seed channels showing the topical overlap between the channels, depicted by a cluster heatmap in Figure 4.10. It helps identify clusters based on topical similarity and connectedness in the network. The upper left cluster<sup>5</sup>, for example, consists mainly of alternative news channels, while the large one in the center contains far-right hate actors.

In summary, the study was the first one to build an abusive language classification model for German Telegram messages and channels and to examine the German hater community on this platform.

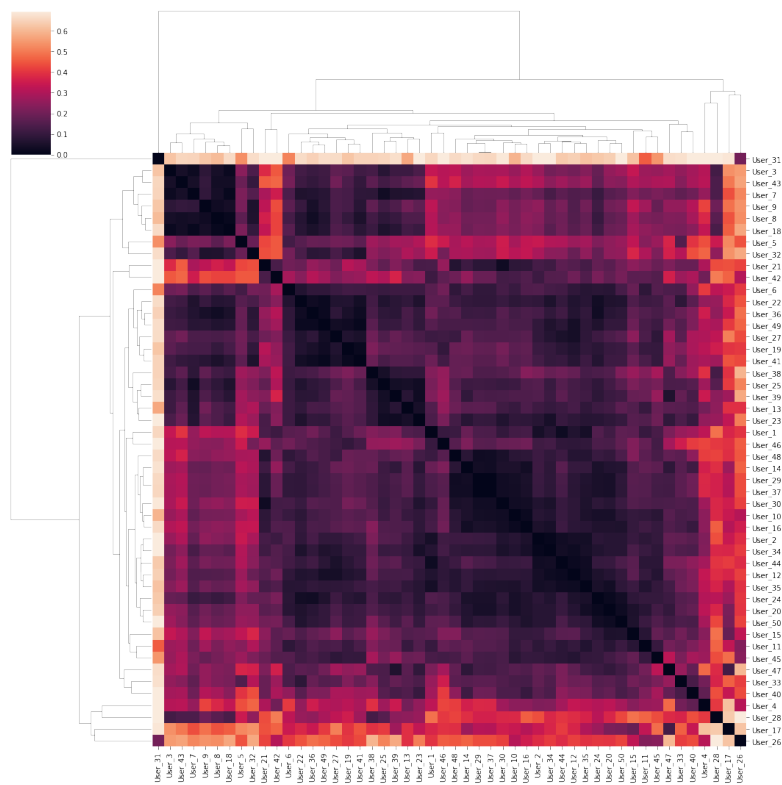


Figure 4.10: Topical similarity and connectedness in the network between seed Telegram channels (figure from Wich, Gorniak, et al., 2022, p. 7).

<sup>5</sup> clusters are the dark red/black areas

### 4.8.3 Discussion

The approach proposed in the study helps to train an abusive language classifier for a new platform based on data gathered from other platforms. We even outperformed Google’s Perspective API, a service to detect toxic comments that is in production (Google, 2021). However, the models still performed better on the test data from the original platform than on the evaluation set from Telegram. That means that the models generalize to a certain degree but with limitations—similar to the observation that was discussed in Section 4.7. Consequently, the approach is suitable to build a classification model for a new platform with a decent performance as long as no platform-specific training data is available. It would be helpful to apply semi-supervised or unsupervised techniques to improve the performance with less or no manual effort in future work. One idea is to pre-train the BERT model with Telegram messages so that it is used to process such messages (Konle and Jannidis, 2020). Alternatively, a small set of labeled data from the target platform can be combined with weakly or semi-supervised approaches to produce training data from the target platform (L. Gao, Kuppersmith, and R. Huang, 2017; Rosenthal et al., 2021).

The decrease of the classification performance drop between data from Twitter and Telegram can be explained by the fact that abusive language differs between these platforms. On Telegram, the users are in their echo chambers with like-minded people because one user has to actively join a group or channel. On Twitter, there are also echo chambers. But users also get displayed tweets from people with other opinions when someone from their echo chamber likes or replies to such a tweet. Consequently, personal attacks (e. g., insulting another user) should appear more often on Twitter than on Telegram. Such characteristics should be investigated in future work.

The dataset annotated to evaluate the models’ performance is a further valuable contribution to research. It is the first German abusive language dataset containing Telegram messages. However, it has two limitations. First, it consists of 1,149 messages, which is smaller than most other German abusive language datasets. Second, it uses the same binary annotation schema as the COVID-19 dataset from Section 4.7. That means that the labels are less fine-grained than the ones of other German datasets (Mandl, Modha, Kumar M, et al., 2020; Mandl, Modha, Majumder, et al., 2019; Struß et al., 2019; Wiegand, Siegel, and Ruppenhofer, 2018). These limitations did not affect the study because the dataset was only used for evaluation and classifiers were trained on a binary task. No classifier was trained on this dataset. However, if the dataset is used for other purposes in the future, the limitations can be relevant.

Regarding the prevalence of abusive content, an increase from 2.4% to 3.4% was observed between January 2019 and February 2021. The average prevalence was 3.1% for this period. Considering only the seed channels classified as haters by Fielitz and K. Schwarz (2020), we observed a prevalence of 5.3%. However, the numbers have to be treated with caution as the classification

models used to detect abusive content still have room for improvement. If we consider the context, the numbers make sense, though. The share of abusive content on Twitter is estimated at up to 3% (A. Founta et al., 2018). The overall prevalence in the collected Telegram channels is in the same order of magnitude. The higher prevalence in seed channels should not surprise anyone as they were classified as hate actors and Telegram does not moderate content on its platform in contrast to Twitter. This finding combined with the increase of the overall prevalence shows that Telegram has a problem with abusive content, which our society may not neglect.

The classification model to detect hateful Telegram channels could be a helpful contribution to this challenge. It demonstrated that it could identify hater channels based on the network structure and the topics of the channels' messages. Such a model can be used to map the opaque landscape of Telegram channels. The model is not the first one that aims to identify hateful accounts on social media platforms based on graph embeddings (Das et al., 2021; Li et al., 2021; M. Ribeiro et al., 2018). But it is the first one that applies this to Telegram with comparable performance.

To sum up, Telegram is a melting pot for hate actors, especially in Germany. The study's findings show that the research community should pay more attention to the messenger and examine the abusive content distributed there. This is necessary to avoid Telegram becoming a larger hotbed of extremists and haters than it is already.



## DISCUSSION

---

The presented studies addressed various weak points of abusive language detection on social media platforms. In this section, the studies' results are overarchingly discussed, and it is outlined how the studies contribute to the research objectives, as indicated in Figure 4.1.

### 5.1 LIMITED PERFORMANCE AND GENERALIZABILITY

The first problem is limited performance and generalizability of abusive language classifiers. One research objective that targets this problem is to increase quantity and quality of training data (A), addressed by five studies address. Study I conducted an in-depth analysis of alt-right fringe communities from Reddit, 4chan, and 8chan. The findings contribute to a better understanding of how hate is expressed in these fringe communities and how manifold it is. Furthermore, the study assembled a fine-grained abusive language dataset collected from platforms that were hardly covered by other datasets. Studies VIII, XI, and XII also produced abusive language datasets, but each addressed a different weak point. The dataset of Study VIII is one of a few that comprise social network data besides the (pseudo-)labeled tweets. Study XI focused on COVID-19-related abusive language—the first German dataset with such a focus. The dataset produced by Study XII contains messages from the messenger platform Telegram, which has not received much attention from the research community yet. But all three have in common that tweets and messages are in German. Even though these three datasets would have benefited from a more granular labeling schema than the binary one, they are a valuable contribution to the research community. Study II has not produced any dataset. But it provides a helpful tool, the bias and comparison framework, to examine the quality of abusive language datasets. A topic that has to be discussed in this context is the claim for a "commonly accepted benchmark corpus" (A. Schmidt and Wiegand, 2017, p. 7). Based on the findings from the conducted studies, a fixed benchmarking corpus would not improve abusive language detection. Such a dataset would not reflect current trends of abusive language and would be enormously large to cover all facets. A better solution could be a framework that combines existing datasets and can be dynamically extended by new datasets. The framework would have a labeling schema so that it can bring together abusive language datasets with deviating schema. A hierarchical approach would be the best as it provides a certain degree of flexibility. Additionally, it would contain a scoring system that considers the datasets' various peculiarities (e. g., different sizes, class imbalance) and calculates an overall classification performance. To enable the comparability of the results, particular sets of abusive language

datasets would be defined and updated regularly. Researchers would report their results based on these sets, making the results comparable and more expressive because the sets cover a broad range of abusive language. The first step in this direction was made by Risch, P. Schmidt, and Krestel (2021) who published a collection of more than 40 abusive language datasets. But aspects like a scoring system or a way to combine labeling schema are still missing.

Four studies pursued Research Objective B, which aims to integrate context data into the classification models to improve performance. The first one is the previously mentioned Study VIII that provides a dataset with social network data and the corresponding data collection methodology to build it. The study is a necessary interim stage because almost no abusive language datasets containing social network data are available. The dataset enables examining the integration of social network data into the abusive language classification. In Study IX and X, classification models have been developed that investigate the benefit of incorporating social network data and the users' previous posts. The results show that these additional data sources improve classification performance. The approach of the classification model from RQ<sub>3</sub> of Study XII is different from the other two studies. The model classifies a Telegram channel as a hater or non-hater instead of a post or another text form. It uses only the relations between the channels (social network) as input features. This approach intends to detect users who systematically distribute abusive content based on patterns in the network and not on textual data.

The latter study investigating the German hater community on Telegram also contributes to improving generalizability of classification models (Research Objective C). Since no German abusive language dataset from Telegram was available, a classifier for Telegram messages was developed that combines models trained on existing datasets from other social network platforms. The approach helps to build classification models that can be used on platforms other than the one where the training data is from. The results are promising, but there is room for improvement. The research community should pay more attention to improving generalizability of the models. Due to deplatforming activities of traditional social media platforms, many hate actors move to alternative platforms (Fielitz and K. Schwarz, 2020; Rogers, 2020). Since these sometimes differ concerning the form of communication (e. g., tweets are limited to 280 characters, while Telegram has a limit of 4,096 characters), it would be beneficial to build more robust and generalizable classifiers, which can handle text data from various platforms. Otherwise, abusive language datasets that have been annotated with a lot of manual effort might become useless in the future.

## 5.2 MISSING TRANSPARENCY AND INTERPRETABILITY

After discussing how the studies address the problem of limited performance and generalizability, the following paragraphs deal with the studies concerning the problem of missing transparency and interpretability. The first research objective that approaches this problem is to uncover unintended bias within training data and models (D). Six studies examined unintended bias from different perspectives and contribute to this objective. Study II's bias and comparison framework provides a toolset to identify various bias types in abusive language datasets. Researchers had already investigated different forms of unintended bias. But no one had provided a tool to examine and compare datasets systematically, which is done by Study II. However, the framework does not assert the claim to be exhaustive. The subsequent studies on annotators bias (III, IV, and V) show that the framework partially neglects this kind of bias. The methods based on the bias matrix proposed by Study V could be a reasonable extension of the framework to alleviate the gap. The approaches in Study III and IV provide interesting findings. However, they are less appropriate for the framework. The approach based on the annotators' demographic characteristics (Study III) requires collecting personal data from the annotators, which is usually not done to ensure privacy. The graph-based approach (Study IV) needs many annotators, making the method not applicable for datasets with few annotators. Methods using the bias matrix do not exhibit these limitations. Therefore, they are suitable for the framework. Study VI investigated the impact of politically biased data on abusive language classification. The results show that such a bias can impair classification performance. However, these findings have to be treated with caution because the political bias was only simulated. The study can be rather seen as a pre-study to estimate potential impact. But the results are promising, which is why this phenomenon should be examined in a future study, as proposed in 4.4.3. The last of the six studies addressing unintended bias is Study VII, which examined annotator bias from an ethical perspective. It is a valuable addition because annotator bias is not a purely technical issue. It is related to ethical challenges, such as the ethical requirements for annotating abusive language datasets. Therefore, this study completes the examination aiming to uncover unintended bias.

The second research objective addressing the missing transparency and interpretability is to make predictions of abusive language classification models more understandable for humans (E). Four of the twelve studies share this objective. The bias and comparison framework proposed in Study II contains a method using the SHAP framework to explain predictions. The method aims to compare classification models trained on different abusive language datasets on a text document level. Users of the framework can find differences in the relevance of single words between the classifiers. Study VI, which investigated the impact of political bias, applied the SHAP framework for a similar purpose. It used explanations of predictions to compare politically biased classification models and identify differences in how the models

weigh single words concerning the prediction. This capability improves the granularity of comparing the biased models. Without XAI, it would only be possible to compare the predictions on a text document level, impairing the examination of political bias. While both studies used the SHAP methods out of the box, Study IX and X developed explainable methods based on SHAP to handle the multimodal abusive language classification models that leverage social network data and previous posts of users. Comparing the explainable capabilities from these two studies, we observed that Study X is more enhanced because it can handle more complex models (e. g., GraphSAGE) and provides more granular explanations (e. g., the relevance of specific edges in the social network). These methods are linked by the fact that they are local explainability approaches, meaning they explain only single instances and not the entire model. This gap is addressed by the t-SNE-based explanation method proposed by Study IX. However, local and global explanations methods, especially for NLP use cases, are in an early stage, as shown by the studies. They produce reasonable explanations but require further research and development to become more accurate and user-friendly. Nevertheless, the four studies contribute to the objective of making predictions of abusive language models more understandable for humans, as demanded by many researchers (Kiritchenko, Nejadgholi, and Fraser, 2021; Mishra, Yannakoudakis, and Shutova, 2020, 2021; Vidgen, Harris, et al., 2019).



## CONCLUSION

---

The overarching goal of the dissertation was to improve abusive language classification regarding various perspectives. Abusive language classification may not be considered a pure engineering competition with the aim of pushing the F1 score or accuracy to the next level. It also comes with ethical challenges because a system that is meant to become part of our daily lives faces more requirements than a decent F1 score (e.g., fairness and transparency). That is why the dissertation focuses on two problems to solve: (1) limited performance and generalizability and (2) missing transparency and interpretability of the classification models. Twelve studies were conducted that approached various weak points of the abusive language classification models to contribute to solving these problems.

Five studies help increase the quantity and quality of abusive language datasets by analyzing hate speech from fringe communities, creating new datasets, and providing a framework to analyze and compare such datasets. Additionally, four studies improve the classification performance by integrating context data (e.g., social network data, previous users' posts) into the classification models as additional features. One study focused on enhancing the generalizability of the models so that good classification performance can be achieved for messages or comments from other platforms. The findings of these studies are mainly conducive to solving the first problem.

To approach the second problem, six studies deal with uncovering unintended bias within abusive language datasets. They provide methods to identify and measure various bias types (e.g., annotator bias, political bias) and their impact on the classification. However, one of these studies differs from the rest because it is an ethical examination of the annotator bias and its influence. Two of these six studies and another two studies combine XAI techniques with the classification models to make the predictions more understandable for humans.

These twelve studies address various weak points of abusive language detection and contribute to resolving them. They also lay the foundation for interesting future work. The results of integrating context data into abusive language classifiers show that this approach is promising because it circumvents the hurdles of a purely text-based classification (e.g., implicit language, sarcasm, irony). Another aspect that should be followed up by the research community is the development and improvement of XAI techniques for such classifiers as it is necessary to earn the users' trust and acceptance. Furthermore, researchers should continue to uncover unintended bias in the datasets and find ways to mitigate it.

This dissertation contributes not only to academia but also to provide solutions for a challenge of today's society. The findings of the conducted

studies support the fight against online hate to make the internet safer and friendlier.

## APPENDIX



## PUBLICATIONS •

---

Publications in Appendix A are relevant for examination in accordance with Exhibit 6 of the regulations for the award of doctoral degree.

### A.1 STUDY II

©2021 The Author(s), published under Creative Commons CC-BY 4.0 License<sup>1</sup>.

Maximilian Wich, Tobias Eder, Hala Al Kuwatly, and Georg Groh (July 2021). "Bias and comparison framework for abusive language datasets." In: *AI and Ethics*. ISSN: 2730-5961. DOI: [10.1007/s43681-021-00081-0](https://doi.org/10.1007/s43681-021-00081-0)

---

<sup>1</sup> <https://creativecommons.org/licenses/by/4.0/>

### *Publication Summary*

"Recently, numerous datasets have been produced as research activities in the field of automatic detection of abusive language or hate speech have increased. A problem with this diversity is that they often differ, among other things, in context, platform, sampling process, collection strategy, and labeling schema. There have been surveys on these datasets, but they compare the datasets only superficially. Therefore, we developed a bias and comparison framework for abusive language datasets for their in-depth analysis and to provide a comparison of five English and six Arabic datasets. We make this framework available to researchers and data scientists who work with such datasets to be aware of the properties of the datasets and consider them in their work." (Wich, Eder, et al., 2021, p. 1)

### *Author Contributions*

Maximilian Wich headed the research project. He developed the initial idea, the methodology, and the framework. Furthermore, he proposed a large number of used analysis methods and contributed to the implementation of the framework's code. Additionally, he wrote most of the manuscript. Tobias Eder implemented the framework as part of his master's thesis that Maximilian Wich supervised. Additionally, he supported the writing process of the paper. Hala Al Kuwatly supported the analysis of the Arabic datasets. Georg Groh regularly discussed the ideas and concepts with the team and provided feedback on the study.



# Bias and comparison framework for abusive language datasets

Maximilian Wich<sup>1</sup> · Tobias Eder<sup>1</sup> · Hala Al Kuwatly<sup>1</sup> · Georg Groh<sup>1</sup>

Received: 5 May 2021 / Accepted: 8 July 2021  
© The Author(s) 2021

## Abstract

Recently, numerous datasets have been produced as research activities in the field of automatic detection of abusive language or hate speech have increased. A problem with this diversity is that they often differ, among other things, in context, platform, sampling process, collection strategy, and labeling schema. There have been surveys on these datasets, but they compare the datasets only superficially. Therefore, we developed a bias and comparison framework for abusive language datasets for their in-depth analysis and to provide a comparison of five English and six Arabic datasets. We make this framework available to researchers and data scientists who work with such datasets to be aware of the properties of the datasets and consider them in their work.

**Keywords** Hate speech detection · Abusive language detection · English · Arabic · Bias

## 1 Introduction

The last few years have seen an increase in popularity for abusive language detection as a classification problem. This growing interest brought along the release of a more significant number of labeled datasets. Although this increase in the available data has made research more accessible, no real benchmark dataset for abusive language detection has been established with a unique set of problems in the domain, chiefly encompassing comparability issues between systems trained and evaluated on various datasets [25, 38]. These problems emerge from differences between the datasets, such as context, platform, sampling process, and labeling, with even the task definition being subtly different in many cases [14].

A further aspect that impairs the dataset comparability is biased data. We define bias as a phenomenon in which a system “systematically and unfairly discriminate[s] against

certain individuals or groups of individuals in favor of others” [17, p. 332]. In the context of abusive language, bias can be materialized in different forms. One example is topic bias [47]. Let us assume that we have an abusive language dataset with a neutral and an abusive class. If the abusive class is dominated by a particular topic that is not abusive per se (e.g., sports) and the neutral class does not contain many documents about this topic, a classification model learns to use terms from this topic to distinguish between both classes [47]. Consequently, the classifier systematically discriminates documents related to this topic. Therefore, it is necessary to uncover bias in datasets and make them transparent.

Recently, frameworks for documenting datasets characteristics have been proposed, such as [18] and [5]. Transparency in the processes to create new datasets can be realized using the guidelines outlined in these frameworks, but they do not solve the problem for existing datasets and comparisons beyond the mostly discreet metrics within specific tasks. Even where such information is available for a single dataset, it is hard to quantify how differences in data collection or labeling choices manifest themselves in the systems built on top of them. In the worst case, this blind spot can lead to strongly biased systems, which inherit some of the systemic problems stemming from the underlying data without this becoming evident from evaluation according to common metrics. Consequently, the further use of these systems is also highly problematic from an ethical standpoint. Without insight into the training data’s actual properties and

---

✉ Maximilian Wich  
maximilian.wich@tum.de

Tobias Eder  
tobias.eder@in.tum.de

Hala Al Kuwatly  
hala.kuwatly@tum.de

Georg Groh  
grohg@in.tum.de

<sup>1</sup> Technical University of Munich, Munich, Germany

distribution, there are no guarantees that the system performs fairly in a real-world setting.

Therefore, the paper aims to provide a framework to compare abusive language datasets and uncover their inherent properties (e.g., different forms of bias). Furthermore, we can help the research community bring order and structure to the variety of abusive language datasets by providing two main contributions:

1. A structured framework for analyzing and comparing abusive language datasets across various fine-grained metrics. The chosen metrics apply across multiple dimensions, capturing meta-information, semantic information, annotations, and derivative measures based on state-of-the-art classification evaluated on these datasets. The framework is not limited to the English or Arabic language. It can also be applied to datasets in other languages.
2. The paper provides an excellent comparison of five English and six Arabic datasets from the abusive language domain, which illustrates their differences, highlights their focus, and reveals potential biases and hidden properties.

The paper is structured as follows: Sect. 2 discusses related work and explains why our work fills a demand that other researchers have not covered. In Sect. 3, we describe our developed framework and outline the reasons for adding the various methods. Afterward, we introduce our data selection for the two case studies: (1) English and (2) Arabic datasets. The results of the case studies are presented in Sects. 5 and 6. Section 7 contains a discussion about our findings and challenges of current and prospective abusive language datasets. Finally, in Sect. 8, we conclude our work.

## 2 Related work

The growing number of abusive language datasets has led to a range of dataset surveys in recent years. However, most of the early research on hate speech or abusive language data was done as part of an overview of the emerging field's methodology, including reviews such as [14, 38], and [24] discussing key properties of a few selected datasets used in existing systems.

A more comprehensive study on abusive language datasets was published by [39], compiling 51 datasets. They proposed a more involved descriptive framework, including information on the target of abuse, the level of annotation, and the class distribution. Further surveys followed, including [32] on 49 datasets, which was the first to include a short specific lexical analysis of the dataset contents to identify topic bias. [25] has recently given an overview of 17 datasets

to evaluate them on their ability to function as benchmark datasets in the abusive language domain, assessing aspects such as availability, class imbalance, exact task definition, and label conflation. All dataset surveys have in common that they conduct a high-level comparison (e.g., number of documents, source, and data collection strategy) and do not look beyond the surface except the limited lexical analysis in [32]. Consequently, these surveys are satisfactory for identifying in broad strokes how different datasets compare on an annotation level. However, they do not provide details of the dataset contents, as they rely mainly on second-order descriptions about the data, principally compiled for the release of a specific dataset.

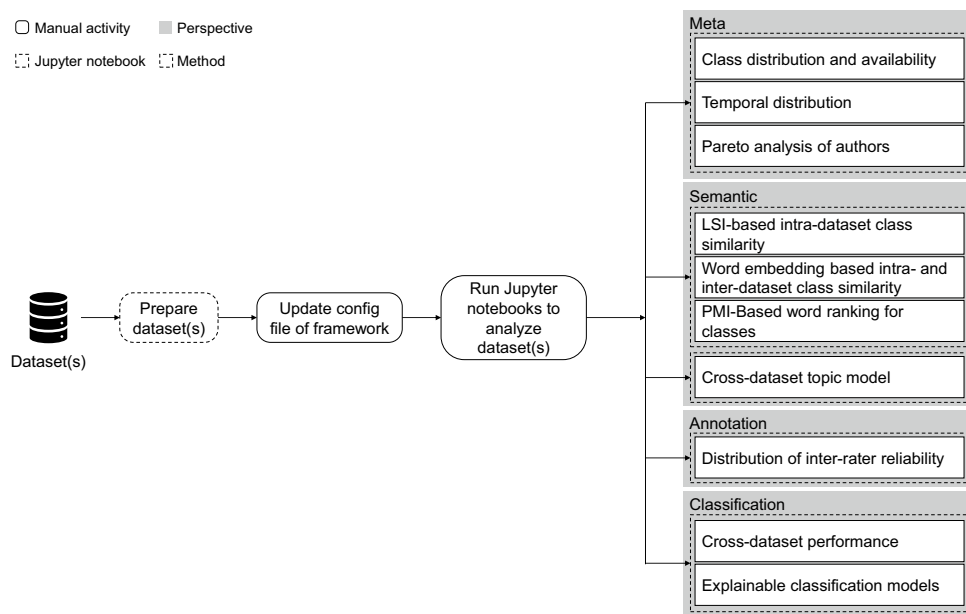
[18] and [5] proposed datasheets for datasets to document their characteristics because the machine learning or NLP communities do not have a standardized approach. These data sheets are necessary and make it easier to compare datasets. However, they cannot be applied to already published datasets, and in-depth analysis and comparison of different data sets are not possible. The type of work required is necessary to be done by the original authors of the dataset. Furthermore, the only recourse for a practitioner trying to work with a dataset for which no datasheet was released would be to contact the original authors and ask about the specific creation information. Furthermore, they do not reflect all characteristics of abusive language datasets, being a very general framework for all data types.

A range of other research has dealt with evaluating specific datasets or systems to uncover bias problems with the underlying data. [30] evaluated gender bias on the [44] and [16] datasets. [13] investigated unintended bias with respect to identify terms and proposed a method to debias the training data. [9] reported problems with the association of minority group language with hate in their data, while [47] have done work on the influence of different biases in the sampling of popular abusive language datasets (e.g., topic and author bias). [46] analyzed how political bias influence hate speech classification models. [37] proposed social bias frames, which is a formalism that “aims to model the pragmatic frames in which people project social biases and stereotypes onto others” [37, p. 1]. Another form of bias in abusive language datasets that researchers have addressed is annotator bias. [36] investigated annotator bias concerning the Afro-American English (AAE) dialect. They showed that a classifier trained on a standard abusive language dataset discriminates documents in AAE. [1] identified annotator bias by splitting annotators according to their demographic characteristics. The challenge of this approach is that it requires demographic data of the annotators. [45] addressed this problem by identifying annotator bias purely on similarities in the annotation behavior. Lastly, [15] compared six popular datasets according to their differences in class labeling via similarity in a common word embedding



**Table 1** Bias framework for abusive language datasets

Perspective	Method	Problem
1. Meta	(a) Class distribution and availability	Degradation
	(b) Time distribution	Temporal bias
	(c) Pareto analysis of authors	Author bias
2. Semantic	(a) LSI-based intra-dataset class similarity	Similarity/dissimilarity of classes
	(b) Word embedding based intra- and inter-dataset class similarity	Similarity/dissimilarity of classes
	(c) Cross-dataset topic model	Topic bias
	(d) PMI-Based word ranking for class	Topic bias
3. Annotation	(a) Distribution of inter-rater reliability	Annotator bias
4. Classification	(a) Cross-dataset performance	Generalizability
	(b) Explainable classification models	Generalizability



**Fig. 1** Overview of the framework’s methods and the required data

space and further classifying them with the Perspective API framework.

To the best of our knowledge, no one has developed a framework to conduct an in-depth comparison of abusive language datasets focusing on various forms of bias.

### 3 Framework

Since there is a need for systematical in-depth analysis and comparison of abusive language datasets going beyond high-level properties, we propose the following framework. It consists of three perspectives that contain methods addressing various challenges. Table 1 provides an overview of the framework and the challenges that are addressed by the methods. Figure 1 displays the steps how the framework is

applied. We also published our code<sup>1</sup> to encourage researchers to use the framework for their research. The framework is a collection of Python scripts and modular Jupyter notebooks that contain the individual parts of the analysis and use a unified framework to handle data input for all parts of the analysis.

<sup>1</sup> Code (including documentation and instruction for installation) available on GitHub: <https://github.com/mawic/abusive-language-dataset-framework>

### 3.1 Meta perspective

The first perspective focuses on the metadata of the documents within the datasets. It leaves the textual data out and emphasizes aspects, such as class distribution or author distribution.

#### 3.1.1 Class distribution and availability

The first method investigates the class distribution and availability of data. We do this since some datasets, especially those collected from Twitter, often contain only references to the documents (e.g., tweet IDs) due to platform policies' restrictions. Sharing only references is problematic because documents on online platforms can be deleted over time [41]. Particularly, documents with hateful or abusive content are prone to removal since they often violate the platform policies. This degradation impairs both the quality and quantity of datasets. To avoid degradation, some researchers publish datasets containing the text—sometimes anonymized (e.g., removing usernames from the documents)—instead of a reference to the original resource. It solves the degradation issue but exacerbates other data analyses (e.g., temporal or author distribution). Therefore, we include the analysis of class distribution and availability in our framework.

#### 3.1.2 Temporal distribution

Another challenge of abusive language detection is evolving language [14, 33, 41]. Words and expressions that are unproblematic today might have an abusive connotation tomorrow. Consequently, a classification model trained on an older dataset can perform worse on new datasets because the model does not recognize new language patterns [41].

If the collected data was created quickly (e.g., only in a few weeks), it can indicate that the abusive data contains only current abusive language patterns and covers only current topics (e.g., refugee crisis). As a result, classification models trained on such a dataset might perform worse on other datasets from other periods or with another topical focus. Thus, it is interesting to investigate when the documents were created and to identify a temporal bias.

#### 3.1.3 Author distribution

An aspect that is also of interest is whether the dataset has an author bias [47]. That means that a small number of users created a large portion of documents from one or more classes. The problem of author bias is that a classification model trained on such a dataset tends to memorize the author's writing style or the topics they are writing about but not actual indicators of hateful language [47]. Therefore, we propose a Pareto analysis of the authors combined with

a class distribution to make this transparent and uncover author bias. It is a method based on quality management and supports root cause analysis [43]. In our case, we count the number of documents for each user and rank them according to the number of documents. Consequently, we can figure out whether a large portion of documents is produced by a small number of authors, signifying author bias.

### 3.2 Semantic perspective

After investigating the metadata, we focus on the semantic level of the datasets. Then we analyze the class similarities within and across the datasets and the topics addressed by them.

#### 3.2.1 LSI-based intra-dataset class similarity

Before explaining the method, we explain why class similarity is relevant to our framework. Firstly, the more similar documents within a class are, the easier it is for a classification model to distinguish it from other classes. Secondly, the more dissimilar the two classes, the easier it is for a classifier to distinguish between both. Therefore, similarity scores can also act as an indicator of the classifier's generalizability.

The first method focuses on the inter- and intra-class similarities within a dataset. It applies Latent Semantic Indexing (LSI) [11] and uses cosine distance to compute the similarity within a class and between the classes because it does not require any pre-trained word embeddings, and we do not want to rely on word embedding based methods in this perspective. The result is a matrix for each dataset, representing the homogeneity and similarity of the dataset's classes. The findings are comparable between different datasets, but they do not demonstrate the similarity of different datasets classes, which is addressed by the following method.

#### 3.2.2 Word embedding based inter- and intra-dataset class similarity

In order to compare the class similarities across the datasets, we apply a variant of the method proposed by [15]. After preprocessing (e.g., removing URLs, usernames), we use a pre-trained FASTTEXT embedding depending on the dataset's language to compute a document vector for each document [26]. This step is slightly different from Fortuna et al.'s methods [15]. Instead of averaging the word vectors of a document to get the document vector, we use FASTTEXT's sentence embedding feature. Then, all word vectors of a class are averaged, obtaining a centroid for the class. In the last step, Principle Component Analysis (PCA) [31] is applied to compute a 2-dimensional representation, visualizing the similarity between the classes across the datasets, as proposed by [15].

### 3.2.3 PMI-based word ranking for classes

The third method of the semantic perspective produces a listing of the most relevant terms for each class of a dataset proposed by [47]. The intention is to provide an impression of what the class is about and what classification models learn. In order to calculate the relevance of the terms, we use pointwise mutual information (PMI) [8]. However, instead of computing the PMI between two words, our pair consists of the word  $w_i$  and class  $c_j$ :

$$pmi(w_i, c_j) = \log \frac{p(w_i, c_j)}{p(w_i)p(c_j)} \quad (1)$$

As a result, we obtain a value representing the relevance of the word for the class. We can identify the most relevant terms by ranking them. However, this method does not describe the class to its full extent but is a good indication.

### 3.2.4 Overarching topic modeling

The fourth method of the semantic perspective analyzes topic bias in datasets. It is often caused by the way how the datasets are collected. Some abusive language datasets, for example, are gathered through a keyword-based approach (e.g., hashtag-based filtering of tweets). But if the keywords are too specific, the resulting dataset can exhibit topic bias. The focus on topic bias is motivated as follows: If an entire dataset focuses on one or a few specific topics, the model's generalizability is impaired, meaning it performs poorly on other datasets. Let us assume that we have an abusive language dataset that mainly contains COVID-19-related content. A model trained on the dataset might perform worse on an abusive language dataset with a sports focus. Therefore, it is necessary to identify topic bias in a dataset. We suggest the following topic model-based method to investigate this phenomenon.

In the first step, we sample  $n$  documents from each of the  $m$  datasets to be analyzed and merge into one dataset. Two different sampling strategies are proposed: (1) sampling according to the actual class distribution, (2) sampling an equal number from each class. The first one delivers a more representative result, while the second gives more weight to underrepresented classes. In the second step, we use CluWords to generate a topic model of the merged dataset with  $l$  topics. CluWords is a topic model algorithm that uses word embeddings and non-probabilistic matrix factorization and works well on short texts [42]. As word embedding, we use the same FASTTEXT model as in the previous method. Besides the  $l$  topics, CluWords outputs a one-dimensional vector for each document, signifying the document's topic distribution. Additionally, we generate for each topic a

one-dimensional vector as representation. In the third step, we apply t-SNE to project the  $l$ -dimensional vectors of the documents and the topic centroids to a two-dimensional representation [23]. After coloring each document data point depending on its dataset, we can use the plot to visualize the topic distribution and uncover topic bias.

## 3.3 Annotation perspective

The third perspective deals with the annotations of the data provided by humans. As studies have shown [1, 36, 37, 45], biased annotations can impair classification performance. Consequently, it must be addressed by our framework.

### 3.3.1 Distribution of inter-rater reliability

We recommend examining the distribution of the annotator's inter-rater reliability to uncover potential annotator bias. Low inter-rater reliability implies "systematically biased coders" [35, p. 673]. Therefore, we suggest analyzing the overall inter-rater reliability of a dataset and the individual inter-rater reliability of each annotator. The overall metric indicates the quality of the annotations and a potential annotations bias. Moreover, the individual metrics help us understand whether a general disagreement between the annotators causes low inter-rater reliability or a few strongly biased annotators. Krippendorff's alpha is utilized as an inter-rater reliability metric because it can handle missing annotations where each annotator's vote is required to conduct the analysis [19]. However, most datasets only provide an aggregated gold standard, making it impossible to apply this method.

## 3.4 Classification perspective

The fourth perspective compares and investigates the classification models separately trained on the different datasets and evaluated on all test sets. The goal is to assess the generalizability and to identify blind spots of the underlying datasets.

### 3.4.1 Cross-dataset performance

The goal of abusive language detection research is to build classification models that reliably detect abusive language. One key component to reach this is training data that covers the diversity and multifacetedness of abusive language. Using such data, we can build more generalizable models. This aspect is related to bias in a dataset: the more significant and stronger the bias is in a dataset, the less generalizable its trained model. Therefore, we integrate it into our framework and propose the following method to compare the generalizability of datasets.

**Table 2** Selected abusive language datasets (class names in bold are the abusive categories)

Lang.	Name	Source	Size	Labels	Ref.
English	Waseem	Twitter	16,907	None, sexism, racism	[44]
	Davidson	Twitter	24,783	Offensive, hate, neither	[10]
	Founta	Twitter	99,996	Normal, abusive, hateful, spam	[16]
	Zampieri	Twitter	14,100	Hierarchical labels: (1) not offensive, offensive (2) if offensive: targeted insult, untargeted insult (3) if targeted: individual target, group target, other	[48]
	Vidgen	Twitter	20,000	Hostility, criticism, counter speech, discussion of East Asian prejudice, neutral	[40]
Arabic	Alsafari	Twitter	5341	3-class: clean, offensive, hate; 6-class: clean, offensive, religious hate, gender hate, nationality hate, ethnicity hate	[3]
	Alshalan	Twitter	8958	Hate, non-hate	[4]
	Albadi	Twitter	6136	Hierarchical labels: (1) neutral, religious hate (2) if religious hate: Muslims, Jews, Christians, Atheists, Sunnis, Shia, other	[2]
	Chowdhury	Twitter, Facebook, YouTube	4000	Hierarchical labels: (1) non-offensive, offensive (2) if offensive: vulgar, hate, only offensive	[7]
	Mubarak	Twitter	9996	Hierarchical labels: (1) non-offensive, offensive (2) if offensive: hate speech, not hate speech	[28]
	Mulki	Twitter	5846	Normal, abusive, hate	[29]

We train a classifier for each dataset and test it on the test sets of the other datasets. Firstly, we sample an equal number of documents from each dataset and split them into training and test set (80:20). Identical training and test set sizes are necessary to receive comparable results. Subsequently, we merge the classes so that we get a binary task (*neutral* and *abusive*). It is necessary because there is no standard labeling schema for abusive language. Most datasets can be converted to binary tasks. After preprocessing the documents, we train classification models for each dataset. For the classifier, we fine-tune a pre-trained BERT model depending on the language of the datasets for our task. After training the classifiers, we evaluate them on all test sets and a combined test set that consists of equal samples of documents from all tests. The results show how well a classifier trained on one dataset performs on unfamiliar datasets, demonstrating the generalizability of a classifier and its corresponding dataset.

### 3.4.2 Explainable classification models

The previous method provides a useful overview of the dataset's generalizability. We recommend a method to analyze the classifiers with an explainable AI technique to study the classifiers and uncover their blind spots or weak points.

Therefore, the models trained by the previous method are combined with the SHAP framework, which provides a set of different methods to explain predictions [22]. Concretely, we apply the Partition SHAP method—a model-agnostic local explainability method that relies on Owen values to explain single predictions [21].

Our method's outcome is the following: For a given document from the combined test set, we receive a prediction and

an explanation for each dataset. The explanation shows how each word contributes to the prediction of the classifier. So, we can compare the different classification models in-depth and identify weak points of the classifiers because we see what is relevant for the classifier and what is not. These insights also help to uncover bias. For example, a classification model classifies a nonabusive document as abusive, and the explanation marks the word *Islam* as highly relevant for the prediction. This can indicate a religious bias in the data, caused by the fact that the word *Islam* occurs more frequently in the abusive class than in the neutral class.

## 4 Data

Our developed framework is meant to be a tool for researchers and data scientists that work with abusive language datasets or create such datasets. In order to demonstrate its usage, we apply the framework to five English and six Arabic datasets listed in Table 2. We selected English because most abusive language resources are written in English, and Arabic because it fundamentally differs from English. In contrast to other dataset reviews, we compare only a small number of datasets due to our comprehensive, in-depth analyses; considering more datasets would go beyond the scope of this article.

Our dataset selection focuses on Twitter as the primary data source to ensure the comparability of the datasets. A further criterion is the size of the dataset. Since we draw even samples from all datasets for some analyses, the smallest dataset has 4,000 tweets.

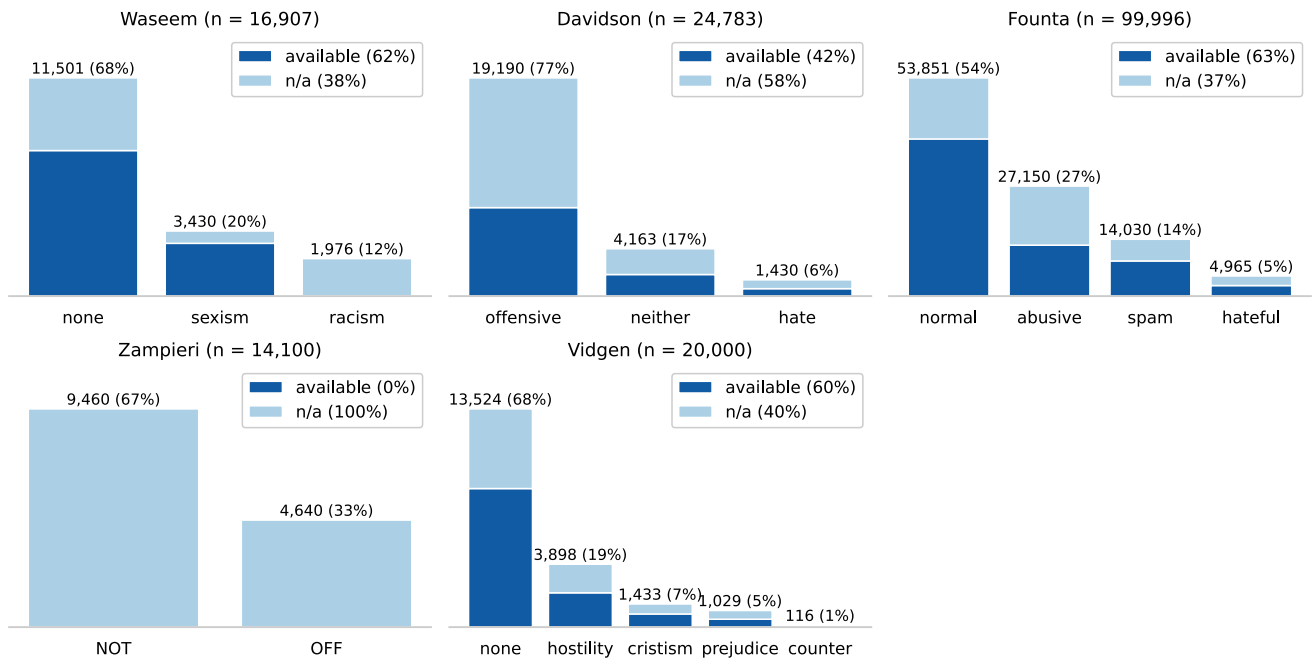


Fig. 2 Class distribution and platform availability of English datasets (*available* means that the online resource, e.g. tweet, is still accessible)

The first three English datasets are commonly used by the research community [32], Zampieri is from the shared task OffensEval 2019 [48], and Vidgen is a relatively new dataset about COVID-19-related abusive language [40]. The latter was selected because it comprises an entirely different context. Regarding the Arabic datasets, we picked the Twitter datasets that we found and are not too small. One of them is also from a shared task—Mubarak (OSACT4 Arabic Offensive Language Detection Shared Task) [28]. Finally, Chowdhury consists of tweets and comments from Facebook and YouTube, making it an interesting dataset to compare to the others [7].

Some proposed methods (e.g., classification) require a unified labeling schema. Therefore, we convert the labels to a binary labeling schema—*neutral* and *abusive*. The *abusive* class comprises all classes that refer to abusive, offensive, or hateful language. Moreover, the bold-faced classes in Table 2 are those that are labeled as *abusive*.

## 5 Case study 1- english datasets

### 5.1 Meta perspective

#### 5.1.1 Class distribution and availability

Figure 2 presents the class distributions and data available on the social media platform (Twitter) of the English datasets. The number next to the dataset name is the total number of documents in the dataset. The percentage value on top of

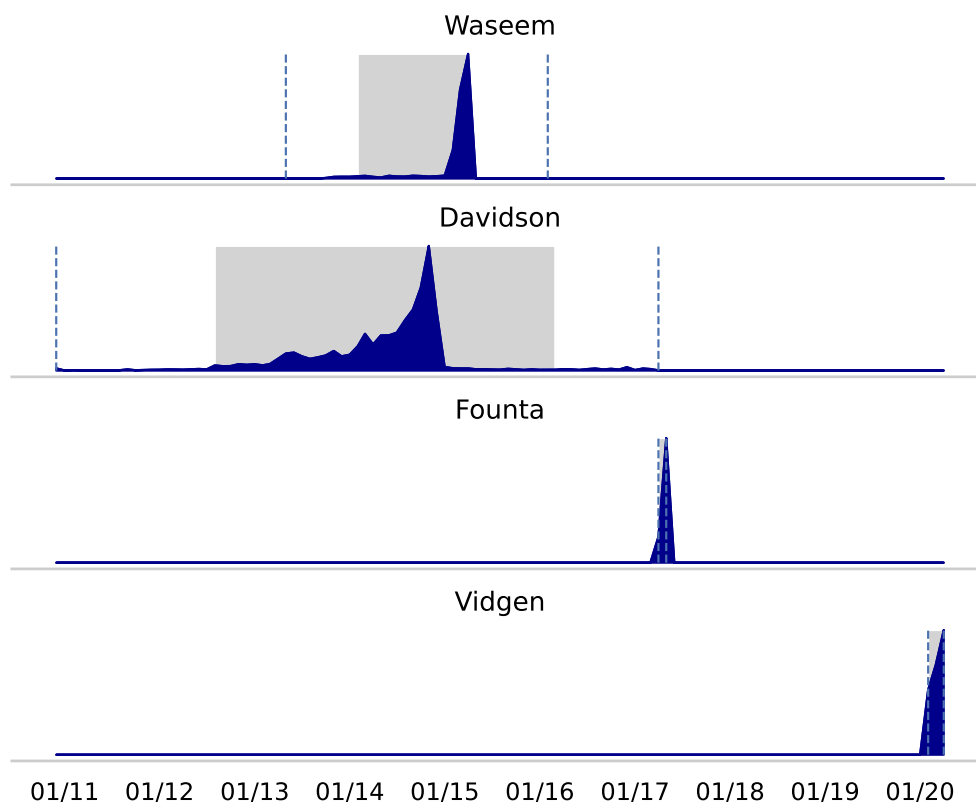
each bar and in the legend states the class’s relative share and reflects how much of the entire dataset is still accessible online, respectively.

The first observation is that all datasets are imbalanced. In all datasets, except Davidson, the abusive language-related classes are underrepresented. In the case of Davidson, the *offensive* class has a share of 77%, while *neither*’s share is 17%, and *hate*’s is only 6%. In regards to the data availability, we observe a degradation between 37% and 58%. It is not surprising that the hate-related classes (e.g., *racism* of Waseem, *offensive* of Davidson, *abusive* of Founta) are more affected by degradation because these tweets violate Twitter’s community guidelines. The 0% availability of the Zampieri is not representative because we do not know how many tweets are still accessible online due to the missing tweet IDs. This is also why we cannot perform the following two analysis methods on the Zampieri dataset.

#### 5.1.2 Temporal distribution

Figure 3 visualized the distribution when the tweets were posted. The dotted lines represent the timestamps of the first and last tweets, while the gray area marks the 95% percentile.

While all documents from Founta and Vidgen were created in a short period of time, the ones from Waseem and especially the ones from Davidson cover a more extended period. The latter is beneficial for training generalizable classifiers because it comprises linguistic traits from various periods—especially in the context of quickly evolving



**Fig. 3** Temporal distribution of the tweets from English datasets with tweet IDs

day-to-day languages. Further observations are that the datasets are from different years and that there are approximately five years between the oldest and newest. If the data is too old, it can have a negative impact because classifiers trained on this data struggle to identify recent abusive language expressions. Therefore, abusive language datasets should be up-to-date.

### 5.1.3 Author distribution

Figure 4 illustrates the Pareto analysis of the documents' authors. Due to degradation, the analysis only considers the tweets that are still available on Twitter<sup>2</sup>. We observe that the Waseem dataset has an obvious author bias because nearly all racist tweets were created by one author and a large portion of the sexism tweets by two authors. In contrast, the other datasets do not contain an imbalance with respect to the authors and their tweets.

## 5.2 Semantic perspective

### 5.2.1 LSI-based intra-dataset class similarity

Figure 5 displays the results of the LSI-based intra-dataset class similarity. The scores are between zero and one. The higher the score, the more homogeneous or similar the two classes.

The first observation is that the LSI scores of Zampieri are higher than those of the other datasets. That means that the classes are more homogeneous than those from the other datasets, but both classes are also similar. A contrast to these two classes is the racism and sexism classes of Waseem. They are more homogeneous by themselves than they are similar to each other. Concerning Founta, we see that the spam and normal class are very similar, while the abusive one is distinguishable from these two classes. The hateful class is less homogeneous than the other three and is also similar to the other three. Finally, Vidgen exhibits constant LSI scores both within and between the classes, suggesting a balanced dataset composition.

<sup>2</sup> For Waseem, we use the dataset provided by [27] because it contains all tweets author name.

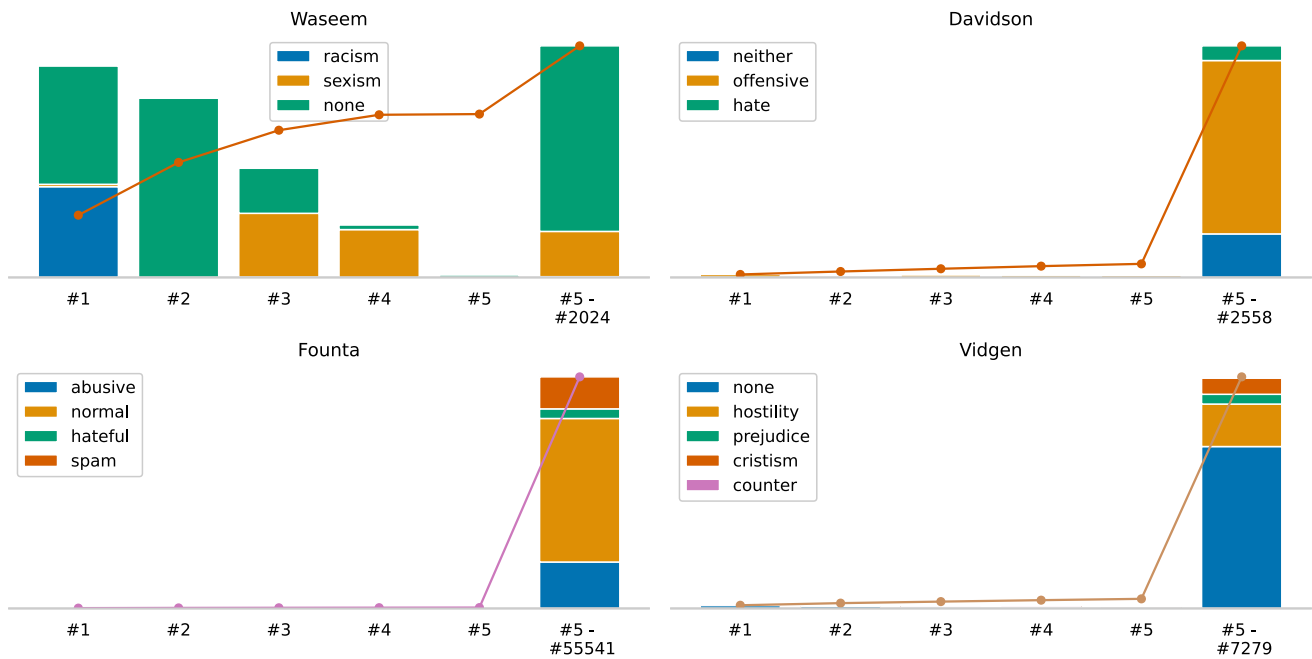


Fig. 4 Pareto analysis showing how many tweets (incl. classes) were created by the top authors of each dataset

### 5.2.2 Document embedding based intra- and inter-dataset class similarity

Figure 6 visualizes the similarities between the classes of all datasets based on the averaged FASTTEXT document vectors and PCA. We can observe that each dataset’s classes are approximately grouped, signifying coherence within the dataset. One outlier is the *spam* class of Founta. It was a good decision from the authors of the Founta dataset to introduce a spam class. Otherwise, the documents would have fallen in the normal class, making it easier for classifiers to distinguish between abusive and normal content without actually learning the differences between these classes. Furthermore, the racism class of Waseem and the prejudice from Vidgen seem to be quite similar. As racism often contains prejudice, this similarity should not surprise us. Besides that, Vidgen’s other classes are separated from the rest, which can be traced back to the topical focus of the dataset. Additionally, we can see that some hate-related classes (*sexism* of Waseem, *hate* and *offensive* of Davidson, and *hateful* and *hateful* of Founta) exhibit a certain degree of similarity. We can observe this grouping effect also at the neutral classes of Vidgen, Founta, and Davidson.

### 5.2.3 PMI-based word ranking for classes

Table 3 presents the words with the highest PMI from the abusive classes, demonstrating what the classes represent. It is not surprising that the abusive classes of Davidson and Founta contain many swearwords. Furthermore, we can see

that the *racism* class of Waseem focuses on religious topics, especially Islam. Another interesting observation is the dominance of political terms in the Zampieri. Similar to Zampieri, the hostility class of Vidgen steps out the lines. The most relevant phrases are related to viruses and China, which we can trace back to the dataset’s topical focus. But the missing offensive words indicate that the hate within the Vidgen dataset might be more implicit.

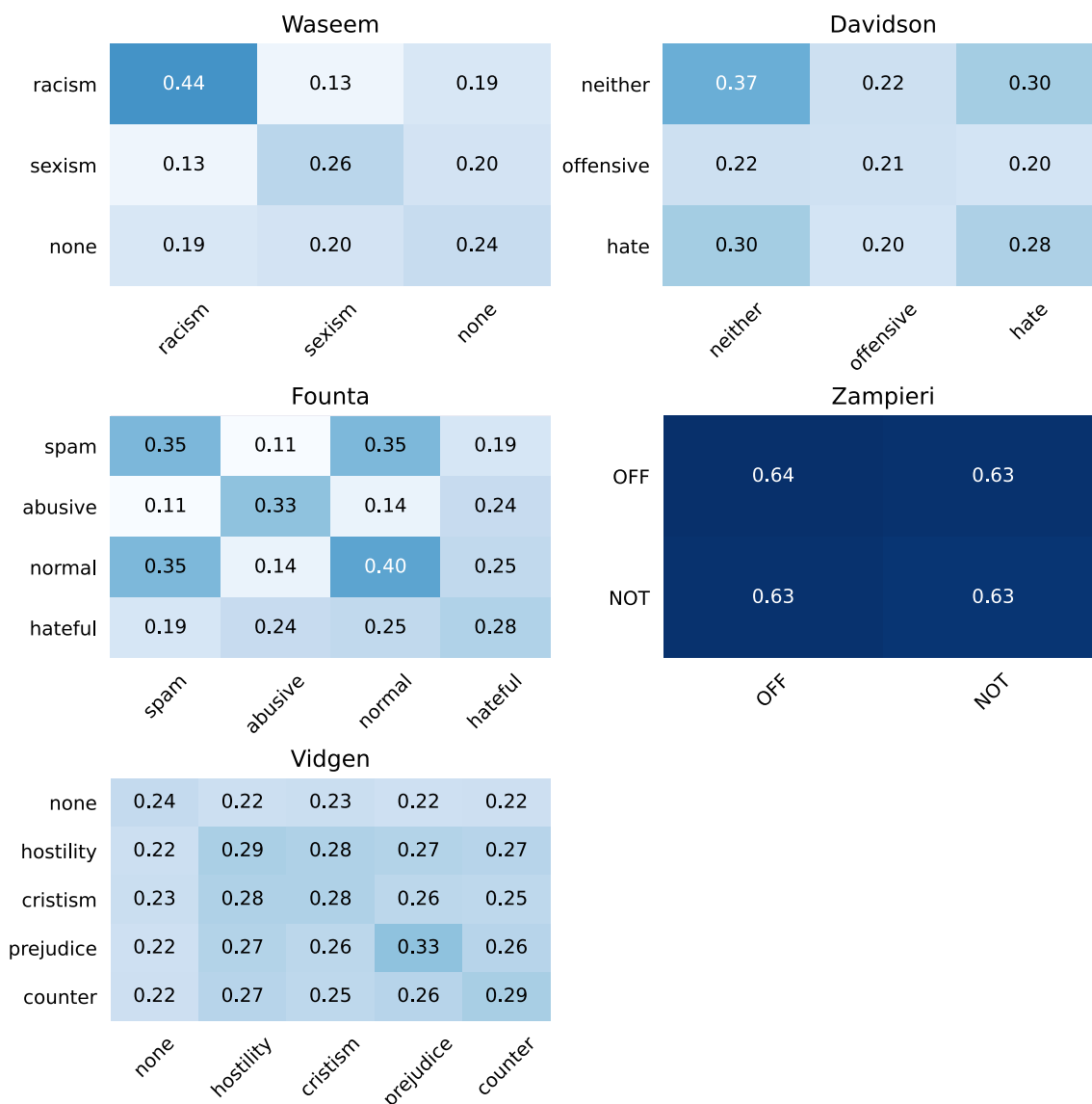
### 5.2.4 Overarching topic modeling

Figure 7 shows the result of the topic model-based analysis on all classes. The black dots represent the centroids of the 20 identified topics. We can observe different topic biases of the datasets: A large portion of Vidgen is about viruses and China, which is not surprising due to the focus on COVID-19 (T17). Many tweets from Waseem deal with Islam. Zampieri exhibits a political focus (T3, T5: e.g., liberals, democrats, conservatives). In contrast, Davidson and Founta contain several tweets with swearwords (T2, T4, T19: e.g., bitch, asshole, nigga). These findings are in accord with the ones from the previous analysis.

## 5.3 Annotation perspective

### 5.3.1 Distribution of inter rater reliability

Unfortunately, only one dataset provides the raw annotations that are necessary for conducting this analysis. Figure 8 displays the distribution of the inter-rater reliability



**Fig. 5** LSI-based similarity of classes within English datasets (the higher the score, the more similar are the two classes)

(Krippendorff's alpha) of each annotator from the Vidgen dataset, sorted from highest to lowest. The horizontal line shows the overall inter-rater reliability of all annotators, where the first observation introduces the overall Krippendorff's alpha value as 0.543. It is not an optimal value, but it is comparable to other abusive language datasets [20]. 10 of the 26 annotators achieve an individual inter-rater reliability score over 0.80 between the annotators and the dataset gold standard, which is relatively good. The outlier is the last annotator with an inter-rater reliability score of 0.564. Since at least two coders annotated each document of Vidgen, one outlier cannot cause an annotator bias. Overall, the annotations of Vidgen seem to have decent quality. Based on the results of this analysis, we are not able to identify any annotator bias.

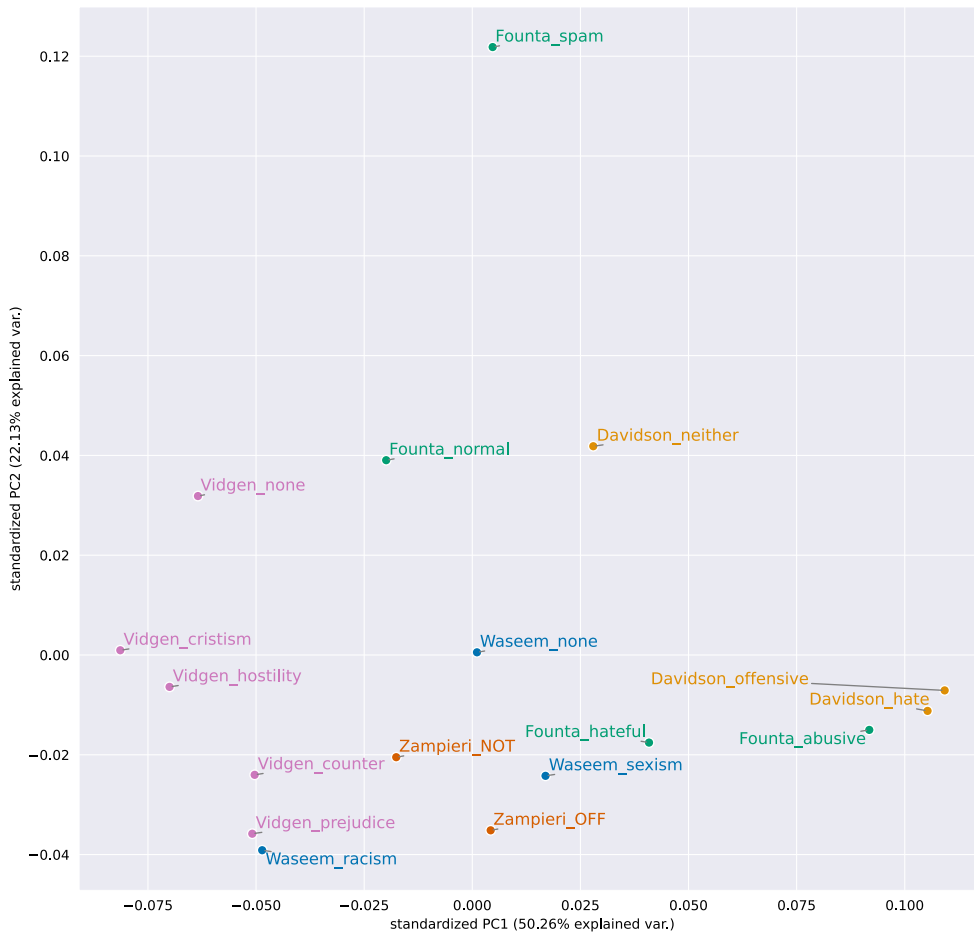
## 5.4 Classification perspective

### 5.4.1 Cross-dataset performance

Figure 9 presents the macro F1 scores of the classifiers that were trained on different datasets and tested on all test sets. Hate labels were unified on each dataset for the purpose of cross-classification.<sup>3</sup> As the basis for the classification model, we use the English pre-trained BERT model `bert-base-uncased` [12].

<sup>3</sup> Labels in bold in Table 2 are assigned to *abusive*, the others to *neutral*.





**Fig. 6** FASTTEXT sentence embedding vectors averaged for each class of English datasets and visualized with PCA (the closer the points, the more similar the classes)

**Table 3** Words with highest PMI for each class of the selected abusive English datasets

	Words with highest PMI
Waseem - sexism	sexist, women, kat, girls, like, call, female, men, think, woman
Waseem - racism	islam, muslims, muslim, mohammed, religion, jews, prophet, isis, quran, like
Davidson - hate	bitch, faggot, like, ass, nigga, white, fuck, nigger, trash, fucking
Davidson - offensive	bitch, bitches, hoes, like, pussy, hoe, ass, got, fuck, get
Founta - abusive	fucking, fucked, like, ass, bitch, fuck, get, bad, shit, know
Founta - hateful	hate, niggas, fucking, nigga, like, people, idiot, get, amp, ass
Zampieri - OFF	liberals, like, control, gun, people, shit, antifa, get, conservatives, one
Vidgen - hostility	china, world, chinese, virus, people, ccp, us, wuhan, spread, rt

Vidgen delivers the worst performance, but this should not be surprising due to the topic focus. Davidson, Founta, and Zampieri show comparable results that are better than the ones from Waseem. Even if Davidson has the highest F1 score on the combined test set, the classifiers trained on Founta and Zampieri provide more stable results across all test sets. Therefore, these two datasets are most suitable for training generalizable classifiers.

### 5.4.2 Explainable classification models

Figure 10 shows the SHAP explanations for the classification of a selected tweet for each classifier. The numbers in bold represent how likely the document is classified as *abusive*. The words in red contribute to the classification as abusive, while the blue ones support the classification as *neutral*. We observe that classifiers trained on the Founta

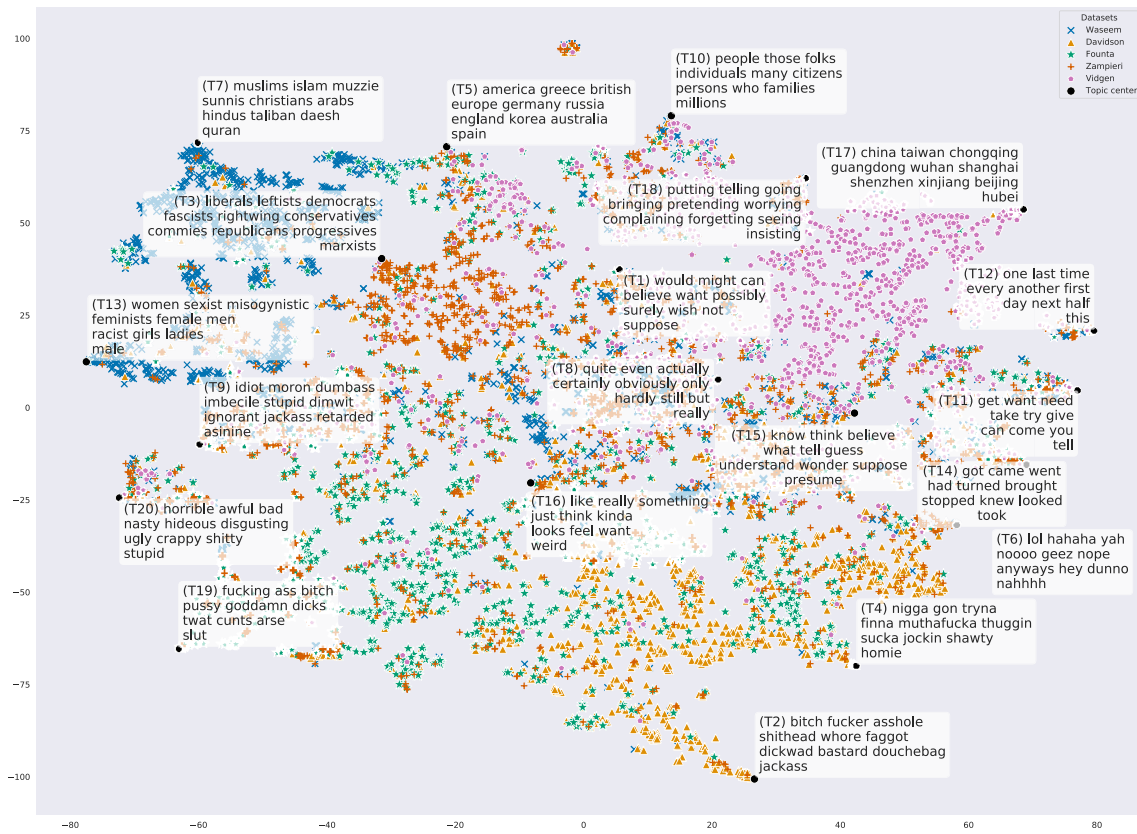


Fig. 7 Topic model on the abusive classes of English dataset selection

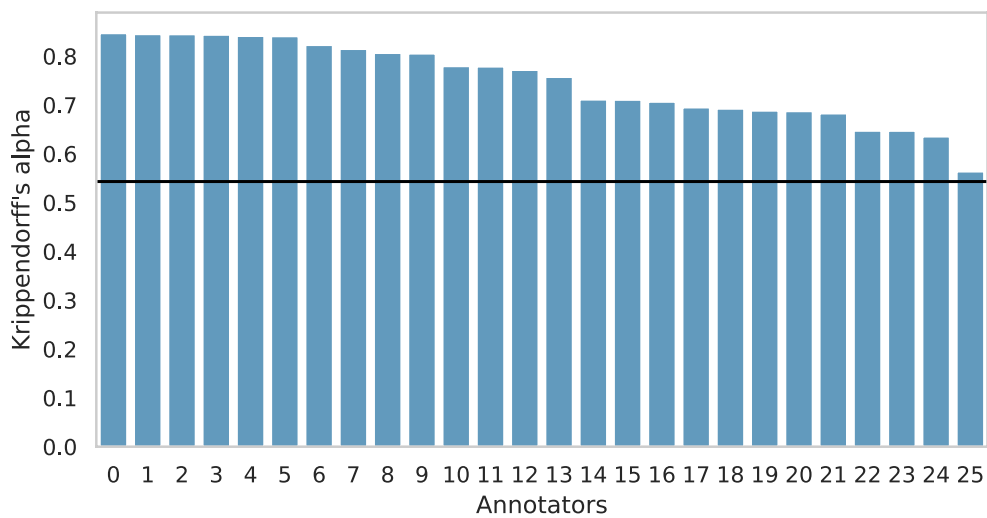


Fig. 8 Annotators' inter-rater reliability scores and overall inter-rater reliability score (black line) of Vidgen dataset

Classifiers	Test sets					Combined test set
	Waseem	Davidson	Founta	Zampieri	Vidgen	
Waseem	81.1%	65.9%	58.4%	58.3%	55.3%	70.5%
Davidson	56.3%	90.9%	89.1%	65.2%	52.8%	79.3%
Founta	65.2%	73.1%	93.1%	74.0%	57.6%	79.9%
Zampieri	61.4%	75.6%	91.3%	77.0%	61.2%	79.2%
Vidgen	45.2%	14.7%	41.3%	41.3%	80.7%	45.3%

Fig. 9 Cross-dataset classification performance (macro F1 scores)

and Vidgen datasets misclassified the *abusive* tweet, while the other three correctly classified it with high confidence. In the case of Vidgen, the result should not be surprising because the dataset focuses on COVID-19-related topics and not on sexism. In contrast to that, it is unexpected that Founta seems to have a blind spot on sexism because it appears to be diverse.

## 6 Case study 2- Arabic datasets

### 6.1 Meta perspective

#### 6.1.1 Class distribution and availability

Figure 11 shows the class distributions and data availability of the Arabic datasets on the social media platforms. Similar to the English datasets, all datasets, except Mulki, are imbalanced and dominated by the neutral class. Overall, the dataset sizes are of the same magnitude and range between 4,000 and 9,996 documents. The dataset classes are more coherent than the ones from the English datasets. Regarding data availability, we can only analyze three of the six datasets because only those contain tweet IDs. We can observe a similar data degradation as for the English datasets. All classes are affected, but mainly the abusive classes. The overall range of degradation is between 34% and 42%. In the case of Albadi, we received the full dataset from the authors, which is employed for the rest of the case study. The other three datasets (Chowdhury, Mubarak, and Mulki) provide only the full text but no reference to the online resource. Therefore, we cannot consider them in the following two analysis methods.



Fig. 10 SHAP explanations of an abusive tweet that is misclassified by two of the five English classification models

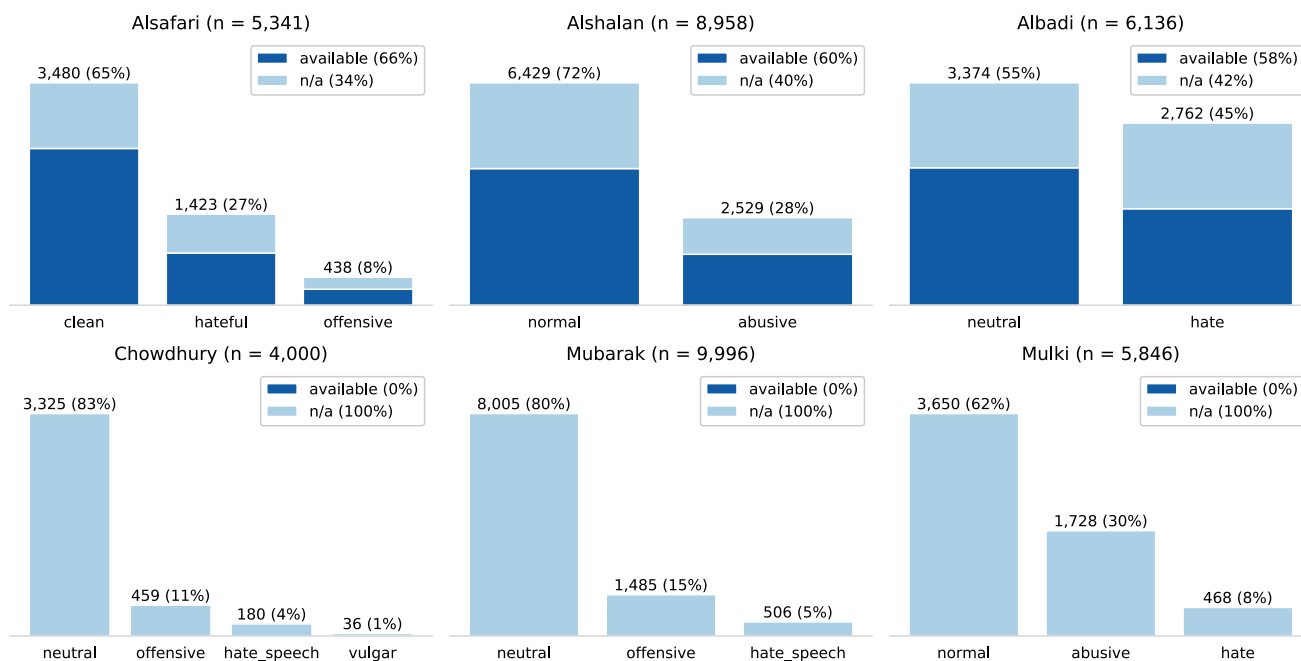
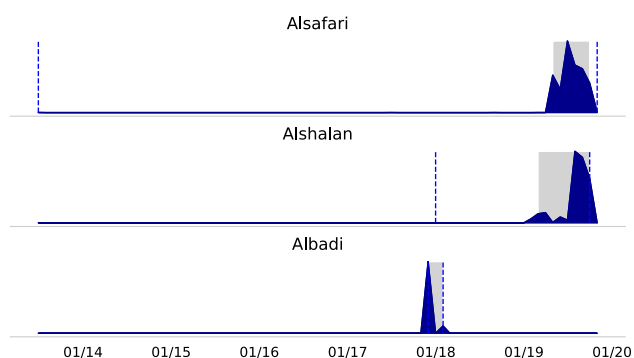


Fig. 11 Class distribution and platform availability of Arabic datasets (*available* means that the online resource, e.g. tweet, is still accessible)



### 6.1.2 Temporal distribution

Figure 12 visualized the distribution when users posted the tweets. Overall, each dataset’s largest portions cover a short period similar to most of the English datasets. Furthermore, 95% percentile from Alsafari and Alshalan mainly come from the same period. In the context of the data degradation findings, it is surprising that the degradation rate of Albadi, which is approximately two years older than the other two, is only 2 and 8 pp higher.

Fig. 12 Temporal distribution of the tweets from Arabic datasets with tweet IDs

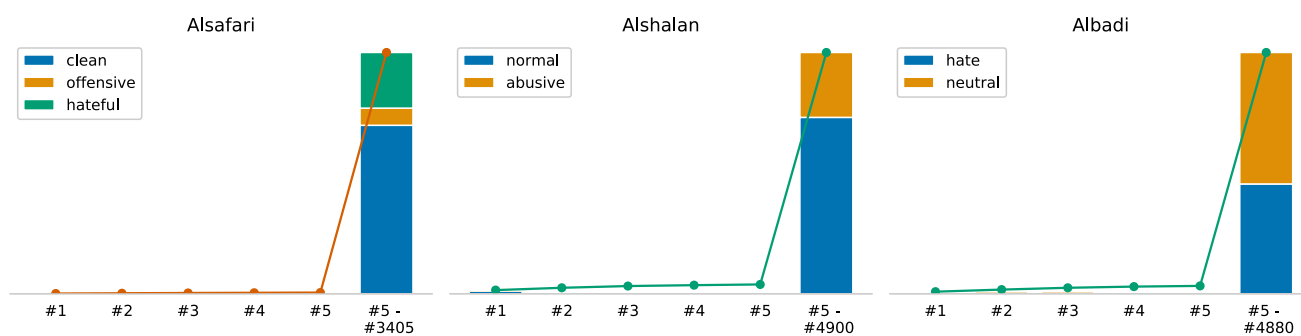


Fig. 13 Pareto analysis showing how many tweets (incl. classes) from Arabic datasets were created by the top authors of each dataset

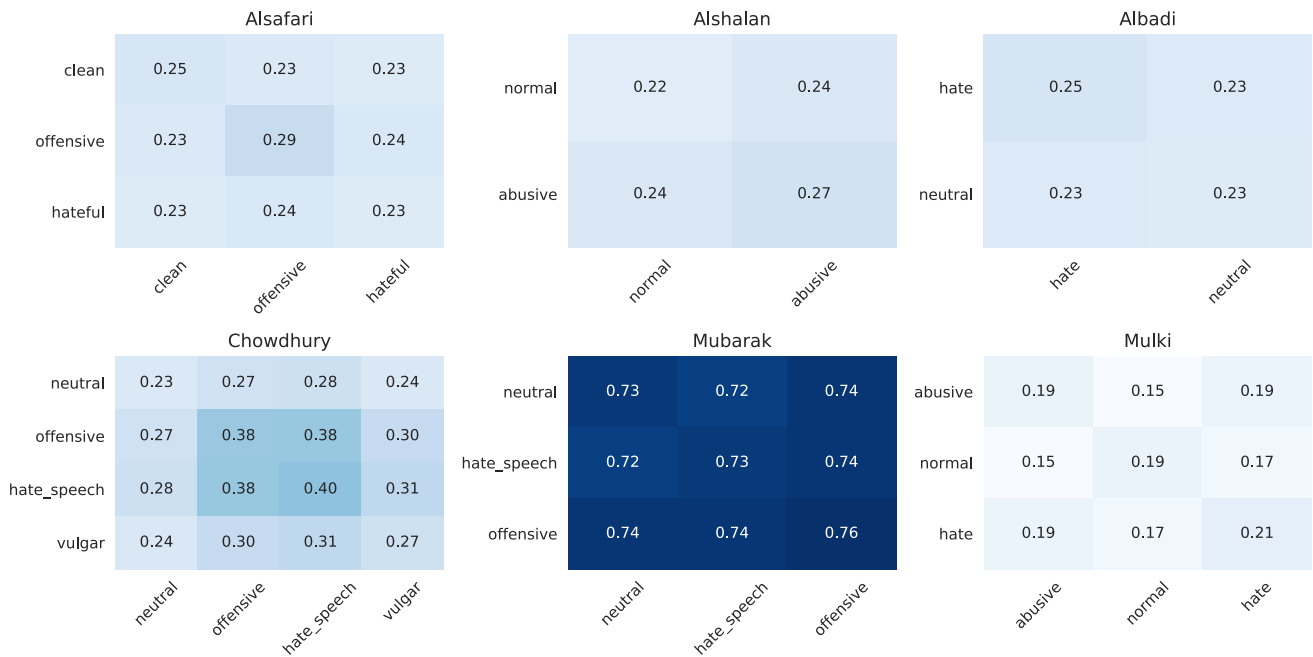


Fig. 14 LSI-based similarity of classes within Arabic datasets (the higher the score, the more similar are two classes)

### 6.1.3 Author distribution

Figure 13 shows the Pareto analysis on the authors of the Arabic datasets that contain tweet IDs. In the case of Albadi, we received the original dataset. Thus the chart includes all authors. In contrast, the charts from Alsafari and Alshalan contain only the author data from 66%, respectively, 60% of the tweets. Overall, none of the datasets have a small group of authors that created a larger portion of the tweets, resulting in no author bias.

## 6.2 Semantic perspective

### 6.2.1 LSI-based intra-dataset class similarity

Figure 14 presents the results of the LSI-based intra-dataset class similarity of the Arabic datasets. The first observation is that the LSI scores of Mubarak are higher than those of the other datasets. That means that the classes are more homogeneous by themselves than those from the other datasets. But the intra-class similarity of all classes is also in the same range. Furthermore, we can observe that the Albadi dataset is similarly homogenous. Alternatively, the *offensive* class of Alsafari and the *offensive* and *hate speech* class of Chowdhury stand out. All three classes are more homogeneous than the other classes. In the case of Chowdhury, both classes are also quite similar compared with the other two dataset classes. Based on Mulki, we can observe that the

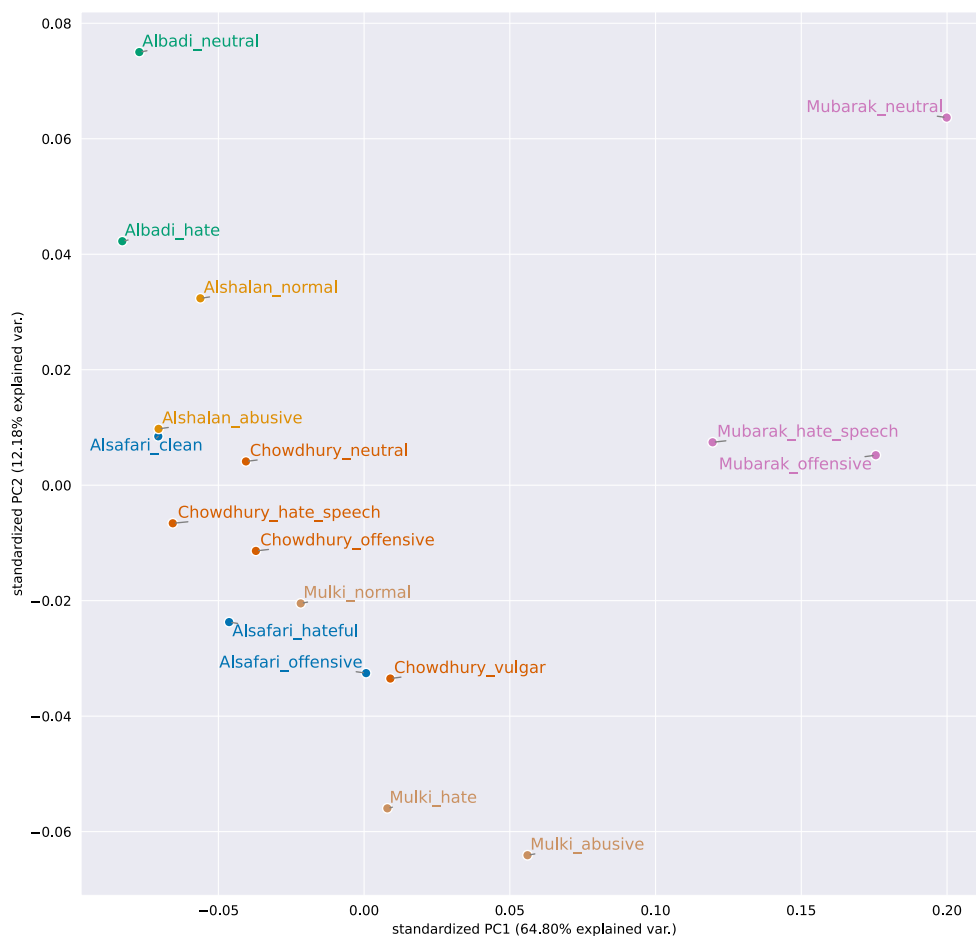
*normal* class distinguish from the *abusive* and *hate* class, while these two are similar.

### 6.2.2 Word embedding based inter- and intra-dataset class similarity

For calculating the inter- and intra-dataset class similarities, we used the Arabic FASTTEXT word embeddings. Figure 15 visualizes the results. The first observation is that the abusive classes of a dataset are closer to each other than to the neutral class, which should not be surprising. But there is one exception—the *vulgar* class from Chowdhury. The Mubarak dataset is an outlier in this analysis because all its classes strongly differentiate from all others.

### 6.2.3 Most relevant terms of abusive classes

In Table 4, we report the words with the highest PMI for each class in each dataset. High PMI words in hateful classes differ for each dataset: while those words in Albadi’s hate class are just religious names, in Chowdhury, they are country names, and in Mulki, they are related to Lebanese politics. The same observation can be seen in the offensive and abusive class of each dataset. In Chowdhury, the highest PMI words in the offensive class are political, while in Mubarak, they are related to sports. The highest PMI words in the abusive class of the Alshalan dataset are not abusive, while those in Mulki’s abusive



**Fig. 15** FASTTEXT sentence embedding vectors averaged for each class of Arabic datasets and visualized with PCA (the closer the points, the more similar the classes are)

class are abusive and are also specific to the Levantine dialect.

#### 6.2.4 Overarching topic modeling

Figure 16 exhibits the topic model-based analysis results on all classes. We can observe varying topic biases from the different datasets. For example, topics identified in Albadi have a religious aspect, which should not be surprising because the dataset focuses on religious hate. One of these topics is about religions, as its words contain religious names (e.g., T15 contains words such as Jews, Muslims, Christians, and Secularism). Another topic in Albadi is about different Islamic Sects (e.g., T6 contains words like Sunnis, Shia, and Salafis), and another one is about religious ideologies and doctrines (T14). Albadi seems to be the most separable dataset in terms of topics, as most of its data points fall near religious topics. Mulki and Mubarak share many topics, specifically those related to different Arabic dialects like Egyptian (T13), Levantine (T15), and standard Arabic

(T10). In addition, Alsafari exhibits topics related to people from different Arabian nationalities (T1, contains words like Egyptians, Palestinians, Saudis, Lebanese) and topics related to females (T10), which is also apparent in Alshalan. Another identified topic in Alshalan is political words (e.g., T7 is reflected by words like democratic, society, union, local, and organizations).

#### 6.3 Annotation perspective

Unfortunately, none of the authors released the raw annotation data. Thus we are not able to conduct this analysis for the Arabic datasets.

#### 6.4 Classification perspective

##### 6.4.1 Cross-dataset performance

Figure 17 presents the macro F1 scores of the classifiers that were trained on different datasets and tested on all test

**Table 4** Words with highest PMI for each class of the selected abusive Arabic datasets

Dataset	Words with highest PMI	Translation
Alsafari - hateful	الله، شعب، العرب، الذكور، والله، الشيعة، البنات، اليهود، الحريم، الشواما	God, People, Arabs, Males, Shia, Girls, Jews, Women, Shamis
Alsafari - offensive	الله، الرجال، منك، انك، النساء، والله، وجهك، العرب، كلب، مثل	God, Men, From you, That you, Women, Your face, Arabs, Dog, Like
Alshalan - abusive	الله، السعودية، نيزك، النسويات، تركيا، والله، مثل، هذه، المجتمع، قطع	God, Saudi Arabia, Meteor, Feminine, Turkey, Like, This, Society, Cut
Albadi - hate	الله، اليهود، الشيعة، الاسلام، اهل، المسلمين، السنة، ايران، الاتحاد، الملحدون	God, Jews, Shia, Islam, People, Muslims, Sunni, Iran, Atheism, Athiests
Chowdhury - offensive	الله، قناة، والله، الجزيرة، مصر، الجزيرة، قطر، العالم، العرب، هذه	God, Channel, Aljazeera, Egypt, Pigs, Qatar, World, Arabs, This
Chowdhury - hate_speech	الله، قطر، العرب، ايران، السعودية، قناة، الاخوان، والله، المسلمين، العربية	God, Qatar, Iran, Saudi Arabia, Channel, Muslim Brotherhood, Muslims, Arabia
Chowdhury - vulgar	الله، فيه، زق، واحد، اساس، بن، سلمان، قطر، محمد، ابن	God, In, One, Base, Son, Salman, Qatar, Mohammed
Mubarak - offensive	الله، ابن، والله، كلب، دي، الكلب، ولاد، ايه، ابو، منك	God, Son, Dog, Sons, Father, From you
Mubarak - hate_speech	الله، ابن، ابو، كلب، الاهلي، والله، عبيد، ال، الاتحاد، عبد	God, Son, Father, Dog, Ahli, Slaves, Itihad, Slave
Mulki - hate	كلب، كلاب، الله، لبنان، باسيل، جبران، مثل، قطر، انتو، حزب	Dogs, Dog, God , Lebanon, Baseel, Jebran, Like, Qatar, You, Party
Mulki - abusive	كول، هوا، باسيل، جبران، كلب، الله، عم، حمار، خراس، روح	Eat, Air, Baseel, Jebran, Dog, God, Donkey, Shut up, Go

sets. As the basis for the classification model, we use the Arabic pre-trained BERT model *asafaya/bert-base-arabic* [34].

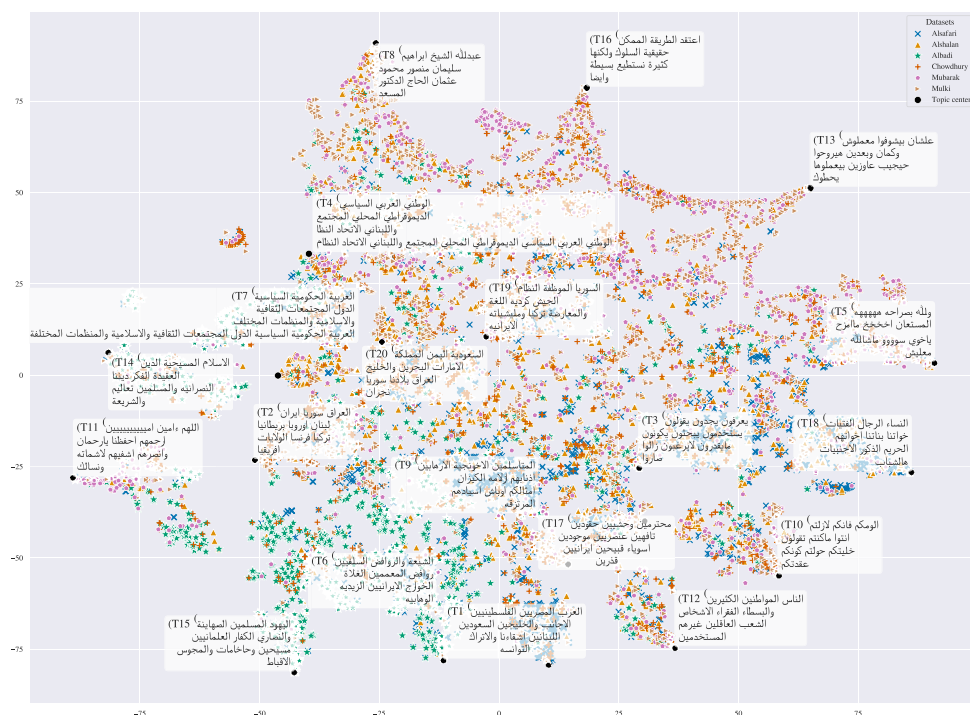
The model that performs best on its own test set is trained on the Alsafari training set. Its performance on the combined test set is slightly worse than the top classification model (Mulki). Consequently, these datasets are more suitable to train generalizable classification models. Interestingly, all classification models except the one trained on Albaldi struggle on the Albadi test set, while the Albadi classifier still provides a comparable F1 score on the combined test set. Overall, the F1 scores on the combined test set are less volatile than those from the English datasets. An explanation can be that the labeling tasks are

similar, and there are no particular focuses on topics (e.g., Vidgen focuses on COVID-19-related tweets). Even if the F1 scores are lower on average than the ones from the English datasets, it is impossible to derive any conclusion from that because we use a different pre-trained model and a different number of training data.

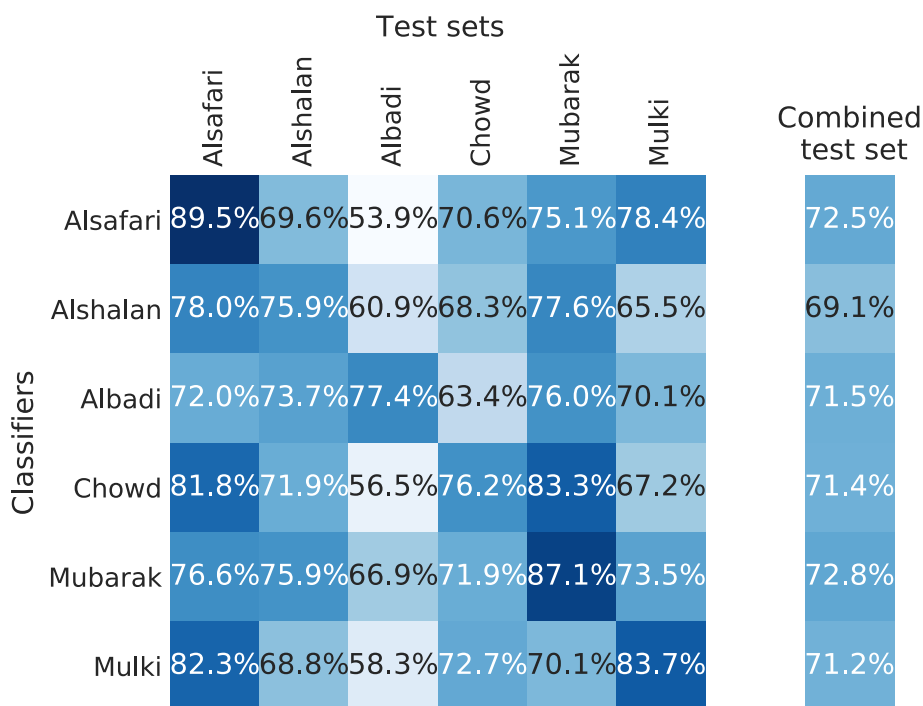
#### 6.4.2 Explainable classification models

Figure 18 shows the SHAP explanations for the classification of a selected tweet for each classifier. The numbers in bold represent how likely the document is classified as *abusive*. The words in red contribute to the classification

**Fig. 16** Topic model on tweets from abusive classes of Arabic datasets



**Fig. 17** Cross-dataset classification performance (macro F1 scores) of Arabic datasets



as abusive, while the blue ones support the classification as *neutral*.

The tweet shown in the figure translates to: “All Moroccans are cuckolds, and God is my witness.” This tweet is misclassified by both Alsafari’s and Alshalan’s classifier but due to different reasons. The figure shows that Alsafari’s classifier does not correlate the word “cuckold” with the

abusive class, while Alshalan’s classifier does but the presence of the word “witness” (which has the same writing as the word “martyr” in Arabic) plays a significant role toward classifying the tweet wrongly. Other classifiers classify the tweet correctly, and for all of them, the word “cuckold” is the one that plays the most prominent role.



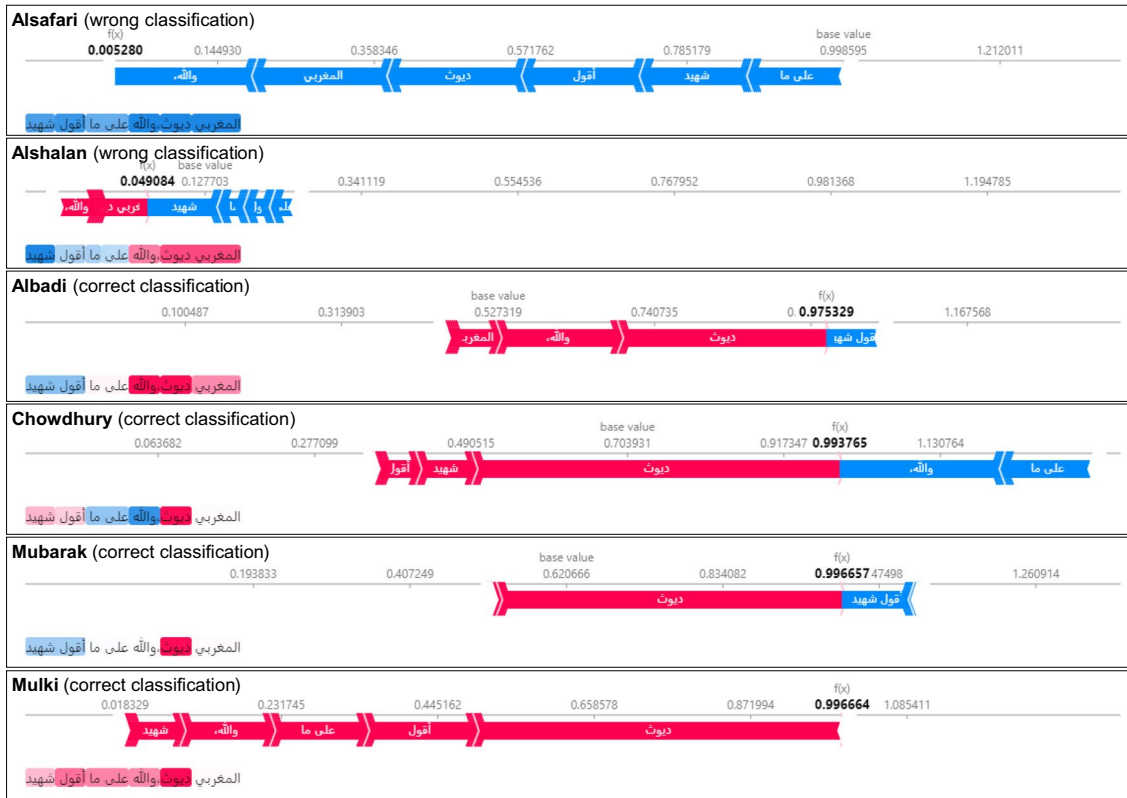


Fig. 18 SHAP explanations of an abusive tweet that is misclassified by two of the six Arabic classification models

## 7 Discussion

The results show significant differences between the compared abusive language datasets, and we identified different bias types. In both case studies, no dataset is free of common problems, and none stands out particularly positively in this regard. Each dataset comes along with advantages and disadvantages. Regarding the English datasets, Founta, for example, is a comprehensive dataset with good diversity, but it covers only a short period. Vidgen has a topic bias due to its focus on COVID-19. But such focused datasets are necessary because they address current trends. Concerning the Arabic datasets, we observed that these datasets are, on average smaller than the English ones—which could be due to resource limitations. Overall, the applied labeling schemata are more coherent than one of the English datasets. Similar to English, the datasets exhibit topical focuses (e.g., religious and political conflicts in the Middle East). Two datasets that look promising are Alsafari and Mubarak. The datasets have a decent size, the classes of both seem to be homogeneous, and they use similar labeling schemata, making them compatible. However, this has to be addressed in future work.

One may criticize that the datasets partially differentiate in the task/ labeling schema, but they all contribute to the

overarching goal of fighting against abusive online language, and some are often used together in papers to evaluate classifiers. Most relevantly, all datasets are listed as abusive language classification datasets and nominally used for the same type of task.

Previous approaches to comparing abusive language data, as discussed in the Related Work section, predominantly rely on surface-level descriptive features to distinguish individual datasets. One of the main propositions of this article is that this falls short of describing the real differences of these data sets, which vary much more than can be described by these surface-level features. In light of the common challenges abusive language detection systems face, systematic bias in training data is often at the core of these issues and very hard to detect or even measure. Therefore, part of the framework’s main contribution in a structured way is to make differences in data visible on a systematic basis that goes beyond descriptive attributes and basic statistics.

In this way, the proposed framework can analyze existing datasets and relate them to other well-known datasets from the field. This is relevant for the specific analysis shown here and with other datasets used in a similar context. The focus here is not on ranking the viability of a specific dataset but on providing context for comparisons. In general, there

are theoretically some aspects of the data that should be an indicator for better generalizability, such as balanced authorship or timescale and a variety of topics in both abusive and nonabusive content. However, they don't necessarily guarantee better data distribution alone.

As this study presents a tool to evaluate and compare abusive language datasets, we are also aware of the problematic ethical circumstances of the field. The framework proposed here stems partially from the need to analyze these datasets further to discover potentially hidden problems and biases. In this role, our framework understands itself as a tool to help researchers verify the integrity of and problems present in their data and help discover potential issues with newly created datasets by comparing them to already existing data.

The system tries to be conscious of various potential biases, but it cannot claim to cover every potential bias or guarantee ethical conduct in parts of the data collection and annotation process. Therefore, researchers need to see the results in conjunction with similar methods and their efforts to guarantee data integrity. In the worst case, the framework can fail to detect bias that is not covered in the evaluation metrics. For this reason, the tool is not presented as a check to guarantee that the data is bias-free but to offer a systematic way to uncover some of the most common problems present.

Understanding the underlying data that goes into potential classification systems is an essential part of systems development. In the past, abusive language detection has shown to be easily biased against minority groups, even though this technology is meant to protect them. Therefore, we see the necessity of being conscious about the different kinds of bias within data and having a framework for analyzing and comparing them. The proposed framework's metrics are predominantly based on either explicit metadata or evaluated in relation to the content of other datasets. This is a conscious choice to detect differences in a more fluid framework that does not rely on explicit prescriptions toward the data but should emerge from it.

Since biases can be exhibited in different ways, the methods in the framework all target various sources of conflict with the data. One of the main takeaways of the comparisons is how easy it is to separate different datasets for abusive language by some other attribute and see the dataset distinction reappear. Furthermore, these similarity measures often show more considerable variations between two different datasets than between the positive and negative classes within one of these datasets, making it clear how classifiers trained on such data might have difficulty generalizing between them.

Additionally, some of the proposed checks make it very clear what potential sources of bias might be predominant in the dataset and where to look for when assessing problems. It's also possible to find matching datasets to "patch-up" weaknesses in a known existing one. Overall, the most

important contribution of the framework lies in making potential blind spots, tendencies, and differences visible to engage with them critically in building systems.

We propose that researchers and data scientists who use different datasets to build abusive language classifiers should be aware of the datasets' differences and biases and consider these findings during the analysis of results. So, they reduce the risk of unintended and unfair behavior of their models. Moreover, there should be increased awareness of the types of issues present in models and an incentive for dataset creators to already these parameters in mind when designing the data collection process.

Apart from general data collection issues, a few other key issues are often observed with abusive language data and become apparent when working with the datasets and using the framework.

A severe issue that we observed during our analysis is the dataset degradation that has been already mentioned by other researchers [41]. The problem is related to the procedure that some researchers publish only references/links (e.g., tweet IDs) instead of the actual text. Over time, fewer documents are available because some of them are deleted. Consequently, abusive language is deleted over time, which is good and reflects moderation efforts by social media sites. However, it impairs the reproducibility of research, reusability of data, and advances in abusive language research. This procedure also has an advantage: it preserves the user's right to delete data. Nevertheless, we argue with respect to the degradation rate that researchers should release the text so that the datasets are persistently available. In order to address the mentioned conflict of interest between user privacy and research, we suggest anonymizing the tweets, meaning anonymous identifiers should replace the author and all usernames appearing in the text. Thus, user privacy can be preserved, and researchers can still apply our proposed framework. Furthermore, we assume that most of the tweets that are no longer available were deleted by Twitter due to policy violations. Therefore, our proposed approach should be a suitable solution.

Similarly, there is still a general problem with the ill-defined term of abusive language in general. Most often, the definition of what is considered abusive is up to the dataset creators in their labeling process. However, sometimes the criteria for the labeling is also not immediately apparent, either when done by experts or in crowdsourcing procedures. Abusive language definitions applied when labeling a dataset are therefore neither in accordance with a legal definition such as hate speech nor necessarily with the policies employed by the social media site's data. Conversely, this is advantageous since it doesn't tie the definition of abuse to the legal framework of one specific country or the policies of whatever social media site the content is located. However, it leads to potentially ill-defined cases where either too much,

too little, or entirely different kinds of content get labeled as abusive. Therefore, it would be valuable to create very explicit datasets about the definition employed, going so far as giving examples of edge cases and how a decision was being made. In some cases, it can also make sense to define the type of task more precisely than just generalized abuse if the data is very focused on the type of content it collects.

Keeping these considerations in mind, researchers and data scientists that release new abusive language datasets should consider the following guidelines:

- They should apply our framework or a comparable one on their dataset to compare with abusive language datasets.
- The dataset should contain the full text, the mentioned metadata, and the reference to the original resource (e.g., tweet ID). To protect users' privacy, they can anonymize the usernames in the metadata (e.g., hashing) and remove them from the full text.
- Besides an aggregated version of the annotations, they should include all raw annotations for each annotator—in the best case with metadata about the annotators.

The latter refers to annotators bias, a form of bias that we could not investigate due to missing data. Investigations into annotator bias require a great deal of transparency from the creator of a dataset, ideally encompassing descriptions of each annotator, their backgrounds, and potential biases, as well as a detailed overview of which annotator assigned what label to a particular data instance. Releasing this type of data would enable further insights into how different people annotate the same type of content and by which annotators influence some examples were labeled one way or the other. There is already research that investigates annotator bias in abusive language but requires additional data [1, 6, 45].

## 8 Conclusions

This paper has presented an overview of a framework to describe and compare datasets from the abusive language detection domain to highlight potential problems, biases, and differences better. Therefore, we propose a multiapproach framework to investigate different aspects of the data and make them comparable beyond a discreet description framework based on labels, which has been used predominantly in the past. Our paper contributes toward helping researchers and data scientists to improve data quality and enhance their systems when collecting new data as well as when working with existing data. While our proposal was focused on the domain of abusive language detection, the proposed

framework would also apply to similar NLP tasks relying on labeled data.

**Funding** Open Access funding enabled and organized by Projekt DEAL. This research has been partially funded by a scholarship from the Hanns Seidel Foundation financed by the German Federal Ministry of Education and Research.

**Data availability** We used only datasets that a publicly available.

## Declarations

**Conflict of interest** The authors declare that there is no conflict of interest.

**Code availability** The code is published on <https://github.com/mawic/abusive-language-dataset-framework>.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Al Kuwatly, H., Wich, M., Groh, G.: Identifying and measuring annotator bias based on annotators' demographic characteristics. In: Proceedings of the Fourth Workshop on Online Abuse and Harms, pp. 184–190. Association for Computational Linguistics, Online (2020). <https://doi.org/10.18653/v1/2020.alw-1.21>. <https://www.aclweb.org/anthology/2020.alw-1.21>
2. Albadi, N., Kurdi, M., Mishra, S.: Are they our brothers? analysis and detection of religious hate speech in the Arabic Twittersphere. Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018 pp. 69–76 (2018). <https://doi.org/10.1109/ASONAM.2018.8508247>
3. Alsafari, S., Sadaoui, S., Mouhoub, M.: Hate and offensive speech detection on Arabic social media. Online Soc. Netw. Media (2020). <https://doi.org/10.1016/j.osnem.2020.100096>
4. Alshalan, R., Al-Khalifa, H.: A Deep Learning approach for automatic hate speech detection in the Saudi Twittersphere. Appl. Sci. **10**(23) (2020). <https://doi.org/10.3390/app10238614>. <https://www.mdpi.com/2076-3417/10/23/8614>
5. Bender, E.M., Friedman, B.: Data statements for natural language processing: Toward mitigating system bias and enabling better science. Trans. Assoc. Computat. Linguist. **6**, 587–604 (2018) [https://doi.org/10.1162/tac1\\_a\\_00041](https://doi.org/10.1162/tac1_a_00041). <https://www.aclweb.org/anthology/Q18-1041>
6. Binns, R., Veale, M., Van Kleek, M., Shadbolt, N.: Like trainer, like bot? inheritance of bias in algorithmic content moderation. In: Ciampaglia, G.L., Mashhadi, A., Yasseri, T. (eds.) Social

- Informatics, pp. 405–415. Springer International Publishing, Cham (2017)
7. Chowdhury, S.A., Mubarak, H., Abdelali, A., Jung, S.g., Jansen, B.J., Salminen, J.: A multi-platform arabic news comment dataset for offensive language detection pp. 6203–6212 (2020). <https://www.aclweb.org/anthology/2020.lrec-1.761>
  8. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography pp. 76–83 (1989). <https://doi.org/10.3115/981623.981633>. <https://www.aclweb.org/anthology/P89-1010>
  9. Davidson, T., Bhattacharya, D., Weber, I.: Racial bias in hate speech and abusive language detection datasets pp. 25–35 (2019). <https://doi.org/10.18653/v1/W19-3504>. <https://www.aclweb.org/anthology/W19-3504>
  10. Davidson, T., Warmley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. Proceedings of the International AAAI Conference on Web and Social Media **11**(1) (2017). <https://ojs.aaai.org/index.php/ICWSM/article/view/14955>
  11. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. *J. Am. Soc. Inform. Sci.* **41**(6), 391–407 (1990). [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:63c391::AID-ASII3e3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:63c391::AID-ASII3e3.0.CO;2-9)
  12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding pp. 4171–4186 (2019). <https://doi.org/10.18653/v1/N19-1423>. <https://www.aclweb.org/anthology/N19-1423>
  13. Dixon, L., Li, J., Sorensen, J., Thain, N., Vasserman, L.: Measuring and mitigating unintended bias in text classification. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18, p. 67–73. Association for Computing Machinery, New York, NY, USA (2018). <https://doi.org/10.1145/3278721.3278729>
  14. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text **51**, 4 (2018). <https://doi.org/10.1145/3232676>
  15. Fortuna, P., Soler, J., Wanner, L.: Toxic, hateful, offensive or abusive? What are we really classifying? an empirical analysis of hate speech datasets. In: Proceedings of the 12th language resources and evaluation conference, pp. 6786–6794. European Language Resources Association, Marseille, France (2020). <https://www.aclweb.org/anthology/2020.lrec-1.838>
  16. Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., Kourtellis, N.: Large scale crowdsourcing and characterization of twitter abusive behavior. Proceedings of the International AAAI Conference on Web and Social Media **12**(1) (2018). <https://ojs.aaai.org/index.php/ICWSM/article/view/14991>
  17. Friedman, B., Nissenbaum, H.: Bias in computer systems. *ACM Trans. Inf. Syst.* **14**(3), 330–347 (1996). <https://doi.org/10.1145/230538.230561>
  18. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K.: Datasheets for datasets. arXiv preprint [arXiv:1803.09010](https://arxiv.org/abs/1803.09010) (2018)
  19. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology. Sage, Oaks (2004)
  20. Kurrek, J., Saleem, H.M., Ruths, D.: Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In: Proceedings of the Fourth Workshop on Online Abuse and Harms, pp. 138–149. Association for Computational Linguistics, Online (2020). <https://doi.org/10.18653/v1/2020.alw-1.17>. <https://www.aclweb.org/anthology/2020.alw-1.17>
  21. Lundberg, S.: shap.PartitionExplainer – SHAP latest documentation (2020). <https://shap.readthedocs.io/en/latest/generated/shap.PartitionExplainer.html>
  22. Lundberg, S.M., Lee, S.I.: A Unified Approach to interpreting model predictions. In: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds.) Advances in neural information processing systems **30**, pp. 4765–4774. Curran Associates, Inc. (2017). <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>
  23. Maaten, Lvd, Hinton, G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**(Nov), 2579–2605 (2008)
  24. MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: Challenges and solutions. *PloS One* **14**(8), e0221152 (2019)
  25. Madukwe, K., Gao, X., Xue, B.: In data we trust: A critical analysis of hate speech detection datasets. In: Proceedings of the Fourth Workshop on Online Abuse and Harms, pp. 150–161. Association for Computational Linguistics, Online (2020). <https://doi.org/10.18653/v1/2020.alw-1.18>. <https://www.aclweb.org/anthology/2020.alw-1.18>
  26. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pre-training distributed word representations. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). Eur. Lang. Resources Assoc. (ELRA), Miyazaki, Japan (2018). <https://www.aclweb.org/anthology/L18-1008>
  27. Mishra, P., Del Tredici, M., Yannakoudakis, H., Shutova, E.: author profiling for abuse detection. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1088–1098. Association for computational linguistics, Santa Fe, New Mexico, USA (2018). <https://www.aclweb.org/anthology/C18-1093>
  28. Mubarak, H., Darwish, K., Magdy, W., Elsayed, T., Al-Khalifa, H.: Overview of OSACT4 arabic offensive language detection shared task. Proceedings of the 4th Workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection, 48–52 (2020). <https://www.aclweb.org/anthology/2020.osact-1.7>
  29. Mulki, H., Haddad, H., Bechikh Ali, C., Alshabani, H.: L-HSAB: A levantine twitter dataset for hate speech and abusive language pp. 111–118 (2019). <https://doi.org/10.18653/v1/W19-3512>. <https://www.aclweb.org/anthology/W19-3512>
  30. Park, J.H., Shin, J., Fung, P.: Reducing gender bias in abusive language detection. In: Proceedings of the 2018 Conference on empirical methods in natural language processing, pp. 2799–2804. Association for computational linguistics, Brussels, Belgium (2018). <https://doi.org/10.18653/v1/D18-1302>. <https://www.aclweb.org/anthology/D18-1302>
  31. Pearson, K.: LIII. On lines and planes of closest fit to systems of points in space *The London, Edinburgh, and Dublin Philosoph. Magaz. J. Sci.* **2**(11), 559–572 (1901)
  32. Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., Patti, V.: Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation* pp. 1–47 (2020)
  33. Raisi, E., Huang, B.: raisi2016cyberbullying. In: Proceedings of the 2016 ICML Workshop on #Data4Good: Machine Learning in Social Good Applications. New York, NY, USA (2016)
  34. Safaya, A., Abdullatif, M., Yuret, D.: KUISAIL at SemEval-2020 Task 12: BERT-CNN for offensive speech identification in social media. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 2054–2059. International Committee for Computational Linguistics, Barcelona (online) (2020). <https://www.aclweb.org/anthology/2020.semeval-1.271>
  35. Salkind, N.: Encyclopedia of research design. Sage, California (2010)
  36. Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N.A.: The risk of racial bias in hate speech detection. In: Proceedings of the 57th Annual Meeting of the association for computational linguistics, pp. 1668–1678. Association for Computational Linguistics,

- Florence, Italy (2019). <https://doi.org/10.18653/v1/P19-1163>. <https://www.aclweb.org/anthology/P19-1163>
37. Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N.A., Choi, Y.: Social bias frames: Reasoning about social and power implications of language. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5477–5490. Association for Computational Linguistics, Online (2020). <https://doi.org/10.18653/v1/2020.acl-main.486>. <https://www.aclweb.org/anthology/2020.acl-main.486>
  38. Schmidt, A., Wiegand, M.: A Survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pp. 1–10. Association for Computational Linguistics, Valencia, Spain (2017). <https://doi.org/10.18653/v1/W17-1101>. <https://www.aclweb.org/anthology/W17-1101>
  39. Vidgen, B., Derczynski, L.: Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS One* **15**, 1–32 (2021). <https://doi.org/10.1371/journal.pone.0243300>
  40. Vidgen, B., Hale, S., Guest, E., Margetts, H., Broniatowski, D., Waseem, Z., Botelho, A., Hall, M., Tromble, R.: Detecting east asian prejudice on social media. In: Proceedings of the Fourth Workshop on Online Abuse and Harms, pp. 162–172. Association for Computational Linguistics, Online (2020). <https://doi.org/10.18653/v1/2020.alw-1.19>. <https://www.aclweb.org/anthology/2020.alw-1.19>
  41. Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., Margetts, H.: Challenges and frontiers in abusive content detection. In: Proceedings of the Third Workshop on Abusive Language Online, pp. 80–93. Association for Computational Linguistics, Florence, Italy (2019). <https://doi.org/10.18653/v1/W19-3509>. <https://www.aclweb.org/anthology/W19-3509>
  42. Viegas, F., Canuto, S., Gomes, C., Luiz, W., Rosa, T., Ribas, S., Rocha, L., Gonçalves, M.A.: CluWords: exploiting semantic word clustering representation for enhanced topic modeling. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pp. 753–761 (2019)
  43. Wang, F.K., Chen, K.S.: Applying Lean Six Sigma and TRIZ methodology in banking services. *Total Quality Management* **21**, 301–315 (2010). <https://doi.org/10.1080/14783360903553248>
  44. Waseem, Z., Hovy, D.: Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In: Proceedings of the NAACL Student Research Workshop, pp. 88–93. Association for Computational Linguistics, San Diego, California (2016). <https://doi.org/10.18653/v1/N16-2013>. <https://www.aclweb.org/anthology/N16-2013>
  45. Wich, M., Al Kuwatly, H., Groh, G.: Investigating Annotator Bias with a Graph-Based Approach. In: Proceedings of the Fourth Workshop on Online Abuse and Harms, pp. 191–199. Association for Computational Linguistics, Online (2020). <https://doi.org/10.18653/v1/2020.alw-1.22>. <https://www.aclweb.org/anthology/2020.alw-1.22>
  46. Wich, M., Bauer, J., Groh, G.: Impact of Politically Biased Data on Hate Speech Classification. In: Proceedings of the Fourth Workshop on Online Abuse and Harms, pp. 54–64. Association for Computational Linguistics, Online (2020). <https://doi.org/10.18653/v1/2020.alw-1.7>. <https://www.aclweb.org/anthology/2020.alw-1.7>
  47. Wiegand, M., Ruppenhofer, J., Kleinbauer, T.: Detection of Abusive Language: the Problem of Biased Datasets. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 602–608. Association for Computational Linguistics, Minneapolis, Minnesota (2019). <https://doi.org/10.18653/v1/N19-1060>. <https://www.aclweb.org/anthology/N19-1060>
  48. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval) pp. 75–86 (2019). <https://doi.org/10.18653/v1/S19-2010>. <https://www.aclweb.org/anthology/S19-2010>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A.2 STUDY V

©2021 The Author(s).

Maximilian Wich, Christian Widmer, Gerhard Hagerer, and Georg Groh (Sept. 2021). "Investigating Annotator Bias in Abusive Language Datasets." In: *Deep Learning for Natural Language Processing Methods and Applications*. Held Online: INCOMA Ltd., pp. 1515–1525. ISBN: 978-954-452-072-4. DOI: [10.26615/978-954-452-072-4\\_170](https://doi.org/10.26615/978-954-452-072-4_170). URL: <https://aclanthology.org/2021.ranlp-1.170>

### *Publication Summary*

"Nowadays, social media platforms use classification models to cope with hate speech and abusive language. The problem of these models is their vulnerability to bias. A prevalent form of bias in hate speech and abusive language datasets is annotator bias caused by the annotator's subjective perception and the complexity of the annotation task. In our paper, we develop a set of methods to measure annotator bias in abusive language datasets and to identify different perspectives on abusive language. We apply these methods to four different abusive language datasets. Our proposed approach supports annotation processes of such datasets and future research addressing different perspectives on the perception of abusive language." (Wich, Widmer, et al., 2021, p. 1)

### *Author Contributions*

Maximilian Wich headed the research project. He developed the initial idea, the concept, and the methodology of the paper. Furthermore, he implemented a substantial part of the code that is partially based on Christian Widmer's code. Additionally, he wrote most of the manuscript. Christian Widmer implemented and conducted some experiments used in the study as part of his master's thesis supervised by Maximilian Wich. He was responsible for a smaller part of the paper. Furthermore, he served as a valuable discussion partner. Gerhard Hagerer provided input on the idea behind the study and reviewed the methodology. In addition, he corrected the proofs and provided feedback on the manuscript. Georg Groh regularly discussed the ideas and concepts with the team and provided feedback on the paper.

# Investigating Annotator Bias in Abusive Language Datasets

**Maximilian Wich, Christian Widmer, Gerhard Hagerer, Georg Groh**

Technical University of Munich, Department of Informatics, Germany

maximilian.wich@tum.de, christian.widmer@tum.de,

ghagerer@mytum.de, grohg@in.tum.de

## Abstract

Nowadays, social media platforms use classification models to cope with hate speech and abusive language. The problem of these models is their vulnerability to bias. One form of bias that is prevalent in abusive language datasets is annotator bias that is caused by the annotator’s subjective perception and the complexity of the annotation task. In our paper, we develop a set of methods to measure annotator bias in abusive language datasets and to identify different perspectives on abusive language. We apply these methods to four different abusive language datasets. Our proposed approaches can support annotation processes of such datasets and future research addressing different perspectives on the perception of abusive language.

## 1 Introduction

A challenge that social media platforms are facing in recent years is the large amount of hate speech and other forms of abusive language (Duggan, 2017). Manual monitoring, however, is no longer possible due to the vast volume of user-generated content. Therefore, machine learning models are trained and used by social media platforms, such as Facebook, to automatically detect such content (Kantor, 2020). According to Rose (2021), these models are a key component of Facebook’s fight against hate speech.

A problem with such machine learning models is that they are vulnerable to bias (Vidgen and Derczynski, 2021; Dixon et al., 2018). Biased models can strongly impair the fairness of a system, which can, for example, lead to discrimination of minorities (Dixon et al., 2018).

Bias in abusive language detection is already a topic that researchers have started to investigate (Vidgen and Derczynski, 2021; Wich et al., 2021). The type of bias we want to focus on in this study

is annotator bias. This form of bias is brought in by the annotators who perceive abusive language differently from each other and have different levels of experience and knowledge (Ross et al., 2016; Waseem, 2016; Geva et al., 2019; Wich et al., 2020).

We aim to investigate two aspects of annotator bias. (1) Assuming that there is only one perspective on whether a text is abusive or not, we want to develop an approach to measure and visualize the annotator bias. Such an approach is supposed to optimize the annotation process (e.g. outlier detection, adapting annotation guidelines in the right way). (2) Assuming that for each text to classify there are multiple valid views (e.g., a group has a more liberal attitude towards abusive texts, while another is stricter), we aim to identify annotator groups to model the different (valid) perspectives. The research questions resulting from these research objectives are the following:

- RQ1: How can we measure and visualize annotator bias in abusive language datasets?
- RQ2: How can we identify and visualize different perspectives on abusive language of the annotators?

Our contributions are the following:

1. An approach to characterize annotators in regard to how liberally or strictly they annotated in comparison to the other annotators. To model this annotator bias, we calculate for each annotator a pessimistic and optimistic score that can be visualized in different ways (e.g., scatter plot, cluster map). We apply it to four English abusive language datasets with different numbers of annotators.
2. A method to use the proposed approach to identify annotator groups with different per-



Name	Documents	Source	Labels	Annotators	Expert check	Reference
<i>Vidgen</i>	20,000	Twitter	<b>hostility</b> , criticism, counter speech, discussion of east Asian prejudice, non-related	26	yes	<a href="#">Vidgen et al. (2020)</a>
<i>Guest</i>	6,567	Reddit	<b>misogynistic</b> , non-misogynistic	6	yes	<a href="#">Guest et al. (2021)</a>
<i>Kurrek</i>	40,000	Reddit	<b>derogatory usage</b> , appropriative usage, non-derogatory usage, homonyms	20	yes	<a href="#">Kurrek et al. (2020)</a>
<i>Wulczyn</i>	115,864	Wikipedia (discussion)	<b>attack</b> , non-attack	4,053	no	<a href="#">Wulczyn et al. (2017)</a>

Table 1: Overview of selected abusive datasets (class names in bold are the abusive categories, the others the neutral ones)

spectives on abusive language. This method is applied to one dataset.

## 2 Related Work

Hate speech and abusive language detection have gained a lot of attention in recent years. A range of different datasets ([Vidgen and Derczynski, 2021](#)) and shared tasks ([Basile et al., 2019](#); [Zampieri et al., 2019, 2020](#)) were published to foster research in this area. Most of the datasets are commonly labeled by crowdworkers or academics with varying expertise ([Vidgen and Derczynski, 2021](#)). However, human annotations tend to be subjective and thus inconsistent ([Aroyo and Welty, 2015](#)), at least if not moderated very strictly. Especially for abusive language [Salminen et al. \(2018\)](#) show that individuals interpret hate differently. One common method to improve the label quality is presenting each sample to multiple annotators and aggregate their results ([Sheng et al., 2008](#)). [Dawid and Skene \(1979\)](#) were the first to propose an approach that incorporates annotator quality into label aggregation. Their expectation-maximization (EM) algorithm uses the bias matrices to estimate the latent truth. In the matrices the annotator quality is encoded. Their seminal work led to further improvements and methods ([Whitehill et al., 2009](#); [Raykar and Yu, 2012](#); [Hovy et al., 2013](#)). For NLP tasks, [Snow et al. \(2008\)](#) used a customized Dawid-Skene algorithm to correct for individual biases of crowdworkers and improve model accuracy. They did, however, not quantify and inspect the bias of the annotators.

In abusive language detection, annotator bias research has focused on how the annotators background influences their annotations. [Waseem](#)

(2016) found models trained on crowd annotations are outperformed by models trained on expert annotations. [Ross et al. \(2016\)](#) emphasized the importance of detailed guidelines to achieve reliable annotations. [Binns et al. \(2017\)](#) showed that classifiers trained on annotations by men or women differ in their performance on test data annotated by men or women. [Al Kuwatly et al. \(2020\)](#) extended this approach to other demographics and found significant differences for annotator’s age group and educational background. [Sap et al. \(2019\)](#) observed that posts in African American dialect are more likely to be labeled offensive. Similarly, [Larimore et al. \(2021\)](#) found that white and non-white workers annotate racially sensitive topics differently. Apart from studying the demographic background, researchers also attempt to find groups of annotators with common annotation behavior. [Wich et al. \(2020\)](#) use graph methods to cluster annotators in groups with higher inter-annotator agreement within groups than across groups. [Akhtar et al. \(2020\)](#) defined a polarization measure to split annotators in two groups that maximize opposing annotations. To the best of our knowledge, no one has quantified annotator bias on annotator-level. Furthermore, the hypothesis of multiple perspective on abusive language is hardly investigated.

## 3 Datasets

We use four different English abusive language datasets to demonstrate our proposed approaches. It was challenging to find appropriate datasets because our experiments require unaggregated annotation data. Most of the abusive language datasets contain only the agreed labels for the documents

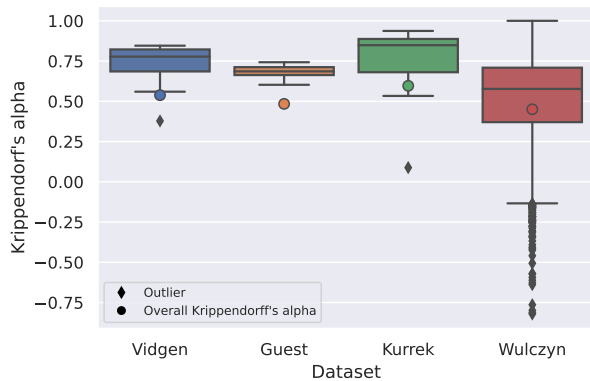


Figure 1: Box plots of the annotators' inter-rater reliability scores

and not the individual votes of the annotators.

Table 1 lists the four datasets with additional information. The first three datasets (*Vidgen*, *Guest*, and *Kurrek*) are similar because they are annotated by small groups of annotators (between 6 and 26). Furthermore, each document of the three datasets was annotated by two annotators. In case of disagreement, an expert reviewed the votes and decided on the gold label. In contrast, the *Wulczyn* dataset was annotated by many crowdworkers – a typical crowdsourcing setup: numerous workers, on average a small number of annotated documents per worker. Each document was annotated by up to 10 annotators. An expert review in case of ambiguous annotations did not take place. The gold label was determined based on majority vote. For our experiment, we convert all datasets to a binary task (abusive/neutral) to compare the results.

Figure 1 shows the distributions of the annotators' inter-rater reliability scores in form of Krippendorff's alpha. The colored dots represent the overall inter-rater reliability score of each dataset. We can see that the overall Krippendorff's alphas are all in the same range. The *Wulczyn* dataset, however, exhibits a considerable variance in contrast to the other three datasets. The reason is that 4,053 crowd workers annotated this dataset, while a small and instructed group of workers annotated the other three datasets. Therefore, we see many outliers. In the case of the *Vidgen* and *Kurrek* dataset, only one annotator strongly differs from the others.

## 4 Methodology

Our analysis of the annotator bias in the selected abusive datasets consists of two parts. In the first part, we characterize the annotation behavior based on the deviations of the annotator votes compared

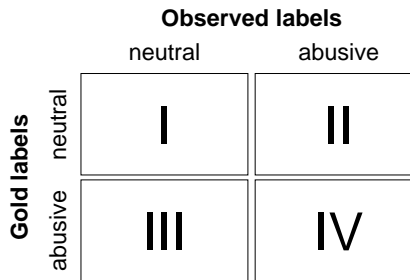


Figure 2: Bias matrix of an annotator

with the gold standard of the dataset. In the second part, we visualize the perspectives of different annotator groups on abusive language with the aid of classification models.

### 4.1 Characterizing Annotator Bias

We define annotator bias as the deviations between the annotator votes and the gold labels of the dataset. The gold labels are either the final labels of the dataset or the majority of the single votes. To measure the annotator bias, we use the concept of the confusion matrix. Figure 2 shows such a matrix for a binary classification task of abusive documents (neutral/abusive). The rows represent the classes of the gold labels, the columns the classes observed by the annotator. We call this matrix bias matrix because it quantifies the deviations between the labels observed by the annotator and the gold labels. Each annotator has one bias matrix.

We use cells II and III, which represent false negatives (type II error) and false positives (type I error) in the classical confusion matrix, to characterize the annotators' behavior, and we assign each annotator a pessimistic and optimistic score. Cell II reflects the number of documents that are neutral according to the gold standard but that are annotated as abusive by the annotator. That means that the annotator is pessimistic in these cases. Cell III is the opposite. It shows the number of documents that are labeled as abusive according to the gold standard but perceived as neutral by the annotator. That means that the annotator is optimistic in these cases. The pessimistic ( $p_i$ ) and optimistic ( $o_i$ ) scores of an annotator ( $i$ ) are entries II and III of row-normalized bias matrix. The concept of annotator's optimism and pessimism was proposed by Dawid and Skene (1979). It also works if we have more than two classes as long as they are ordinal. In this case, the cells above or below the diagonal are summed up. In our paper, however, we consider only binary classification tasks.

To analyze bias matrices, we have these options:

1. We can calculate the bias matrix for a group of annotators or all of them by averaging the bias matrices. The resulting bias matrix tells us whether the selected annotators rather tend to be optimistic or pessimistic. Such a finding can help to adjust annotation guidelines or the training of the annotators.
2. We can use a 2-dimensional scatter plot of the pessimistic and optimistic scores to visualize the annotators and their biases. In contrast to comparing inter-rater reliability scores, this visualization reveals whether annotators are more optimistic or pessimistic than the gold standard. Such information can help to detect outliers in the respective direction and to instruct the identified annotators in the right way.
3. We can calculate a distance between two annotators based on their bias matrices ( $A$  and  $B$ ) with the Frobenius norm (Golub and Van Loan, 2013, p.71):

$$distance(\mathbf{A}, \mathbf{B}) = \sqrt{\sum_{i=1}^n \sum_{j=1}^n (a_{ij} - b_{ij})^2}$$

Visualizing these distances with a hierarchically clustered heatmap helps identify annotator groups with similar annotation behavior and outliers.

4. If the number of annotators is so large that the results of the previously proposed methods is longer manageable, we can apply a hierarchical clustering on the bias matrices based our distance metric. By doing so, annotators with a similar annotation behavior are clustered. If we aggregate the bias matrices according to (2), we receive an impression how the cluster annotates the data in context of the gold standard.
5. If we have additional information about the annotators to characterize them (e.g., demographics, such as age or education), we can use the pessimistic and optimistic scores to test whether there is a significant difference between the annotation behavior of annotators with different characteristics. For this purpose, we apply the two-dimensional version of the Kolmogorov-Smirnov test (Fasano

and Franceschini, 1987; Peacock, 1983) to compare the distributions of the groups' pessimistic and optimistic scores. The output of the test is the Kolmogorov-Smirnov statistic  $D$  and the corresponding significance level  $s$ . If  $D$  is larger than the predefined significance level  $p$  and  $p$  larger than  $s$ , we can reject the null hypothesis that both samples have the same distribution. We use the Python implementation provided by Gabriel Taillon<sup>1</sup>. In the case of the *Wulczyn* dataset, we have such information (Wulczyn et al., 2017). Our predefined significance level  $p$  is 0.05.

#### 4.2 Identifying different perspectives on abusive language

The previous subsection focuses on methods to measure and visualize annotator bias, answering RQ1. The underlying assumption is that there is one truth, and we want to identify outliers deviating from the one truth.

Now we assume that there are more perspectives on abusive language — e.g., a group has a more liberal attitude towards abusive texts, while another group is less liberal. To examine this hypothesis, we run the following experiment: First, we split the annotators into different groups based on the pessimistic and optimistic scores. Second, we create for each group a dataset containing the documents that all groups annotated. The labels of the documents result from the majority vote of the groups' annotators. Third, we train for each group a classifier on its training set and evaluate it on the test sets of all groups. Suppose a classifier performs well on its test set and worse on the other test sets. Well means that the performance is comparable to a baseline classifier trained on the same data with gold labels. In that case, it indicates that this group has a coherent perspective on abusive language that differs from the other groups. This approach is based on the one proposed by Wich et al. (2020).

To split the annotators according to their pessimistic ( $p_i$ ) and optimistic  $o_i$  scores, we apply the following function:

$$group_a(p_i, o_i) = \begin{cases} 0 & \text{if } p_i \geq 3 \cdot o_i \\ 1 & \text{if } o_i \geq 3 \cdot p_i \\ 2 & \text{otherwise} \end{cases}$$

The factor 3 in the function is the result of a trade-off between having a dominating dimension

<sup>1</sup><https://github.com/Gabinou/2DKS>

in the optimistic and pessimistic group and having enough annotators in all three groups. Increasing the factor would strengthen the dominating dimensions but reduce the amount of annotators in the optimistic and pessimistic groups. Decreasing the factor would weaken the dominating dimension but increase the number of annotators in the groups.

For the classification model, we use the pre-trained English DistilBERT model `distilbert-base-uncased` provided by the Transformer library from Huggingface (Wolf et al., 2020). It is a smaller and faster version of BERT (Sanh et al., 2019; Devlin et al., 2019). In the context of abusive language detection, it provides a performance comparable to BERT (Devlin et al., 2019). We train each model for three epochs with a learning rate of  $5 \cdot 10^{-5}$  and a batch size of 32. After the three epochs, we select the model with the lowest validation loss. 60% of the documents annotated by all groups are used as training set, 20% as validation set, and 20% as test set. To compare the different classifiers, we use the macro F1 score.

## 5 Results

### 5.1 Characterizing Annotator Bias

#### Aggregated Bias Matrix

The problem of the inter-rater reliability analysis is that it does not reveal whether the annotators annotated more pessimistically or optimistically. This gap is addressed by the aggregated bias matrices, shown in Figure 3. The annotators of the datasets *Vidgen*, *Guest*, and *Wulczyn* tended to annotate more liberally because the optimistic scores (bottom-left cell) outweigh the pessimistic ones (upper-right cell). On the contrary, the annotators of the *Kurrek* dataset were stricter because 16% of non-derogatory documents were labeled as derogatory (pessimistic score), while only 4.5% (optimistic score) of the derogatory documents were labeled as non-derogatory.

#### Scatter plot of Annotators

To gain a better understanding of the individual annotation behavior, we analyze the annotators based on their pessimistic and optimistic scores, shown in Figure 4. Considering the plots of *Vidgen*, *Guest*, and *Kurrek*, we observe that the annotators of *Vidgen* and *Guest* annotated more liberally due to the higher optimistic scores, while it is the opposite for the *Kurrek* dataset. Comparing the *Guest* dataset

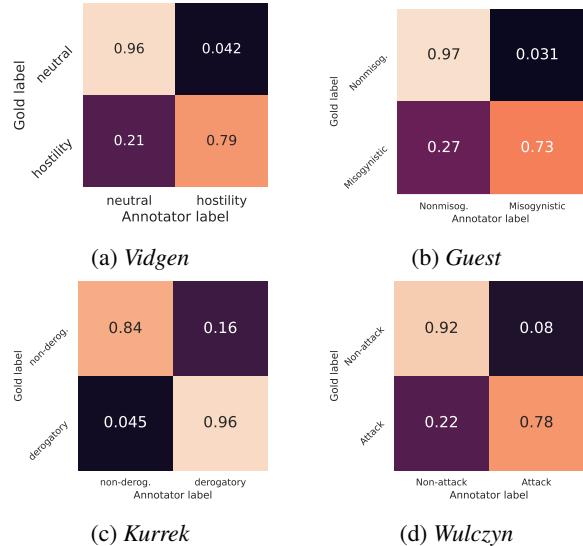


Figure 3: Aggregated bias matrices for the selected datasets

with the other two, we see that the annotators are less widely spread, meaning the annotation behavior is more coherent. Concerning the previously mentioned outliers of *Vidgen* and *Kurrek*, we can use the plots to better understand how they deviate. The outlier of *Vidgen* is the most right point, the one of *Kurrek* the uppermost point. Their positions reveal that the outlier of *Vidgen* annotated more liberally, while the one of *Kurrek* was stricter. These findings can help to instruct the annotators if the method is used during the annotation process. A further observation concerning both datasets is that the density of annotators increases towards the origin of both dimensions. This indicates that most of the annotators have a similar annotation behavior.

In the case of the *Wulczyn* dataset, plotting each annotator as a data point would be confusing because the dataset contain 4,053 annotators. Therefore, we decided to cluster the annotators with a hierarchical clustering approach, making interpreting easier. We chose the agglomerative clustering approach with  $k = 30$  and euclidean distance function. The reason for  $k = 30$  is that it is a manageable amounts of data points on a scatter plot and it has the same order of magnitude as *Vidgen* and *Kurrek*. Figure 4d shows the annotators' clusters. We observe the tendency of the annotators to annotate more liberally, as shown by the aggregated bias matrix in Figure 3d.

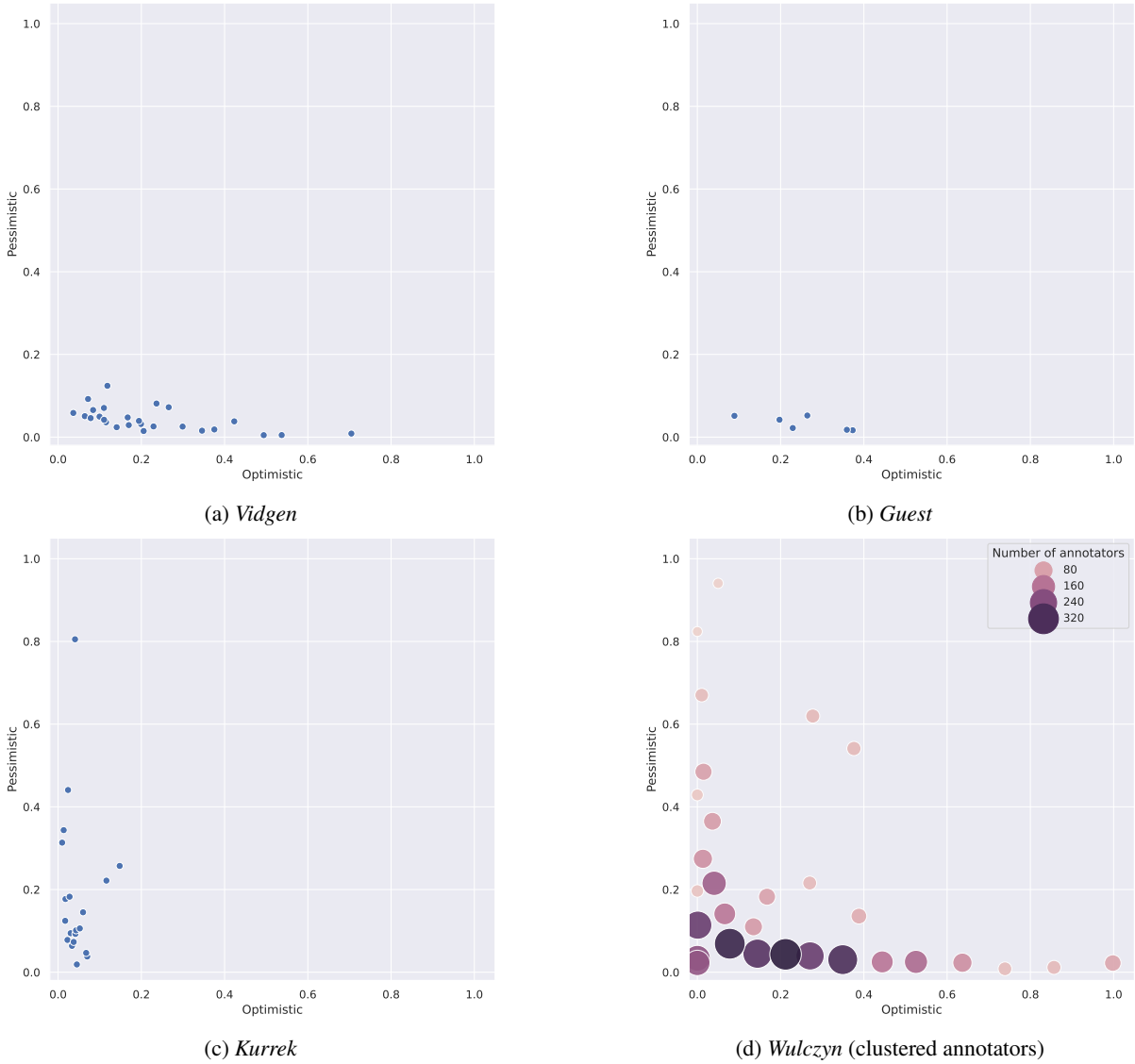


Figure 4: Annotators visualized based on their pessimistic and optimistic scores; in case of *Wulczyn*, annotators are hierarchically clustered

### Cluster Map of Distances between Annotators

A method to identify groups of annotators with similar annotation behavior is the hierarchically clustered heatmap based on the distances between the bias matrices of the annotators. Figure 5 shows the cluster map of the *Kurrek* dataset. The first thing that leaps to the reader’s eye is the first column and row. It shows the outlier of the dataset. Furthermore, we observe that the annotators Ann7, Ann13, Ann15, Ann3, and Ann5 (last five columns and rows) form a group. In Figure 4c, these annotators are the points above a pessimistic score of 0.2 and below 0.6. The other 15 annotators exhibit a more coherent annotation behavior. Due to the page restriction, we do not include this analysis for the other three datasets.

### Different Annotation behavior of Demographic Groups

The *Wulczyn* dataset contains demographic information for 2,190 of the 4,053 annotators (gender, age group, education, and English as the first language). We tested for each demographic feature whether there is a difference between the groups regarding the annotators’ pessimistic and optimistic scores. The result of the two-dimensional Kolmogorov-Smirnov for the demographic feature gender is the following:

$$D_{\text{gender}} = 0.092 \quad s_{\text{gender}} = 0.005$$

That means that we can reject the null hypothesis ( $p = 0.05$ ). Consequently, there is a significant difference between the pessimistic and opti-

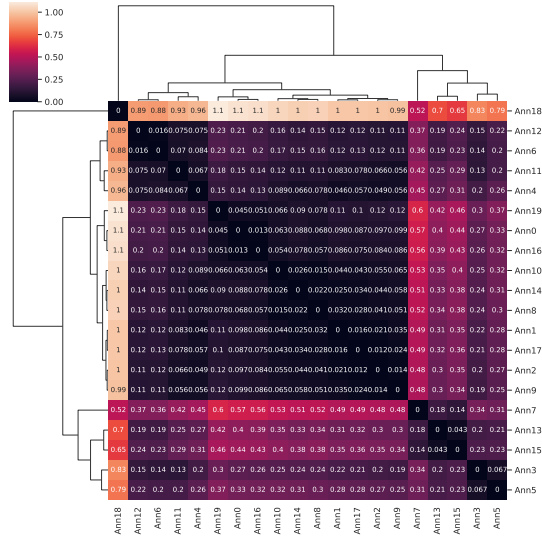


Figure 5: Cluster map of annotators’ distances from *Kurrek* dataset

mistic scores of male and female annotators. Females are more pessimistic than males ( $p_{\text{female}} = 0.107$   $p_{\text{male}} = 0.090$ ), while the optimistic scores are comparable ( $o_{\text{female}} = 0.192$   $o_{\text{male}} = 0.199$ ). For the feature describing whether English is the first language of the annotator or not, we can also reject the null hypothesis:

$$D_{\text{1st language}} = 0.192 \quad s_{\text{1st language}} = 8.9 \times 10^{-8}$$

Native English speakers have a larger pessimistic score ( $p_{\text{native}} = 0.093$   $p_{\text{non-native}} = 0.117$ ) and a lower optimistic score than non-native speakers ( $o_{\text{native}} = 0.160$   $o_{\text{non-native}} = 0.204$ ).

Table 3a shows the output of the two-dimensional Kolmogorov-Smirnov test for the different age groups. We observe that there are significant differences in the distributions of the annotators’ pessimistic and optimistic scores between the age groups — except 30-45 and over 60 and 45-60 and over 60. Interestingly, the largest difference is between the age groups 18-30 and 45-60. While annotators between 45 and 60 are more pessimistic ( $p_{45-60} = 0.146$   $o_{45-60} = 0.128$ ), it is the opposite for annotators between 18 and 30 ( $p_{18-30} = 0.08518$   $o_{18-30} = 0.234$ ).

Table 3b shows the output for the different educational backgrounds. In contrast to the age groups, the scores of the annotators do not distinguish much between the groups. Only the difference between the Bachelors and Masters is significant.

## 5.2 Identifying Different Perspectives on Abusive Language

Since this experiment requires a dataset with a large number of documents and annotators, we could conduct it only with the *Wulczyn* dataset. In the case of the other three datasets, the number of annotators is too small to meaningfully split the annotators into subsets and to have enough documents that were annotated by all subsets.

The results of the experiment to identify different perspectives and to answer RQ2 can be found in Table 2. It shows the different F1 scores for the abusive class of the classifiers that were trained on subsets of annotators (rows) and were evaluated on the test sets of these subgroups (columns).

	Pessimistic	Medium	Optimistic
Pessimistic	80.2	80.6	71.0
Medium	73.5	81.9	83.1
Optimistic	64.3	74.4	87.5

Table 2: F1 scores from classifiers of the different annotator subsets

To answer our research question RQ2 how to identify and visualize different perspectives on abusive language of the annotators, we need to focus on the pessimistic and optimistic rows and columns. We observe that the classifier trained on the annotations of the optimistic annotators performs best on its own test set (87.5%) and worst on the pessimistic test set (64.5%). For the classifier trained on the more pessimistic annotations, it is almost the opposite. It performs most poorly on the optimistic test set (71.0%) and comparable well on its own test set (80.2%). Only on the test set of the medium group, the pessimistic classifier performs slightly better.

But it is more relevant to our research question that the pessimistic and optimistic classifiers work well on their own test set but worse on the test set of the other extreme. The first fact indicates that the annotations are coherent, so that the classifier can learn patterns to identify abusive language. The second aspect shows that the labels of the pessimistic and optimistic subgroups’ dataset are so different that it can cause a difference of 9.2pp or 23.2pp in the F1 score. Consequently, we conclude that the annotators of the pessimistic and optimistic subgroup have two different perspectives on abusive language.

An explanation for the more coherent results of

	Under 18	18-30	30-45	45-60		Pessimistic	Optimistic
Under 18	-	-	-	-		0.080	0.172
18-30	<b>0.261 / 0.040</b>	-	-	-		0.085	0.234
30-45	<b>0.303 / 0.011</b>	<b>0.177 / 0.000</b>	-	-		0.100	0.165
45-60	<b>0.435 / 0.000</b>	<b>0.322 / 0.000</b>	<b>0.216 / 0.000</b>	-		0.146	0.126
Over 60	<b>0.416 / 0.031</b>	<b>0.377 / 0.016</b>	0.248 / 0.249	0.165 / 0.775		0.125	0.140

(a) Age group of annotators

	some	hs	bachelors	masters	doctorate		Pessimistic	Optimistic
some	-	-	-	-	-		0.085	0.210
hs	0.116 / 0.738	-	-	-	-		0.096	0.195
bachelors	0.109 / 0.790	0.059 / 0.341	-	-	-		0.096	0.193
masters	0.141 / 0.520	0.070 / 0.378	<b>0.102 / 0.040</b>	-	-		0.098	0.206
doctorate	0.175 / 0.827	0.217 / 0.407	0.199 / 0.516	0.231 / 0.346	-		0.075	0.216
professional	0.136 / 0.597	0.104 / 0.134	0.070 / 0.530	0.124 / 0.074	0.184 / 0.647		0.109	0.190

(b) Educational background of annotators

Table 3: Results of 2-dimensional Kolmogorov-Smirnov test for split according to demographic features and corresponding pessimistic and optimistic scores (*Wulczyn* dataset); first number in cells is  $D$ , second  $s$ ; bold means rejected

the optimistic classifier can be the larger number of annotators. While it comprises annotations from 1,708 annotators, the pessimistic subset contains only 1,312. As we can see, this difference is in line with the finding that the annotators of the *Wulczyn* dataset tended to annotate more liberally.

## 6 Discussion

The first part of our study addressing RQ1 shows that the proposed approach based on the pessimistic and optimistic scores helps to measure and visualize the difference in the annotation behavior of annotators. In contrast to the inter-rater reliability, our method reveals information about the tendency of the annotators: Did they annotate more liberally or stricter than the group average? These findings can be used to understand outliers better, instruct single annotators in the right direction to align them with the rest of the group or adapt the annotation guidelines. Our approach comprises a range of methods, from an aggregated perspective on all annotations to cluster analyses to evaluations of individual annotators. This variety allows handling datasets with different numbers of annotators.

The proposed approach is unsupervised by itself because it does not require any labeled data. But it can be combined with additional data, as shown by the experiment with the demographic features. We showed that it can help to detect annotator bias caused by different demographic backgrounds. Our results are partially in line with the findings from Al Kuwatly et al. (2020), who

examined the same dataset but with a different approach. We confirmed the differences between native and non-native speakers and between the age groups. In our case, we identified a significant difference between male and female annotators, which Al Kuwatly et al. (2020) did not find. In contrast to us, they observed a stronger difference between the educational backgrounds. The reason for the discrepancy can be the different methods. They trained classifiers on different subsets and compared their performances, as we did for the second part of our study. Furthermore, they had to group the educational backgrounds to have enough data. Consequently, the results can differ. The advantage of our approach over the classifier-based one used by Al Kuwatly et al. (2020) and by Binns et al. (2017) on another dataset is that we do not rely on a classifier as we can use the full dataset.

The underlying assumption for the first part of the study is that there is only one ground truth whether a text is abusive or not. That means that we all share the same understanding. In the second part of the study, we had the controversial assumption that there are different perspectives on the perception of abusive language. Our goal was to use our proposed method to identify different perspectives and to visualize the differences. By splitting the annotators according to the ratio between the pessimistic and optimistic scores and training different classifiers for these annotators subsets, we showed that there are different perspectives on abusive language. The classifiers of the pessimistic

and optimistic annotator subsets perform well on their own test set and poorly on the test set of the other subset. That means that the perception of abusive language within each group is coherent, but it differs from the perception of the other subset.

The multiple perspectives on abusive language are a research object that should be further investigated. Akhtar et al. (2020), for example, showed that balancing different perspectives in the training set can improve the classification performance. But we can also imagine building classification models that model the different perspectives. That means that we would have for each group an own model that is trained on the groups' individual values and perceptions.

## 7 Conclusion

In this paper, we presented a novel approach for measuring and visualizing annotator bias purely on their annotation behavior. The approach can help to better understand the annotation behavior, detect outliers, and gain insights on how to adapt annotation guidelines. Furthermore, we showed that there can be different perspectives on abusive language. Using our proposed approach, we can identify these perspectives and make the differences visible.

## Resources

The code is available under <https://github.com/mawic/annotator-bias-abusive-language>.

## Acknowledgments

This paper is based on a joined work in the context of Christian Widmer's master's thesis (Widmer, 2021). This research has been partially funded by a scholarship from the Hanns Seidel Foundation financed by the German Federal Ministry of Education and Research.

## References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 151–154.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *International conference on social informatics*, pages 405–415. Springer.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Maeve Duggan. 2017. *Online harassment 2017*. Pew Research Center.

Giovanni Fasano and Alberto Franceschini. 1987. A multidimensional version of the kolmogorov-smirnov test. *Monthly Notices of the Royal Astronomical Society*, 225(1):155–170.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *2019 Conference on Empirical Methods in Natural Language Processing*, pages 1161–1166.

Gene H Golub and Charles F Van Loan. 2013. *Matrix computations*, volume 4. The Johns Hopkins University Press.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.



- Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.
- Arcadiy Kantor. 2020. [Measuring Our Progress Combating Hate Speech](#).
- Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. [Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage](#). In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 138–149, Online. Association for Computational Linguistics.
- Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90.
- John A Peacock. 1983. Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society*, 202(3):615–627.
- Vikas C Raykar and Shipeng Yu. 2012. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *The Journal of Machine Learning Research*, 13(1):491–518.
- Guy Rose. 2021. [Community Standards Enforcement Report, First Quarter 2021](#).
- Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wotzki. 2016. [Measuring the reliability of hate speech annotations: the case of the european refugee crisis](#). In *NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*.
- Joni Salminen, Fabio Veronesi, Hind Almerkhi, Soon-Gvo Jung, and Bernard J Jansen. 2018. Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 88–94. IEEE.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. [The risk of racial bias in hate speech detection](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.
- Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622.
- Rion Snow, Brendan Oconnor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.
- Bertie Vidgen and Leon Derczynski. 2021. [Directions in abusive language training data, a systematic review: Garbage in, garbage out](#). *PLOS ONE*, 15(12):1–32.
- Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. [Detecting East Asian prejudice on social media](#). In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.
- Zeerak Waseem. 2016. [Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter](#). In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.
- Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22:2035–2043.
- Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020. [Investigating annotator bias with a graph-based approach](#). In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 191–199, Online. Association for Computational Linguistics.
- Maximilian Wich, Tobias Eder, Hala Al Kuwatly, and Georg Groh. 2021. [Bias and comparison framework for abusive language datasets](#). *AI and Ethics*.
- Christian Widmer. 2021. Investigation of bias in hate speech classifications. Master’s thesis, Technical University of Munich. Advised and supervised by Maximilian Wich and Georg Groh.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*:

*System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447.

## A.3 STUDY VI

©2020 Association for Computational Linguistics, published under Creative Commons CC-BY 4.0 License<sup>2</sup>.

Maximilian Wich, Jan Bauer, and Georg Groh (Nov. 2020). “Impact of Politically Biased Data on Hate Speech Classification.” In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, pp. 54–64. DOI: [10.18653/v1/2020.alw-1.7](https://doi.org/10.18653/v1/2020.alw-1.7). URL: <https://www.aclweb.org/anthology/2020.alw-1.7>

---

<sup>2</sup> <https://creativecommons.org/licenses/by/4.0/>

### *Publication Summary*

"One challenge that social media platforms are facing nowadays is hate speech. Hence, automatic hate speech detection has been increasingly researched in recent years — in particular with the rise of deep learning. A problem of these models is their vulnerability to undesirable bias in training data. We investigate the impact of political bias on hate speech classification by constructing three politically-biased data sets (left-wing, right-wing, politically neutral) and compare the performance of classifiers trained on them. We show that (1) political bias negatively impairs the performance of hate speech classifiers and (2) an explainable machine learning model can help to visualize such bias within the training data. The results show that political bias in training data has an impact on hate speech classification and can become a serious issue." (Wich, Bauer, and Groh, 2020, p. 54)

### *Author Contributions*

Maximilian Wich headed the research project. He developed the initial idea, the concept, and the methodology of the study. Furthermore, he collected the data for the project. Additionally, he wrote most of the manuscript. Jan Bauer developed the models and ran experiments. In addition, he provided feedback on the paper. Georg Groh regularly discussed the ideas and concepts with the team and provided feedback on the paper.

# Impact of Politically Biased Data on Hate Speech Classification

**Maximilian Wich**  
TU Munich,  
Department of Informatics,  
Germany  
maximilian.wich@tum.de

**Jan Bauer**  
TU Munich,  
Department of Informatics,  
Germany  
jan.bauer@tum.de

**Georg Groh**  
TU Munich,  
Department of Informatics,  
Germany  
grohg@in.tum.de

## Abstract

One challenge that social media platforms are facing nowadays is hate speech. Hence, automatic hate speech detection has been increasingly researched in recent years — in particular with the rise of deep learning. A problem of these models is their vulnerability to undesirable bias in training data. We investigate the impact of political bias on hate speech classification by constructing three politically-biased data sets (left-wing, right-wing, politically neutral) and compare the performance of classifiers trained on them. We show that (1) political bias negatively impairs the performance of hate speech classifiers and (2) an explainable machine learning model can help to visualize such bias within the training data. The results show that political bias in training data has an impact on hate speech classification and can become a serious issue.

## 1 Introduction

Social media platforms, such as Twitter and Facebook, have gained more and more popularity in recent years. One reason is their promise of free speech, which also obviously has its drawbacks. With the rise of social media, hate speech has spread on these platforms as well (Duggan, 2017). But hate speech is not a pure online problem because online hate speech can be accompanied by offline crime (Williams et al., 2020).

Due to the enormous amounts of posts and comments produced by the billions of users every day, it is impossible to monitor these platforms manually. Advances in machine learning (ML), however, show that this technology can help to detect hate speech — currently with limited accuracy (Davidson et al., 2017; Schmidt and Wiegand, 2017).

There are many challenges that must be addressed when building a hate speech classifier. First of all, an undesirable bias in training data can cause

models to produce unfair or incorrect results, such as racial discrimination (Hildebrandt, 2019). This phenomenon is already addressed by the research community. Researchers have examined methods to identify and mitigate different forms of bias, such as racial bias or annotator bias (Geva et al., 2019; Davidson et al., 2019; Sap et al., 2019). But it has not been solved yet; on the contrary, more research is needed Vidgen et al. (2019). Secondly, most of the classifiers miss a certain degree of transparency or explainability to appear trustworthy and credible. Especially in the context of hate speech detection, there is a demand for such a feature Vidgen et al. (2019); Niemann (2019). The reason is the value-based nature of hate speech classification, meaning that perceiving something as hate depends on individual and social values and social values are non-uniform across groups and societies. Therefore, it should be transparent to the users what the underlying values of a classifier are. The demand for transparency and explainability is also closely connected to bias because it can help to uncover the bias.

In the paper, we deal with both problems. We investigate a particular form of bias — political bias — and use an explainable AI method to visualize this bias. To our best knowledge, political bias has not been addressed in hate speech detection, yet. But it could be a severe issue. As an example, a moderator of a social media platform uses a system that prioritizes comments based on their hatefulness to efficiently process them. If this system had a political bias, i.e. it favors a political orientation, it would impair the political debate on the platform. That is why we want to examine this phenomenon by addressing the following two research questions:

**RQ1** What is the effect of politically biased data sets on the performance of hate speech classi-

fiers?

**RQ2** Can explainable hate speech classification models be used to visualize a potential undesirable bias within a model?

We contribute to answering these two questions by conducting an experiment in which we construct politically biased data sets, train classifiers with them, compare their performance, and use interpretable ML techniques to visualize the differences.

In the paper, we use hate speech as an overarching term and define it as "any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic" (Nockleby (2000, p.1277)), as cited in Schmidt and Wiegand (2017)).

## 2 Related Work

### 2.1 Biased Training Data and Models

A challenge that hate speech detection is facing is an undesirable bias in training data (Hildebrandt, 2019). In contrast to the inductive bias — the form of bias required by an algorithm to learn patterns (Hildebrandt, 2019) — such a bias can impair the generalizability of a hate speech detection model (Wiegand et al., 2019; Geva et al., 2019) or can lead to unfair models (e.g., discriminating minorities) (Dixon et al., 2018).

There are different forms of bias. A data set, for example, could have a topic bias or an author bias, meaning that many documents are produced by a small number of authors (Wiegand et al., 2019). Both forms impair the generalizability of a classifier trained on such a biased data set (Wiegand et al., 2019). Another form of bias that has a negative impact on the generalizability of classifiers is annotator bias Geva et al. (2019). In the context of hate speech detection, it is caused by the vagueness of the term hate speech, aggravating reliable annotations (Ross et al., 2017). Waseem (2016), for example, compared expert and amateur annotators — the latter ones are often used to label large data sets. They showed that classifiers trained on annotations from experts perform better. Binns et al. (2017) investigated whether there is a performance difference between classifiers trained on data labeled by males and females. Wojatzki et al. (2018) showed that less extreme cases of sexist speech (a form of hate speech) are differently perceived by

women and men. Al Kuwatly et al. (2020) were not able to confirm the gender bias with their experiments, but they discovered bias caused by annotators' age, educational background, and the type of their first language. Another form that is related to annotator bias is racial bias. Davidson et al. (2019) and Sap et al. (2019) examined this phenomenon and found that widely-used hate speech data sets contain a racial bias penalizing the African American English dialect. One reason is that this dialect is overrepresented in the abusive or hateful class (Davidson et al., 2019). A second reason is the insensitivity of the annotators to this dialect (Sap et al., 2019). To address the second problem, Sap et al. (2019) suggested providing annotators with information about the dialect of a document during the labeling process. This can reduce racial bias. Furthermore, Dixon et al. (2018) and Borkan et al. (2019) develop metrics to measure undesirable bias and to mitigate it. To our best knowledge, no one, however, has investigated the impact of political bias on hate speech detection so far.

### 2.2 Explainable AI

Explainable Artificial Intelligence (XAI) is a relatively new field. That is why we can find only a limited number of research applying XAI methods in hate speech detection. Wang (2018) used an XAI method from computer vision to explain predictions of a neural network-based hate speech classification model. The explanation was visualized by coloring the words depending on their relevance for the classification. Švec et al. (2018) built an explainable hate speech classifier for Slovak, which highlights the relevant part of a comment to support the moderation process. Vijayaraghavan et al. (2019) developed a multi-model classification model for hate speech that uses social-cultural features besides text. To explain the relevance of the different features, they used an attention-based approach. (Risch et al., 2020) compared different transparent and explainable models. All approaches have in common that they apply local explainability, meaning they explain not the entire model (global explanation) but single instances. We do the same because there is a lack of global explainability approaches for text classification.

## 3 Methodology

Our approach for the experiment is to train hate speech classifiers with three different politically bi-

ased data sets and then to compare the performance of these classifiers, as depicted in Figure 1. To do so, we use an existing Twitter hate speech corpus with binary labels (offensive, non-offensive), extract the offensive records, and combine them with three data sets each (politically left-wing, politically right-wing, politically neutral) implicitly labeled as non-offensive. Subsequently, classifiers are trained with these data sets and their F1 scores are compared. Additionally, we apply SHAP to explain predictions of all three models and to compare the explanations. Our code is available on GitHub<sup>1</sup>.

### 3.1 Topic Modeling

In order to answer our research questions, we need to ensure that the data sets are constructed in a fair and comparable way. Therefore, we use an existing Twitter hate speech corpus with binary labels (offensive, non-offensive) that consists of two data sets as a starting point - GermEval Shared Task on the Identification of Offensive Language 2018 (Wiegand et al., 2018) and GermEval Task 2, 2019 shared task on the identification of offensive language (Struß et al., 2019). Combining both is possible because the same annotation guidelines were applied. Thus, in effect, we are starting with one combined German Twitter hate speech data set. In the experiment, we replace only the non-offensive records of the original data set with politically biased data for each group. To ensure that the new non-offensive records with a political bias are topically comparable to the original ones, we use a topic model. The topic model itself is created based on the original non-offensive records of the corpus. Then, we use this topic model to obtain the same topic distribution in the new data set with political bias. By doing so, we assure the new data sets' homogeneity and topical comparability. The topic model has a second purpose besides assembling our versions of the data set. The keywords generated from each topic serve as the basis of the data collection process for the politically neutral new elements of the data set. More details can be found in the next subsection.

For creating the topic model, we use the Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003). A downside of LDA, however, is that it works well for longer documents (Cheng et al.,

2014; Quan et al., 2015). But our corpus consists of Tweets that have a maximum length of 280 characters. Therefore, we apply the pooling approach based on hashtags to generate larger documents, as proposed by Alvarez-Melis and Savesk (2016) and Mehrotra et al. (2013).

For finding an appropriate number of topics, we use the normalized pointwise mutual information (NPMI) as the optimization metric to measure topic coherence (Lau et al., 2014). The optimal number of topics with ten keywords each (most probable non-stop words for a topic) is calculated in a 5-fold cross-validation. Before generating the topic model, we remove all non-alphabetic characters, stop words, words shorter than three characters, and all words that appear less than five times in the corpus during the preprocessing. Additionally, we replace user names that contain political party names by the party name, remove all other user names, and apply Porter stemming to particular words<sup>2</sup> (Porter et al., 1980). Only documents (created by hashtag pooling) that contain at least five words are used for the topic modeling algorithm.

### 3.2 Data Collection

After topic modeling of the non-offensive part from the original data set (without augmentations), we collect three data sets from Twitter: one from a (radical) left-wing subnetwork, one from a (radical) right-wing subnetwork, and a politically neutral one serving as the baseline. All data was retrieved via the Twitter API. The gathering process for these three biased data sets is the following:

**1. Identifying seed profiles:** First of all, it is necessary to select for each subnetwork seed profiles that serve as the entry point to the subnetworks. For this purpose, the following six profile categories are defined that have to be covered by the selected profiles: politician, political youth organization, young politician, extremist group, profile associated with extremist groups, and ideologized news website. In the category politician, we select two profiles for each subnetwork — one female and one male. The politicians have similar positions in their parties, and their genders are balanced. For the category political youth organization, we took the official Twitter profiles from the political youth organizations of the parties that the politicians from the previous category are a member of. In the cate-

<sup>1</sup><https://github.com/mawic/political-bias-hate-speech>

<sup>2</sup>*Frauen, Männer, Linke, Rechte, Deutschland, Nazi, Jude, Flüchtling, Grüne*

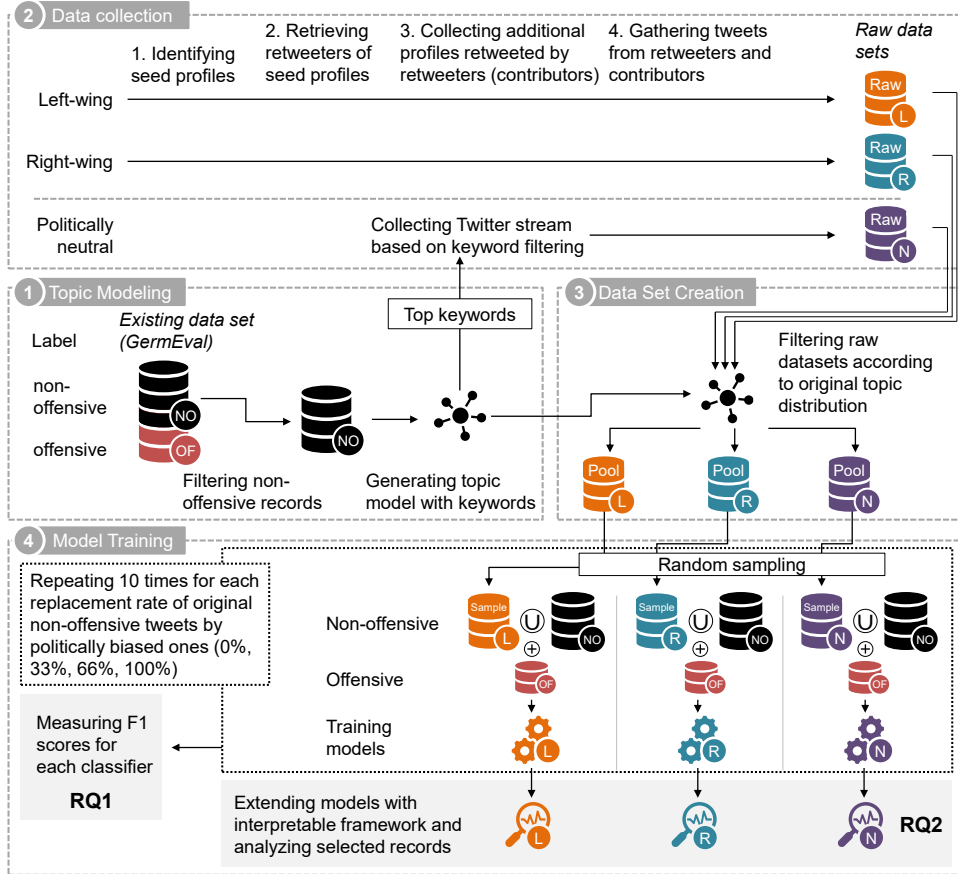


Figure 1: Methodological approach visualized

gory young politician, we selected one profile of a member from the executive board of each political youth organization. For the extremist group, we use official classifications of official security agencies to identify one account of such a group for each subnetwork. Concerning the category profile associated with extremist groups, we select two accounts that associate with an extremist group according to their statements. The statements come from the description of the Twitter account and from an interview in a newspaper. In regards to the ideologized news website, we again rely on the official classifications of a federal agency to choose the Twitter accounts of two news websites. We ensure for all categories that the numbers of followers of the corresponding Twitter accounts are comparable. The seven profiles for each subnetwork are identified based on explorative research.

**2. Retrieving retweeters of seed profiles:** After identifying the seven seed Twitter profiles for each political orientation as described in the previous paragraph, we are interested in the profiles that retweet these seed profiles. Our assumption in this context is that retweeting expresses agree-

ment concerning political ideology, as shown by [Conover et al. \(2011a\)](#), [Conover et al. \(2011b\)](#), and [Shahrezaye et al. \(2019\)](#). Therefore, the retweets of the latest 2,000 tweets from every seed profile are retrieved - or the maximum number of available tweets, if the user has not tweeted more. Unfortunately, the Twitter API provides only the latest 100 retweeters of one tweet. But this is not a problem because we do not attempt to crawl the entire subnetwork. We only want to have tweets that are representative of each subnetwork. After collecting these retweets, we select those of their authors (retweeters) that retweeted at least four of the seven seed profiles. We do this because we want to avoid adding profiles that retweeted the seed profiles but are not clearly part of the ideological subnetwork. Additionally, we remove retweeters that appear in both subnetworks to exclude left-wing accounts retweeting right-wing tweets or vice versa. Moreover, we eliminate verified profiles. The motivation of deleting verified profiles is that these profiles are ran by public persons or institutions and Twitter has proved their authenticity. This transparency might influence the language the users use for this



profile.

**3. Collecting additional profiles retweeted by retweeters (contributors):** Step 3 aims to gather the profiles (contributors) that are also retweeted by the retweeters of the seed profiles. Therefore, we retrieve the user timelines of the selected retweeters (output of step 2) to get their other retweets. From these timelines, we select those profiles that have been retweeted by at least 7.5% of the retweeters. This threshold is pragmatically chosen — in absolute numbers 7.5% means more than 33 (left-wing) and 131 (right-wing) retweeters. The reason for setting a threshold is the same one as in step 2. Besides that, profiles appearing on both sides and verified ones are also deleted.

**4. Gathering tweets from retweeters and contributors:** Additionally to the gathered user timelines from step 3, we collect the latest 2,000 tweets from the selected contributors (step 3), if they are available. Furthermore, the profiles of selected retweeters (step 2) and selected contributors (step 3) are monitored via the Twitter Stream API for a few weeks to collect additional tweets.

The politically neutral data set is collected by using the Twitter Stream API. It allows us to stream a real-time sample of tweets. To make sure to get relevant tweets, we filtered the stream by inputting the keywords from the topic model we have developed. Since the output of the Stream API is a sample of all publicly available tweets (Twitter Inc., 2020), we can assume that the gathered data is not politically biased. The result of the data collection process is a set of three raw data sets - one with a left-wing bias, one with a right-wing bias, and one politically neutral.

### 3.3 Data Set Creation

Having the topic model and the three raw data sets, we can construct the pool data sets that exhibit the same topic distribution as the original non-offensive data set. They serve as pools for non-offensive training data that the model training samples from, described in the next sub-section. Our assumption to label the politically biased tweets as non-offensive is the following: Since the tweets are available within the subnetwork, they conform to the norms of the subnetwork, meaning the tweets are no hate speech for its members. Otherwise, members of the subnetwork could have reported these tweets, leading to a deletion in case of hate speech. The availability of a tweet, however, does

not imply that they conform to the norms of the medium. A tweet that complies with the norms of the subnetwork, but violates the ones of the medium could be only distributed within the subnetwork and does not appear in the feed of other users. Consequently, it would not be reported and still be available.

We compose the pool data sets according to the following procedure for each politically biased data set: In step 1, the generated topic model assigns every tweet in the raw data sets a topic, which is the one with the highest probability. In step 2, we select so many tweets from each topic that the following conditions are satisfied: Firstly, the size of the new data is about five times the size of the non-offensive part from the GermEval corpus. Secondly, tweets with a higher topic probability are chosen with higher priority. Thirdly, the relative topic distribution of the new data set is equal to the one of the non-offensive part from the GermEval corpus. The reason for the increased size of the three new data sets (the three pool data sets) is that we have enough data to perform several iterations in the phase Model Training in order to contribute to statistical validity.

### 3.4 Model Training

In the phase Model Training, we train hate speech classifiers with the constructed data sets to compare performance differences and to measure the impact on the F1 score (RQ1). Furthermore, we make use of the ML interpretability framework SHAP to explain generated predictions and visualize differences in the models (RQ2).

Concerning the RQ1, the following procedure is applied. The basis is the original training corpus consisting of the union of the two GermEval data sets. For each political orientation, we iteratively replace the non-offensive tweets with the ones from the politically biased data sets (33%, 66%, 100%). The tweets from the politically biased data sets are labeled as non-offensive.

For each subnetwork (left-wing, right-wing, politically neutral) and each replacement rate (33%, 66%, 100%), ten data sets are generated by sampling from the non-offensive part of the original data set and the respective politically biased pool data set and leaving the offensive part of the original data set untouched. We then use these data sets to train classifiers with 3-fold cross-validation. This iterative approach produces multiple observa-

tion points, making the results more representative — for each subnetwork and each replacement rate we get  $n = 30$  F1 scores. To answer RQ1, we statistically test the hypotheses, (a) whether the F1 scores produced by the politically biased classifiers are significantly different and (b) whether the right-wing and/or left-wing classifier performs worse than the politically neutral one. If both hypotheses hold, we can conclude that political bias in training data impairs the detection of hate speech. The reason is that the politically neutral one is our baseline due to the missing political bias, while the other two have a distinct bias each. Depending on the results, we might go one step further and might infer that one political orientation diminishes hate speech classification more substantially than the other one. For this, we use the two-sided Kolmogorov-Smirnov test (Selvamuthu and Das, 2018). The null hypothesis is that the three distributions of F1 scores from three sets of classifiers are the same. The significance level is  $p < 0.01$ . If the null hypothesis is rejected, which confirms (a), we will compare the average F1 scores of each distribution with each other to answer (b).

The classifier consists of a non-pre-trained embedding layer with dimension 50, a bidirectional LSTM comprising 64 units, and one fully connected layer of the size 16. The output is a sigmoid function classifying tweets as offensive or not. We used Adam optimization with an initial learning rate of 0.001 and binary cross-entropy as a loss function. We applied padding to each tweet with a maximal token length of 30. As a post-processing step, we replaced each out-of-vocabulary token occurring in the test fold with an `<unk>` token to overcome bias and data leaking from the test data into the training data.

In regards to RQ2, we apply the following procedure. We select one classifier from each subnetwork that is trained with an entirely replaced non-offensive data set. To explain the generated predictions, we apply the DeepExplainer from the SHAP framework for each classifier (Lundberg and Lee, 2017). After feeding DeepExplainer with tweets from the original corpus ( $n = 1000$ ) to build a baseline, we can use it to explain the predictions of the classifiers. An explanation consists of SHAP values for every word. The SHAP values "attribute to each feature the change in the expected model prediction when conditioning on that feature" (Lundberg and Lee, 2017, p. 5). Comparing

the SHAP values from the three different classifiers for a selected word in a tweet indicates how relevant a word is for a prediction w.r.t. to a specific class (e.g., offensive, non-offensive). Figure 3a shows how these values are visualized. This indication, in turn, can reveal a bias in the training data. Therefore, we randomly select two tweets from the test set that are incorrectly classified by the left-wing, respectively right-wing classifier and compare their predictions to answer RQ2.

## 4 Results

### 4.1 Data

The two GermEval data sets are the basis of the experiment. In total, they contain 15,567 German tweets - 10,420 labeled as non-offensive and 5,147 as offensive. The data for the (radical) left-wing subnetwork, the (radical) right-wing one, and the neutral one was collected via the Twitter API between 29.01.2020 and 19.02.2020. We gathered 6,494,304 tweets from timelines and 2,423,593 ones from the stream for the left-wing and right-wing subnetwork. On average, 1,026 tweets ( $median = 869; \sigma^2 = 890.48$ ) are collected from 3,168 accounts. For the neutral subnetwork, we streamed 23,754,616 tweets. After removing retweets, duplicates, tweets with less than three tokens, and non-German tweets, we obtain 1,007,810 tweets for the left-wing raw data set, 1,620,492 for the right-wing raw data set, and 1,537,793 for the neutral raw data set. 52,100 tweets of each raw data set are selected for the data pools according to the topic model and the topic distribution. The input for the 3-fold cross-validation of the model training consists of the 5,147 offensive tweets from GermEval and 10,420 non-offensive ones from GermEval or the collected data depending on the replacement rate.

### 4.2 Results

All three classifiers show significantly ( $p < 0.01$ ) different F1 scores. The one with the worst performance is the one trained with the right-wing data set (78.7%), followed by the one trained with the left-wing data set (83.1%) and the politically neutral one (84.8%).

Figure 2a shows how the F1 scores change depending on the replacement rate. The lines are the average F1 scores of the three classifiers, and the areas around them are the standard deviation of the multiple training iterations. At first glance,

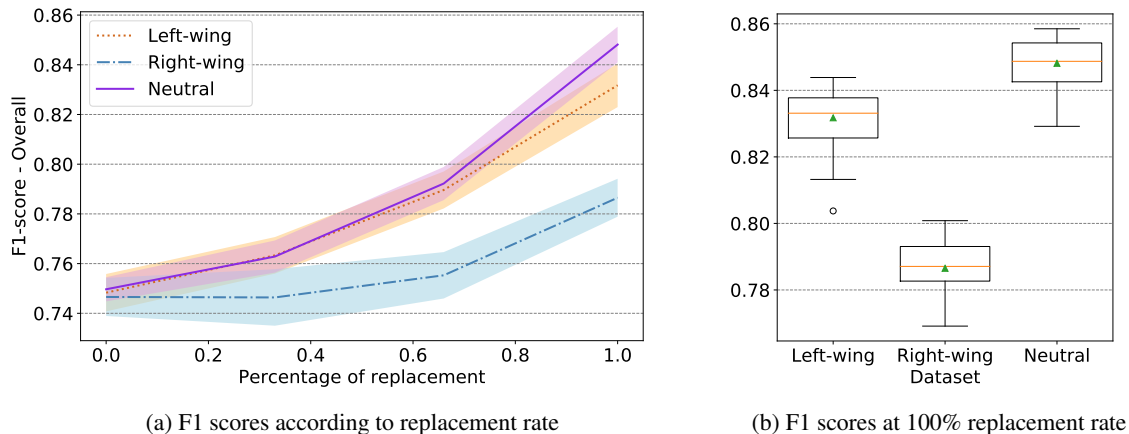


Figure 2: F1 scores of the three classifier subnetworks

the political biases in the data seem to increase the performance due to the improvement of the F1 scores. This trend, however, is misleading. The reason for the increase is that the two classes, of-fensive and non-offensive, vary strongly with the growing replacement rate, making it easier for the classifiers to distinguish between the classes. More relevant to our research question, however, are the different steepnesses of the curves and the emerg-ing gaps between them. These differences reveal that it is harder for a classifier trained with a po-litically biased data set to identify hate speech - particularly in the case of a right-wing data set. While the neutral and left-wing curves are nearly congruent and only diverge at a 100% replacement rate, the gap between these two and the right-wing curve already occurs at 33% and increases. Figure 2b visualizes the statistical distribution of the measured F1 scores at a 100% replacement rate as box plots. The Kolmogorov-Smirnov test confirms the interpretation of the charts. The distributions of the left-wing and politically neutral data set are not significantly different until 100% replacement rate — at 100%  $p = 8.25 \times 10^{-12}$ . In contrast to that, the distribution of the right-wing data set already differs from the other two at 33% replacement rate — at 33% left- and right-wing data set  $p = 2.50 \times 10^{-7}$ , right-wing and neutral data set  $p = 6.53 \times 10^{-9}$  and at 100% left- and right-wing data set  $p = 1.69 \times 10^{-17}$ , right-wing and neutral data set:  $p = 1.69 \times 10^{-17}$ . Thus, we can say that political bias in a training data set negatively im-pairs the performance of a hate speech classifier, answering RQ1.

To answer RQ2, we randomly pick two offensive tweets that were differently classified by the three

interpretable classifiers. Subsequently, we com-pare the explanations of the predictions from three different classifiers. These explanations consist of SHAP values for every token that is fed into the classifier. They indicate the relevance of the tokens for the prediction. Please note: not all words of a tweet are input for the classifier because some are removed during preprocessing (e.g., stop words). A simple way to visualize the SHAP values is de-picted in Figure 3a. The model output value is the predicted class probability of the classifier. In our case, it is the probability of how offensive a tweet is. The words to the left shown in red (left of the box with the predicted probability) are responsible for pushing the probability towards 1 (offensive), the ones to the right shown in blue (right of the box) towards 0 (non-offensive). The longer the bars above the words are, the more relevant the words are for the predictions. Words with a score lower than 0.05 are not displayed.

Figure 3a shows the result of the three inter-pretable classifiers for the following offensive tweet: @<user>@<user> *Natürlich sagen alle Gutmenschen 'Ja', weil sie wissen, dass es dazu nicht kommen wird.* (@<user>@<user> *Of course, all do-gooders say "yes", because they know that it won't happen.*)

The left-wing and neutral classifiers predict the tweet as offensive (0.54, respectively 0.53), while the right-one considers it non-offensive (0.09). The decisive factor here is the word *Gutmenschen*. *Gut-mensch* is German and describes a person "who is, or wants to be, squeaky clean with respect to morality or political correctness" (PONS, 2020). The word's SHAP value for the right-wing classi-fier is 0.09, for the left-wing one 0.45, and for the

neutral one 0.36. It is not surprising if we look at the word frequencies in the three different data sets. While the word *Gutmensch* and related ones (e.g., plural) occur 38 times in the left-wing data set and 39 times in the neutral one, we can find it 54 times in the right-wing one. Since mostly (radical) right-wing people use the term *Gutmensch* to vilify political opponents (Hanisch and Jäger, 2011; Auer, 2002), we can argue that differences between the SHAP values can indicate a political bias of a classifier.

Another example of a tweet that one politically biased classifier misclassifies is the following one (see Figure 3b): @<user>@<user> *Hätte das Volk das recht den Kanzler direkt zu wählen, wäre Merkel lange Geschichte. (If the people had the right to elect the chancellor directly, Merkel would have been history a long time ago.)*

The right-wing (0.10) and neutral classifiers (0.35) correctly classify the tweet as non-offensive, but not the left-wing one (0.96). All three have in common that the words *Volk* (German for people) and *Merkel* (last name of the German chancellor) favoring the classification as offensive, but with varying relevance. For the right-wing classifier, both terms have the lowest SHAP values (*Volk*: 0.05, *Merkel*: 0.04); for the neutral classifier, the scores are 0.34 (*Volk*) and 0.16 (*Merkel*); for the left-wing classifier, they are 0.14 (*Volk*) and 0.31 (*Merkel*). The low values of the right-wing classifier can be explained with relative high word frequency of both terms in the non-offensive training set. Another interesting aspect is that the term *Kanzler* (chancellor) increases the probability of being classified as offensive only in the case of a left-wing classifier (SHAP value: 0.08). We can trace it back to the fact that the term does not appear in the non-offensive part of the left-wing data set, causing the classifier to associate it with hate speech. This example also shows how a political bias in training data can cause misleading classifications due to a different vocabulary.

## 5 Discussion

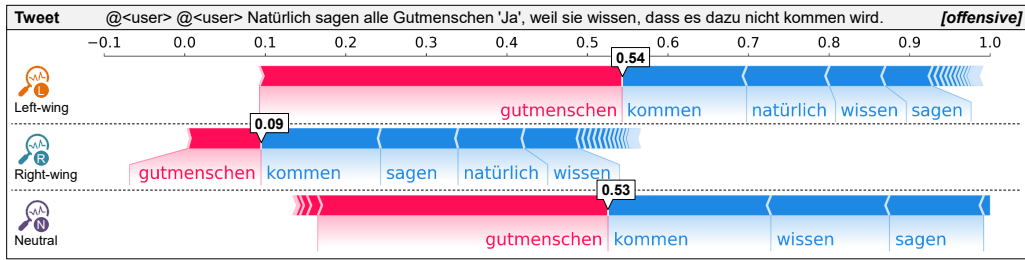
The experiment shows that the politically biased classifiers (left- and right-wing) perform worse than the politically neutral one, and consequently that political bias in training data can lead to an impairment of hate speech detection (RQ1). In this context, it is relevant to consider only the gaps between the F1 classifiers' scores at 100% replace-

ment rate. The gaps reflect the performance decrease of the politically biased classifiers. The rise of the F1 scores with an increasing replacement rate is caused by the fact that the new non-offensive tweets are less similar to the offensive ones of the original data set.

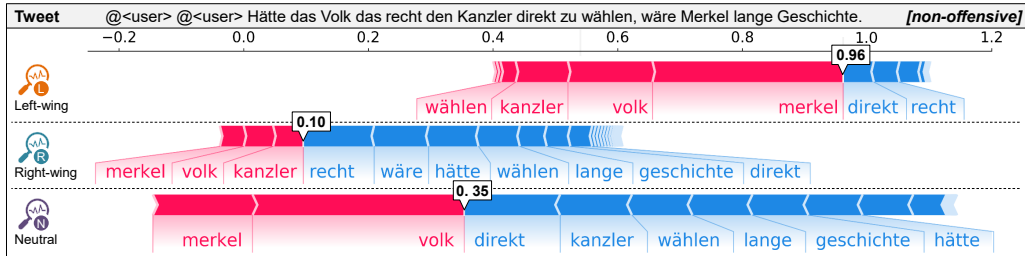
The results also indicate that a right-wing bias impairs the performance more strongly than a left-wing bias. This hypothesis, however, cannot be confirmed with the experiment because we do not have enough details about the composition of the offensive tweets. It could be that right-wing hate speech is overrepresented in the offensive part. The effect would be that the right-wing classifier has more difficulties to distinguish between offensive and non-offensive than the left-wing one even if both data sets are equally hateful. The reason is that the vocabulary of the right-wing data set is more coherent. Therefore, this hypothesis can neither be confirmed nor rejected by our experiment.

Concerning RQ2, we show that explainable ML models can help to identify and to visualize a political bias in training data. The two analyzed tweets provide interesting insights. The downside of the approach is that these frameworks (in our case SHAP) can only provide local explanations, meaning only single inputs are explained, not the entire model. It is, however, conceivable that the local explanations are applied to the entire data set, and the results are aggregated and processed in a way to identify and visualize bias. Summing up, this part of the experiment can be seen rather as a proof-of-concept and lays the foundation for future research.

Regarding the overall approach of the experiment, one may criticize that we only simulate a political bias by constructing politically biased data sets and that this does not reflect the reality. We agree that we simulate political bias within data due to the lack of such data sets. Nevertheless, we claim the relevance and validity of our results due to the following reasons: Firstly, the offensive data part is the same for all classifiers. Consequently, the varying performances are caused by non-offensive tweets with political bias. Therefore, the fact that the offensive tweets were annotated by annotators and the non-offensive tweets were indirectly labeled is less relevant. Furthermore, any issues with the offensive tweets' annotation quality do not play a role because all classifiers are trained and tested on the same offensive tweets. Secondly, we con-



(a) Tweet wrongly classified by right-wing classifier



(b) Tweet wrongly classified by left-wing classifier

Figure 3: SHAP values for the two selected tweets

struct the baseline in the same way as the left- and right-wing data set instead of using the original data set as the baseline. This compensates confounding factors (e.g., different time, authors). Thirdly, we use a sophisticated topic-modeling-based approach to construct the data sets to ensure the new data sets’ topic coherence.

## 6 Conclusion

We showed that political bias in training data can impair hate speech classification. Furthermore, we found an indication that the degree of impairment might depend on the political orientation of bias. But we were not able to confirm this. Additionally, we provide a proof-of-concept of visualizing such a bias with explainable ML models. The results can help to build unbiased data sets or to debias them. Researchers that collect hate speech to construct new data sets, for example, should be aware of this form of bias and take our findings into account in order not to favor or impair a political orientation (e.g., politically balanced set of sources). Our approach can be applied to identify bias with XAI in existing data sets or during data collection. With these insights, researchers can debias a data set by, for example, adjusting the distribution of data. Another idea that is fundamentally different from debiasing is to use these findings to build politically branded hate speech filters that are marked as those. Users of a social media platform, for example, could choose between such filters depending on their preferences. Of course, obvious hate

speech would be filtered by all classifiers. But the classifiers would treat comments in the gray area of hate speech depending on the group’s norms and values.

A limitation of this research is that we simulate the political bias and construct synthetic data sets with offensive tweets annotated by humans and non-offensive tweets that are only implicitly labeled. It would be better to have a data set annotated by different political orientations to investigate the impact of political bias. But such an annotating process is very challenging. Another limitation is that the GermEval data and our gathered data are from different periods. We, however, compensate this through our topic modeling-based data creation.

Nevertheless, political bias in hate speech data is a phenomenon that researchers should be aware of and that should be investigated further. All in all, we hope that this paper contributes helpful insights to the hate speech research and the fight against hate speech.

## Acknowledgments

This paper is based on a joined work in the context of Jan Bauer’s master’s thesis (Bauer, 2020). This research has been partially funded by a scholarship from the Hanns Seidel Foundation financed by the German Federal Ministry of Education and Research.

## References

- Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators’ demographic characteristics. In *Proc. 4th Workshop on Online Abuse and Harms*.
- David Alvarez-Melis and Martin Saveski. 2016. Topic modeling in twitter: Aggregating tweets by conversations. In *10th Intl. AAAI Conf. Weblogs and Social Media*.
- Katrin Auer. 2002. Political Correctness – Ideologischer Code, Feindbild und Stigmawort der Rechten. *Österreichische Zeitschrift für Politikwissenschaft*, 31(3):291–303.
- Jan Bauer. 2020. Political bias in hate speech classification. Master’s thesis, Technical University of Munich. Advised and supervised by Maximilian Wich and Georg Groh.
- Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *International conference on social informatics*, pages 405–415. Springer.
- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.
- Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Proc. 28th WWW Conf.*, pages 491–500.
- Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941.
- Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011a. Predicting the political alignment of twitter users. In *2011 IEEE 3rd Intl. Conf. Privacy, Security, Risk, and Trust and 2011 IEEE 3rd Intl. Conf. Social Computing*, pages 192–199.
- Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011b. Political polarization on twitter. In *5th Intl. AAAI Conf. Weblogs and Social Media*.
- Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proc. 11th ICWSM Conf*.
- Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proc. 2018 AAAI/ACM Conf. AI, Ethics, and Society*, pages 67–73.
- Maeve Duggan. 2017. *Online harassment 2017*. Pew Research Center.
- Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proc. Conf. Empirical Methods in Natural Language Processing*, pages 1161–1166.
- Astrid Hanisch and Margarete Jäger. 2011. Das Stigma ”Gutmensch”. *Duisburger Institut für Sprach-und Sozialforschung*, 22.
- Mireille Hildebrandt. 2019. Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. *Theoretical Inquiries in Law*, 20(1):83–121.
- Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.
- Scott M Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.
- Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 889–892.
- Marco Niemann. 2019. Abusiveness is non-binary: Five shades of gray in german online news-comments. In *IEEE 21st Conference Business Informatics*, pages 11–20.
- John T Nockleby. 2000. Hate speech. *Encyclopedia of the American constitution*, 3(2):1277–1279.
- PONS. 2020. [Gutmensch - Deutsch-Englisch Übersetzung — PONS](#).
- Martin F Porter et al. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.
- Xiaojun Quan, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. 2015. Short and sparse text topic modeling via self-aggregation. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

- Julian Risch, Robin Ruff, and Ralf Krestel. 2020. Offensive language detection explained. In *Proc. Workshop on Trolling, Aggression and Cyberbullying (TRAC@LREC)*, pages 137–143.
- Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proc. 57th ACL Conf.*, pages 1668–1678.
- Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proc. 5th Intl. Workshop on Natural Language Processing for Social Media*, pages 1–10.
- Dharmaraja Selvamuthu and Dipayan Das. 2018. *Introduction to statistical methods, design of experiments and statistical quality control*. Springer.
- Morteza Shahrezaye, Orestis Papakyriakopoulos, Juan Carlos Medina Serrano, and Simon Hegelich. 2019. Estimating the political orientation of twitter users in homophilic networks. In *AAAI Spring Symposium: Interpretable AI for Well-being*.
- Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language. In *Proc. 15th KONVENS*, pages 354–365.
- Andrej Švec, Matúš Pikuliak, Marián Šimko, and Mária Bielíková. 2018. Improving Moderation of Online Discussions via Interpretable Neural Models. In *Proc. 2nd Workshop on Abusive Language Online*, pages 60–65.
- Twitter Inc. 2020. Sample stream - Twitter Developers. [https://developer.twitter.com/en/docs/tweets/sample-realtime/overview/GET\\_status\\_sample](https://developer.twitter.com/en/docs/tweets/sample-realtime/overview/GET_status_sample).
- Bertie Vidgen, Rebekah Tromble, Alex Harris, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proc. 3rd Workshop on Abusive Language Online*, pages 80–93.
- Prashanth Vijayaraghavan, Hugo Larochelle, and Deb Roy. 2019. [Interpretable Multi-Modal Hate Speech Detection](#). In *Intl. Conf. Machine Learning AI for Social Good Workshop*.
- Cindy Wang. 2018. Interpreting neural network hate speech classifiers. In *Proc. 2nd Workshop on Abusive Language Online*, pages 86–92.
- Zeeraq Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proc. 1st Workshop on NLP and Computational Social Science*, pages 138–142.
- Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *NAACL-HLT 2019: Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 602–608.
- Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proc. 14th KONVENS*.
- Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2020. Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, 60(1):93–117.
- Michael Wojatzki, Tobias Horsmann, Darina Gold, and Torsten Zesch. 2018. Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments. In *Proc. 14th KONVENS*.

## A.4 STUDY VIII

©2021 IW<sub>3</sub>C<sub>2</sub> (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License<sup>3</sup>.

Maximilian Wich, Melissa Breitingner, Wienke Strathern, Marlena Naimarevic, Georg Groh, and Jürgen Pfeffer (Apr. 2021). “Are Your Friends Also Haters? Identification of Hater Networks on Social Media: Data Paper.” In: *Companion Proceedings of the Web Conference 2021*. WWW '21. Ljubljana, Slovenia: Association for Computing Machinery, pp. 481–485. ISBN: 9781450383134. DOI: [10.1145/3442442.3452310](https://doi.org/10.1145/3442442.3452310)

---

<sup>3</sup> <https://creativecommons.org/licenses/by/4.0/>



### *Publication Summary*

"Hate speech on social media platforms has become a severe issue in recent years. To cope with it, researchers have developed machine learning-based classification models. Due to the complexity of the problem, the models are far from perfect. A promising approach to improve them is to integrate social network data as additional features in the classification. Unfortunately, there is a lack of datasets containing text and social network data to investigate this phenomenon. Therefore, we develop an approach to identify and collect hater networks on Twitter that uses a pre-trained classification model to focus on hateful content. The contributions of this article are (1) an approach to identify hater networks and (2) an anonymized German offensive language dataset that comprises social network data. The dataset consists of 4,647,200 labeled tweets and a social graph with 49,353 users and 122,053 edges." (Wich, Breitinger, et al., 2021, p. 1)

### *Author Contributions*

Maximilian Wich headed the research project. He developed the initial idea as well as the concept and guided the development of the paper's methodology. Furthermore, he was one of three data annotators and wrote the paper based on the results of Melissa Breitinger's Guided Research. Melissa Breitinger developed the methodology and managed as well as implemented the data collection and analysis as the main objectives of her Guided Research under the guidance of Maximilian Wich. She also managed the annotation process and was one of three data annotators. Additionally, she provided feedback on the paper. Wienke Strathern, Marlena Naimarevic, and Jürgen Pfeffer contributed one of the datasets that was used to identify seed users. Furthermore, Wienke Strathern wrote the corresponding section of the manuscript and gave feedback, which was also provided by Jürgen Pfeffer. Georg Groh regularly discussed the ideas and concepts with the team and provided feedback on the manuscript.

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

Data Paper

Maximilian Wich  
Department of Informatics, Technical  
University of Munich  
Germany  
maximilian.wich@tum.de

Melissa Breitingner  
Department of Informatics, Technical  
University of Munich  
Germany  
melissa.breitingner@tum.de

Wienke Strathern  
Bavarian School of Public Policy  
Technical University of Munich  
Germany  
wienke.strathern@tum.de

Marlena Naimarevic  
Bavarian School of Public Policy  
Technical University of Munich  
Germany  
marlena.n@arcor.de

Georg Groh  
Department of Informatics, Technical  
University of Munich  
Germany  
grohg@in.tum.de

Jürgen Pfeffer  
Bavarian School of Public Policy  
Technical University of Munich  
Germany  
juergen.pfeffer@hfp.tum.de

## ABSTRACT

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

## CCS CONCEPTS

• **Computing methodologies** → **Language resources**; • **Human-centered computing** → **Collaborative and social computing theory, concepts and paradigms**.

## KEYWORDS

hate speech, abusive language, dataset, network analysis, machine learning, classification

### ACM Reference Format:

Maximilian Wich, Melissa Breitingner, Wienke Strathern, Marlena Naimarevic, Georg Groh, and Jürgen Pfeffer. 2021. Are Your Friends Also Haters? Identification of Hater Networks on Social Media: Data Paper. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion)*, April 19–23, 2021, Ljubljana, Slovenia. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3442442.3452310>

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

*WWW '21 Companion*, April 19–23, 2021, Ljubljana, Slovenia

© 2021 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-1-4503-8313-4/21/04.

<https://doi.org/10.1145/3442442.3452310>

## 1 INTRODUCTION

The rise of social media platforms (e.g., Facebook and Twitter) does not only have positive effects on society. A phenomenon showing the dark side of social media is the spread of hate speech [3]. Hate speech is a severe issue because it is not limited to the online world but it can also *spill over* into the offline world, e.g. by causing physical crime [18]. Consequently, the identification of hate speech is an important societal challenge.

Since the users on social media produce enormous amounts of data, it is impossible to manually monitor their content. That is why machine learning models have been developed to automatically detect hate speech. Even if the results look promising, the models have limited accuracy [2, 13].

One challenge is that hate speech is a broad and complex phenomenon and comprises various sub-types (e.g., anti-Semitism, misogyny, racism), making automatic detection difficult. One idea is to integrate additional data into the classification model besides the textual data [11, 12]. The hypothesis behind this is that characteristics about the user and its social network provide additional clues helping to detect hate speech. It is grounded on the fact that according to [8] a high portion of hateful and offensive content is produced by small subnetworks. The problem that the research community is facing here is a lack of datasets to investigate this hypothesis. There are already a lot of abusive language datasets available.

Therefore, we have two research objectives: (1) we aim to develop an approach to identify and collect hater networks on a social media platform (in our case Twitter) and (2) we aim to release the collected data (social media posts and social network data of the authors).

For this purpose, we train an offensive language detection model on a publicly available dataset. In the second step, we select a set of hateful seed users that serves as a starting point. Then, we collect their social networks depending on the offensiveness of the content by pseudo labeling the collected data with our classifier. In the fourth step, we annotate a sample of the gathered data to evaluate our approach.

Contributions:

- Approach: We provide a methodology to identify and gather hater networks on Twitter.
- Dataset: We release an offensive language dataset in German that contains 4,647,200 labeled tweets, 49,353 users and 122,053 edges of the social graph. The 4,647,200 labels are pseudo labels produced by a classification model. Furthermore, human annotators annotated 1,356 tweets for evaluation purposes (included in the dataset). To protect users' privacy, we anonymized the data and replaced all usernames with anonymous identifiers.

## 2 RELATED WORK

Researchers in the hate speech detection community have investigated the relevance of social network data [6] for hate speech classification. Chatzakou et al. [1] integrated user-based and network-based features into their classification model in addition to the textual data. They showed that the additional features improve the classification performance. But the network-based features were limited to aggregated metrics for each user (e.g., number of followers and friends), meaning that the dataset did not contain any information about relations. Other researchers [4, 5] picked up Chatzakou et al.'s [1] approach to integrating aggregated network metrics and confirmed their findings. In contrast to them, [11] used the actual edges of the follower network in form of a node2vec graph embedding to improve the hate speech classification. For this purpose, they used the dataset from Wassem and Hovy [15] and enriched it with social network data. The problem with this approach is that most of the hateful tweets in the dataset were produced by only a few users [16], meaning the network data is not representative. Ribeiro et al. [12] applied a network-centric approach to collect data and to investigate the relevance of network data for hate speech detection. They crawled a sample of Twitter's retweet network and tweets of the discovered users, starting from a seed user. Then, they annotated a sample of the data, trained a classifier using textual and network data, and evaluated the model. Unfortunately, they released only the social graph and the tweets as averaged word embeddings, making it very hard to use this dataset in other models. Their approach, however, is similar to our one - except that we consider more network types and integrate a classification model in our process to crawl the networks more targeted.

## 3 METHODOLOGY

Our approach consists of 4 phases, as depicted in Figure 1. In the first phase, we train an offensiveness classification model. In the second phase, we select the seed users whose social networks are gathered based on the content's offensiveness. Thirdly, we crawl the social networks using an offensiveness classification model to filter offensive users (haters). In the fourth phase, we manually annotate a sample of the collected tweets to evaluate our approach.

### 3.1 Training Classification Model

We need a classification model to detect offensive language in the tweets for identifying hater networks. As the basis, we use a pre-trained German BERT model [10]. In the first step, we fine-tune

the language model of the pre-trained BERT with around 4 million German tweets, which we preprocess beforehand. In the second step, we add a classification head to the model and train it to distinguish between offensive and non-offensive languages. For the training, the datasets of GermEval Shared Task on the Identification of Offensive Language 2018 [17] and GermEval Task 2, 2019 shared task on the identification of offensive language [14] are used. Since both datasets have the same labeling schema, they can be merged to one dataset. The term offense in the context of these datasets covers a wide range of aspects so that a classifier trained on this data is suitable to identify haters. It comprises "abusive language, insults, as well as merely profane statements" [17, p.2].

### 3.2 Selecting Seed Users

In the second phase, we select the seed users that serve as a starting point for the network crawling phase. In total, we select 9 seed users from different sources: (1) GermEval 2019 dataset, (2) German right-wing dataset, and (3) manual exploration of Twitter. By doing so, we ensure to have already classified haters and avoid an author bias. Due to the limitations of the Twitter API, we cannot start with a large number of seeds. Otherwise, crawling would take too long.

*GermEval 2019 Shared Task 2.* The first one is the dataset of GermEval 2019 Shared Task 2 containing 8,952 tweets labeled as offensive or non-offensive that is also used for training the classifier. We select the top 500 users that the largest amount of offensive tweets stems from. After that, we collect from these 500 users their most recent timelines via the Twitter API, limiting the number of tweets to 50. Then, all collected tweets are classified to assign each user an offensiveness score  $o_u$  that is proposed by [7] and calculated as follows:

$$o_u = \frac{1}{1 - \log \left[ \frac{\sum_{i=1}^n p_{i,c_i=1}}{\sum_{i=1}^n p_{i,c_i=1} + \sum_{i=1}^n p_{i,c_i=2}} \right]} \in [0, 1] \quad (1)$$

$p_{i,c_i=c}$ : probability of tweet  $i$  for class  $c$

Subsequently, the users with  $o_u \geq 0.5$ , i.e. offensive users or haters, are manually reviewed with respect to user activeness. Finally, 4 users from list of the most offensive and active users are selected as seed profiles.

*German right-wing dataset.* The second source is a dataset that we have collected, containing German-speaking tweets from right-wing Twitter users. Since the data is not labeled, we classify all tweets with our classifier and apply the same procedure as the one for the GermEval dataset - computing the offensiveness scores, ranking the user accordingly, investigating the user activity of the top ranked. Finally, we select the top 5 of most offensive and active users, while two of them appear already in the seed list from GermEval. The dataset itself was collected as follows: In a first step, we searched Twitter for users whose profile information included two German right wing parties. In a second step, we read about a 100 tweets to study the topics being discussed on Twitter by these two parties. Reading the tweets we filtered seven main categories to which the content could be referred to: ethnicity, nationality, sexuality, gender, religion, disability, class. Next, we manually collected Twitter account names of people who frequently took action in these discussions, i.e., actively posted and interacted. For each party

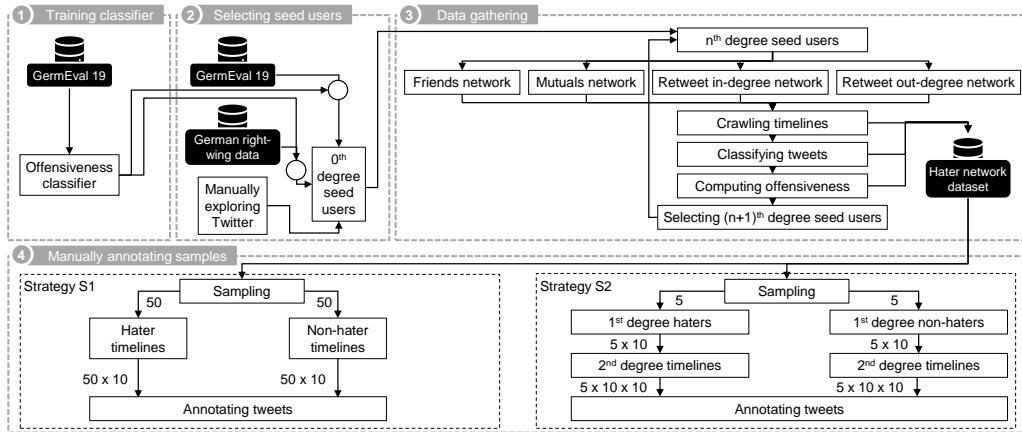


Figure 1: Methodology

we collected 500 followers. In addition to the names, we filtered also the associated profile information. Based on these information, we followed these users on Twitter and collected their posted tweets from February 22 to April 6, 2020, 45 days in total. To further understand content and language of these users, we evaluated qualitatively the top 1,000 tweets that had been re-tweeted most often. By doing so we observed that 90-95% of these tweets could be classified according to the above mentioned categories. Secondly, we took a closer look at the language being used and observed that 90% of these tweets contained offensive words. More precisely, the tweets contained clear offensive words and they were used in the context of directed aggressiveness against a group of one of the above categories. These categories are predominantly topics of hate speech. Hate speech, mostly likely, is used against a certain group or community. Regarding the time period, it should be noted that the first Corona case in Germany became known on January 28, 2020. It can be said in retrospect that the next four weeks were the media starting point of Covid19 reporting. The continuous increase of infected persons started four weeks later, February 25, 2020 – right after the start of our data collection period. From the 1,000 manually collected accounts, there was some overlap between the two right-wing party supporters. 886 accounts remained. Of these, some were no longer active, and we were ultimately able to filter out 858 users. We followed them and in total, we were able to collect about 9,000-10,000 tweets per day. The majority of the tweets (90-95%) were retweets. The data was collected on the basis of UTC-0 timezone.

*Manual exploration.* During our explorative research on Twitter, we identified 2 more hateful profiles that we add to our seed list.

### 3.3 Data Gathering

After selecting the 0<sup>th</sup> seed users, we iteratively collect their social network, as depicted in phase 3 in Figure 1. For this purpose, we use 4 different types of social relations:

- Friends network: users followed by seed user
- Mutual network: intersection of friends and followers of a seed user

- Retweet network (in-degree): retweeters of a seed user
- Retweet network (out-degree): users retweeted by a seed user

We do not consider all types of social relations that are provided by Twitter. We exclude the follower network of seed users because the follower network, in general, could be extensive. The reason is that everyone can nearly follow everyone without permission, making this kind of relationship also less meaningful. The mention network meaning one user mentions another user in a tweet is also not considered since the in-degree mention network (users mentioned a seed user) is not accessible via the standard API.

To collect the retweet network, we extract the 500 most recent tweets of a user and analyze whom they have retweeted and who has retweeted the tweets of the user. The result is a list of usernames that have a relationship to the seed users. In the next step, we gather the 100 most recent tweets from their timeline to classify them with the hate classifier and calculate the users’ offensiveness score.

Since we want to collect data from hater networks and avoid that the amount of data to collect grows exponentially, we cannot crawl the social networks of all collected users. Therefore, we have to limit this number. We do this by selecting all intersecting haters – an intersecting hater has relations to at least two seed users – and 50 other users with the highest offensiveness score  $o_u$ . Haters with a score of 1.0 are excluded because manual exploration has shown that these are either bots or users with only a few tweets. Regarding the non-hater seeds, we define a range for  $o_u$  between 0.25 and 0.5 for intersecting non-haters, aiming to choose seeds that are close to haters. A further restriction is a limit of a maximum of 1,000 followers. It aims to exclude popular profiles that interact with many non-hateful users.

These identified haters serve as seed users for the next cycle. In this paper’s scope, we apply this cycle two times, meaning that we collect the 1<sup>st</sup> and 2<sup>nd</sup> degree hater network.

### 3.4 Manual Annotation

Since a pre-trained offensiveness classification model classifies the collected tweets, we want to evaluate the classification performance

by manually annotating a sample of the data. To increase the portion of offensive and hateful content in our sample, we apply two different sampling strategies:

- S1: We randomly sample 10 tweets from 50 haters and 50 non-haters — in total 1,000 tweets.
- S2: Firstly, we randomly select 5 1<sup>st</sup> degree haters and 5 1<sup>st</sup> degree non-haters. Secondly, we sample 10 tweets from 50 users belonging to the social networks of the 1<sup>st</sup> degree haters. We also apply this for the non-haters. In total, S2 comprises 1,000 tweets.

Besides increasing the portion of offensive content, sampling the data equally from haters and non-haters helps us to test whether the haters' network contains more offensive content than the others. Applying two different sampling strategies aims to get a diverse sample from the dataset.

The sampled data are annotated by three annotators with expert knowledge in hate speech. Most of the data is annotated by two persons. The third person annotates only these tweets that received diverging annotations from the other two annotators. Since the annotators are allowed to skip a tweet and tweets containing only link(s) are ignored, some sampled tweets are not annotated and others have only one annotation instead of two or three. The inter-rater reliability of the annotators is measured with Krippendorff's Alpha [9].

## 4 RESULTS

### 4.1 Classification Model

Our fine-tuned BERT model for identifying offensive language in German tweets reached a macro F1 score of 78.6%. It is 1.5 pp better than the best model submitted to GermEval 2019 [14]. The other evaluation metrics can be found in Table 1.

**Table 1: Evaluation metrics of trained BERT classifier**

Acc.	Prec.	Recall	Micro F1	Macro F1	Weighted F1
0.821	0.753	0.654	0.82	0.786	0.817

### 4.2 Collected Data

Starting from the 9 seed users, we partially captured the 1<sup>st</sup> and 2<sup>nd</sup> degree network of these users between May 15, 2020 and August 15, 2020. Due to the size of the network and our goal to identify hater networks, we focused on offensive content and offensive users. In total, we collected 49,353 users, the mentioned social relations of these users (friends network, the intersection of follower and friends, retweet in- and out-degree network), and 4,647,200 tweets. 396 (0.8%) of the users were classified as haters ( $o_u \geq 0.5$ ) and 289,780 of the tweets (6.2%) as offensive. Further details can be found in Table 3.

Table 4 shows how many users were gathered depending on the network type and the subnetwork and how large the hater percentage was. In this context, subnetwork means a part of the collected social network. For example, "Degree 1 (H)" comprises all users that have any kind of relations to the 0<sup>th</sup> degree seed haters. Degree 2 (H and NH) refers to the subnetwork that was collected

based on the hate and non-hate seed users of degree 1. Note: Since 0<sup>th</sup> degree contains only haters, there is no Degree 1 (H).

The first finding is that the subnetworks that have only haters as seed — Degree 1 (H) and Degree 2 (H) — have for all types of networks a higher percentage of haters than the others. The second finding is that the percentage of haters also depends on the type of network. While the retweet in-degree has on average the lowest percentage, the retweet out-degree network seems to be the best network for identifying connected haters.

### 4.3 Evaluation of Classifications

To evaluate the quality of the pseudo labels that are assigned to the gathered tweets by the classifier, three annotators annotated 1,356 tweets containing 270 offensive ones. The inter-rater reliability in form of the Krippendorff's alpha is 48.9%. It is not the best one, but it is comparable to other hate speech datasets (e.g., [19] with  $\alpha = 0.45$ ). The data to be annotated was sampled by two strategies — 1,000 tweets from S1 and 1,000 tweets S2. Since annotators could skip tweets (e.g., tweets containing only URLs, missing context), S1 produced 857 annotated tweets, S2 499.

To measure the classification performance, we calculated the classification metrics between the pseudo labels provided by the classifier and our annotations. The results can be found in Table 2. The macro F1 score of the classifier on all annotated tweets (S1 and S2) is 75.3%, which is only 3.9 pp lower than on the original test set. The macro F1 score on the S2 data is only 65.9%. This could be related to the fact that dataset is smaller and more imbalanced than the S1 dataset. All in all, the classification performance on the GermEval 2019 test set (Table 1) and on the annotated test set of S1 and S2 (Table 2) are comparable.

**Table 2: Classification performance on the manually annotated test data (total and split into strategy S1 and S2)**

	S1 and S2	S1	S2
Accuracy	0.828	0.812	0.856
Precision	0.555	0.620	0.324
Recall	0.693	0.728	0.522
Micro F1	0.828	0.812	0.856
Macro F1	0.753	0.769	0.659
Weighted F1	0.835	0.817	0.870
Test data	1356	857	499
– Offensive	270	224	46
– Non-offensive	1086	633	453

## 5 DISCUSSION

We presented an approach of identifying and collecting hate networks on Twitter and showcased the utility of our approach. We found that the out-degree retweet network is the best of our four selected social relations to uncover hater networks, which partially confirms the finding from [12]. Unfortunately, we could not consider all kinds of social relations offered by Twitter due to missing endpoints. A type that is also interesting is the mention network because it reflects which users interact. In general, our approach

**Table 3: Overview of gathered data by network degree**

	Degree 0	Degree 1	Degree 2	Total	Hater/offensive
Number of users	9	14,084	35,260	49,353	396 (0.8%)
Number of tweets	700	1,367,441	3,279,059	4,647,200	289,780 (6.2%)

**Table 4: Number and percentage of classified haters by network type, network degree, and split between hater (N) and non-hater (NH) seeds**

	Degree 1 (H)		Degree 2 (H)		Degree 2 (NH)		Degree 2 (H and NH)	
	Total	Hater per.	Total	Hater per.	Total	Hater per.	Total	Hater per.
Friends	3,250	1.45%	12,423	1.57%	28,547	0.37%	36,933	0.71%
Mutuals	1,796	1.61%	6,410	1.61%	7,003	0.73%	11,581	1.11%
Retweet In-Degree	10,332	0.57%	4,062	1.08%	896	0.11%	4,590	0.98%
Retweet Out-Degree	2,419	2.77%	1,070	3.83%	4,757	0.21%	5,488	0.89%

should be applicable to other social networks that allow extracting social relations.

A point of criticism can be that our dataset mainly contains pseudo labels provided by a classification model. Firstly, it was not possible to manually annotate all 4.6 million tweets due to limited resources. Secondly, our manually annotated test data showed that the classifier provides valid and reliable classification performance to some extent because the metrics on the annotated sample are comparable to the ones on the test set. Thirdly, the focus of this paper was to provide a hate speech dataset with social network data so that other researchers can integrate this additional data into hate speech detection.

A possible improvement of our approach for future work is to work with several classifiers trained on different datasets to cover more aspects of hate speech (e.g., personal attack, sexism misogyny, anti-Semitism). Besides that, increasing the number of annotators and annotated data would also improve our findings' reliability.

## 6 CONCLUSION

We developed an approach to identifying and collecting hater networks on Twitter that applies a pre-trained classification model to focus on offensive users. We showed that our method produces the desired results. Furthermore, we collected a dataset comprising around 4,647,200 million tweets from 49,353 users (including social relations) that the research community can use to investigate social network data's relevance in hate speech detection. All tweets were pseudo-labeled, and a small sample was manually annotated. An additional finding was that the retweet out-degree network is the most appropriate network type of the investigated networks to detect hater networks.

## RESOURCES

Our code is available under <https://github.com/mawic/hater-network-identification>. Concerning the data, please contact us via e-mail or <https://in.tum.de/social/team/maximilian-wich/>.

## REFERENCES

- [1] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proc. 2017 ACM on Web Science Conference*.

- [2] Thomas Davidson, Dana Warmesley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proc. 11th ICWSM Conf.*
- [3] Maeve Duggan. 2017. *Online harassment 2017*. Pew Research Center.
- [4] Elise Fehn Unsvåg and Björn Gambäck. 2018. The Effects of User Features on Twitter Hate Speech Detection. In *Proc. 2nd Workshop on Abusive Language Online*. 75–85.
- [5] Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proc. 10th ACM Conference on Web Science*. 105–114.
- [6] Marina Hennig, Ulrik Brandes, Jürgen Pfeffer, and Ines Mergel. 2012. *Studying Social Networks. A Guide to Empirical Research*. Campus Verlag.
- [7] Linda Jahn. 2020. *Leveraging Social Network Data for Hate Speech Detection*. Master's thesis. Technical University of Munich. advised and supervised by Maximilian Wich and Georg Groh.
- [8] Philip Kreißel, Julia Ebner, Alexander Urban, and Jakob Guhl. 2018. Hass auf Knopfdruck. Rechtsextreme Trollfabriken und das Ökosystem koordinierter Hasskampagnen im Netz. *Institute for Strategic Dialogue* (2018).
- [9] K. Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Sage.
- [10] MDZ Digital Library. 2020. dbmdz BERT models. <https://github.com/dbmdz/berts> (accessed on 22.4.2020).
- [11] Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author Profiling for Abuse Detection. In *Proc. 27th International Conference on Computational Linguistics*. 1088–1098.
- [12] Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgilio A. F. Almeida, and Wagner Meira. 2018. Characterizing and Detecting Hateful Users on Twitter. *arXiv preprint arXiv:1803.08977* (2018).
- [13] Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proc. 5th Intl. Workshop on Natural Language Processing for Social Media*. 1–10.
- [14] Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval Task 2, 2019 shared task on the identification of offensive language. In *Proc. 15th KONVENS*. 354–365.
- [15] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proc. NAACL student research workshop*. 88–93.
- [16] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 602–608.
- [17] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proc. 14th KONVENS*.
- [18] Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2020. Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. *The British Journal of Criminology* 60, 1 (2020), 93–117.
- [19] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proc. 26th International Conference on World Wide Web*. 1391–1399.

## A.5 STUDY X

©2021 Springer Nature Switzerland AG

Reprinted with permission from:

Maximilian Wich, Edoardo Mosca, Adrian Gorniak, Johannes Hingerl, and Georg Groh (Sept. 2021). “Explainable Abusive Language Classification Leveraging User and Network Data.” In: *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*. Ed. by Yuxiao Dong, Nicolas Kourtellis, Barbara Hammer, and Jose A. Lozano. Cham: Springer International Publishing, pp. 481–496. ISBN: 978-3-030-86517-7. DOI: [10.1007/978-3-030-86517-7\\_30](https://doi.org/10.1007/978-3-030-86517-7_30)

This thesis includes the accepted version of publication and not the final published version.

### *Publication Summary*

"Online hate speech is a phenomenon with considerable consequences for our society. Its automatic detection using machine learning is a promising approach to contain its spread. However, classifying abusive language with a model that purely relies on text data is limited in performance due to the complexity and diversity of speech (e.g., irony, sarcasm). Moreover, studies have shown that a significant amount of hate on social media platforms stems from online hate communities. Therefore, we develop an abusive language detection model leveraging user and network data to improve the classification performance. We integrate the explainable AI framework SHAP (SHapley Additive exPlanations) to alleviate the general issue of missing transparency associated with deep learning models, allowing us to assess the model's vulnerability toward bias and systematic discrimination reliably. Furthermore, we evaluate our multimodal architecture on three datasets in two languages (i.e., English and German). Our results show that user-specific timeline and network data can improve the classification, while the additional explanations resulting from SHAP make the predictions of the model interpretable to humans." (Wich, Mosca, et al., 2021, p. 1)

### *Author Contributions*

Maximilian Wich headed the research project. He developed the initial idea, the concept, and the methodology of the study. Additionally, he wrote most of the paper. Edoardo Mosca implemented the first version of the explainable multimodal model as part of his master's thesis supervised by Maximilian Wich. Additionally, he provided extensive feedback during the writing process of the manuscript. Adrian Gorniak and Johannes Hingerl implemented an enhanced version of the explainable multimodal model during their NLP Lab Course project supervised by Maximilian Wich. They also provided feedback on the manuscript. Georg Groh regularly discussed the ideas and concepts with the team and provided feedback on the manuscript.



# Explainable Abusive Language Classification Leveraging User and Network Data

Maximilian Wich<sup>[0000-0002-9149-9454]</sup> ✉, Edoardo Mosca<sup>[0000-0003-4045-5328]</sup>,  
Adrian Gorniak<sup>[0000-0002-6165-5807]</sup>, Johannes Hingerl<sup>[0000-0002-8260-032X]</sup>, and  
Georg Groh<sup>[0000-0002-5942-2297]</sup>

Technical University of Munich, Germany  
{maximilian.wich, edoardo.mosca, adrian.gorniak, johannes.hingerl}@tum.de  
grohg@in.tum.de

**Abstract.** Online hate speech is a phenomenon with considerable consequences for our society. Its automatic detection using machine learning is a promising approach to contain its spread. However, classifying abusive language with a model that purely relies on text data is limited in performance due to the complexity and diversity of speech (e.g., irony, sarcasm). Moreover, studies have shown that a significant amount of hate on social media platforms stems from online hate communities. Therefore, we develop an abusive language detection model leveraging user and network data to improve the classification performance. We integrate the explainable AI framework SHAP (SHapley Additive exPlanations) to alleviate the general issue of missing transparency associated with deep learning models, allowing us to assess the model’s vulnerability toward bias and systematic discrimination reliably. Furthermore, we evaluate our multimodel architecture on three datasets in two languages (i.e., English and German). Our results show that user-specific timeline and network data can improve the classification, while the additional explanations resulting from SHAP make the predictions of the model interpretable to humans.

**Keywords:** Hate speech · Abusive language · Classification model · Social network · Deep learning · Explainable AI

**Warning:** This paper contains content that may be abusive or offensive.

## 1 Introduction

Hate speech is a severe challenge that social media platforms such as Twitter and Facebook face nowadays. However, it is not purely an online phenomenon and can spill over to the offline world resulting in physical violence [36]. The Capitol riots in the US at the beginning of the year are a tragic yet prime example. Therefore, the fight against hate speech is a crucial societal challenge.

The enormous amount of user-generated content excludes manual monitoring as a viable solution. Hence, automatic detection of hate speech becomes the key component of this challenge. A technology to facilitate the identification

is *Machine Learning*. Especially in recent years, *Natural Language Processing* (NLP) has made significant progress. Even if these advances also enhanced hate speech classification models, there is room for improvement [29].

However, gaining the last points of the F1 score is a massive challenge in the context of hate speech. Firstly, abusive language has various forms, types, and targets [32]. Secondly, language itself is a complex and evolving construct; e.g., a word can have multiple meanings, people create new words or use them differently [29]. This complexity exacerbates classifying abusive language purely based on textual data. Therefore, researchers have started to look beyond pure text-driven classification and discovered the relevance of social network data [10]. Kreißel et al. [11], for example, showed that small subnetworks cause a significant portion of offensive and hateful content on social media platforms. Thus, it is beneficial to integrate network data into the model [3, 22, 15, 5, 6]. However, to the best of our knowledge, no one has investigated the impact of combining the text data of the post that is meant to be classified, the user’s previous posts, and their social network data.

An issue with such an approach is its vulnerability to bias, meaning that a system ”systematically and unfairly discriminate[s] against certain individuals or groups of individuals in favor of others” [7, p. 332]. *Deep Learning* (DL) models often used in NLP are particularly prone to this issue because of their black-box nature [17]. Conversely, a system combining various data sources and leveraging user-related data has a more considerable potential of discriminating individuals or groups. Consequently, such systems should integrate *eXplainable AI* (XAI) techniques to address this issue and increase trustworthiness.

We address the following two research questions in our paper concerning the two discussed aspects:

**RQ1** Can abusive language classification be improved by leveraging users’ previous posts and their social network data?

**RQ2** Can explainable AI be used to make predictions of a multimodal hate speech classification model more understandable?

To answer the research questions, we develop an explainable multimodal classification model for abusive language using the mentioned data sources<sup>1</sup>. We evaluate our model on three different datasets—WASEEM [33], DAVIDSON [4], and WICH [35]. Furthermore, we report findings of integrating user and social network data that are relevant for future work.

## 2 Related Work

Most work in the abusive language detection domain has focused on developing models that only use the text data of the document to be classified [29, 16, 24]. Other works, however, have started to integrate context-related data into abusive language detection [29, 24, 18]. One promising data source is the users’

<sup>1</sup> Code available on <https://github.com/mawic/multimodal-abusive-language-detection>

social network because it has been shown that hater networks on social media platforms cause a considerable amount of online hate [11, 8]. Combining network and text data from Twitter was already successfully applied to predict whether an account is verified [2] or to identify extremist accounts [38]. In the case of abusive language, Papegnies et al. [19] built a classification model using local and global topological measures from graphs as features for cyberbullying detection (e.g., average distance, betweenness centrality). A similar approach has been applied by Chatzakou et al. [3], but they also integrated user-related data (e.g., number of posts, account age) and textual data (e.g., number of hashtags). This approach was picked up and extended by other researchers [6, 5] (e.g., integrating users’ gender, geolocation) who confirmed the usefulness of additional context-related data sources. They all have in common that the network features are only topological measures and do not contain any information about the relations. Mishra et al. [15] addressed this downside and modeled the users’ follower network with a node2vec embedding that serves as an additional input for the classification model. Ribeiro et al. [22] developed a similar model; they, however, used the graph embedding GraphSAGE to model the retweet network and combined it with a document embedding for the text data [9]. For this purpose, they collected a dataset that has a fully connected network. Unfortunately, they released only the network data and the document embeddings but not the raw text. Recently, Li et al. [12] refined this approach.

Another data source that supports abusive language detection is the user’s history of previous posts. Qian et al. [20] improved a hate speech classifier for tweets by adding the previous tweets of the author. Raisi and Huang [21] proposed a model that leverages the user’s history of posts and the post directed to the user to calculate a bully and victim score for each user. However, to the best of our knowledge, no one has integrated user’s previous posts and social networks into abusive language detection.

Besides multimodality, XAI in abusive language detection is another topic that we have to consider in this section. Since XAI is a relatively new field, it has not been frequently applied to abusive language detection with some exceptions [14, 34, 31, 27, 30, 18]. All models use only the text as input, except [30]. Their model also relies on network data. But the network submodel is very simple; it is only a binary vector encoding whether the user follows pre-defined hater accounts. Furthermore, the explanations for this submodel are not detailed. Hence, the explainable model that we propose is an advancement.

### 3 Data

For our experiment, we use three abusive language datasets that are from Twitter. Table 1 provides an overview of the datasets’ characteristics. Figure 1 visualizes the social network graph of the datasets.

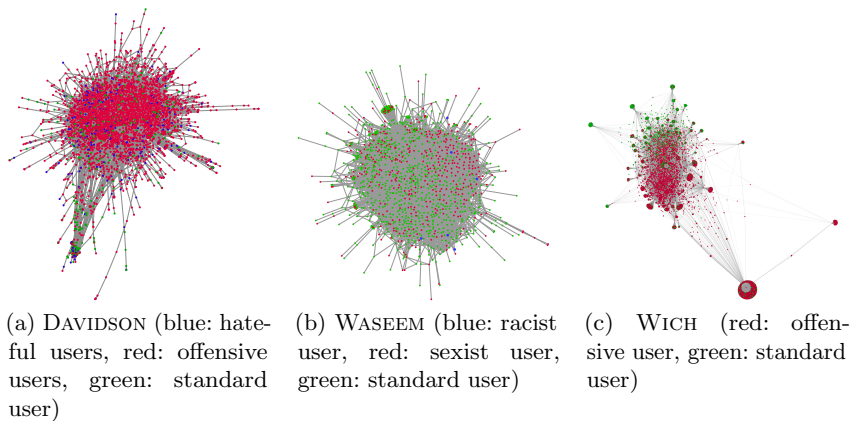
**DAVIDSON** Davidson et al. [4] released an English abusive language dataset containing 24,783 tweets annotated as hate, offensive, or neither. Unfortunately,

**Table 1.** Overview of the datasets’ statistics

	Davidson			Waseem			Wich	
Number of tweets	14,939			16,907			68,443	
Number of users	6,725			2,024			939	
Avg. number of tweets per user	2.22			8.35			72.9	
Class distribution	hate	offensive	neither	sexism	racsim	none	offensive	non-offensive
	814	11,800	2,325	3,430	1,976	11,501	26,205	42,238
Network: avg. degree	1.85			3.44			1.63	
Network: graph density	0.0005			0.0034			0.0002	

the dataset does not contain any data about the user or the network. Therefore, we used the Twitter API to get the original tweets and the related user and network data. Since not all tweets are still available on Twitter, our dataset has shrunk to 14,939 tweets.

WASEEM Waseem et al. [33] published an English abusive language dataset containing 16,907 tweets annotated as sexist, racist, or none. Similar to DAVIDSON, the dataset does not provide any user- or network-related data. The authors of [15] shared their enriched WASEEM dataset with us containing the user and network data.



**Fig. 1.** Visual comparison of the network topologies. Standalone nodes or very small subnetworks that do not connect to the main graph for DAVIDSON and WASEEM are excluded.

WICH Wich et al. [35] released a German offensive language dataset containing 4,647,200 tweets annotated as offensive or non-offensive. Most of the tweets are pseudo-labeled with a BERT-based classifier; a smaller portion of the dataset

is also manually annotated. The difference between this dataset and the other two is the way it was collected. Wich et al. applied a snowball sampling strategy focusing on users. Starting from seed users, the authors collected the connected users and their tweets based on their offensiveness. Hence, the network graph has a star-shaped network topology contrary to the other two, as depicted in Figure 1c. We select only 68,443 tweets and the related user and network information to better handle the data. The manually annotated tweets are used as a test set.

## 4 Methodology

The section is split into two subsections. The first one deals with the model architecture and training of the multimodal classification model. The second one considers the XAI technique that we use to explain the predictions of our multimodal model.

### 4.1 Multimodal Classification Model

**Architecture** The multimodal classification model for abusive language consists of three submodels that process the different inputs:

1. **Text model:** It processes the text data of the tweet that is meant to be classified. For this purpose, we use DistilBERT with a classification head.
2. **History model:** It processes the tweet history of the user.
3. **Network model:** It processes the social network data of the tweet’s user. To model the network data, we use the vector embedding framework GraphSAGE.

The three models’ outputs are combined in a linear layer, which outputs the prediction for the tweet to be classified.

*Text model* The text data of the tweet is fed into a pre-trained DistilBERT model with a classification head. DistilBERT is a lighter and faster version of the transformer-based model BERT [23]. Despite the parameter reduction, its performance is comparable to BERT in general [23] and in the context of abusive language detection [28]. In order to implement the model, we use the Transformers library from Hugging Face<sup>2</sup> and its implementation of DistilBERT [37]. As pre-trained models, we use `distilbert-base-uncased` for the English datasets and `distilbert-base-german-cased` for the German one. Before tokenizing the text data, we remove username mentions from the tweets, but we keep the ”@” from the mention<sup>3</sup>. The purpose of this procedure is to avoid the classifier memorizing the username and associating it with one of the classes. But the classifier should recognize that the tweet addresses another user.

<sup>2</sup> <https://huggingface.co/transformers/>

<sup>3</sup> If a user is mentioned in a tweet, an ”@” symbol appears before the user name.

*History model* We use a bag-of-words model to model the user’s tweet history, comprising the 500 most common terms from the dataset based on term frequency-inverse document frequency (tf-idf). For each user, it is a 500-dimensional binary vector that reflects which of the most common terms appear in the user’s tweet history.

*Network model* In order to model the user’s social network, we apply the inductive representation learning framework GraphSAGE [9]. The advantage of an inductive learning framework is that it can be applied to previously unseen data, meaning the model can generate an embedding for a new user in a network, which is a desirable property for our use case. Our GraphSAGE model is trained on the undirected network graph of the social relations. Furthermore, we assign to each user/node a class derived from the labels of their tweets. The output of the model is a 32-dimensional graph embedding for each user. The graphs are modeled as follows:

- DAVIDSON: An edge between two users exists if at least one follows the other. A user is labeled as hater, if he or she has at least one hate tweet; as offensive, if he or she has at least one offensive tweet, but no hate tweet; as neither, if he or she has only neither tweets.
- WASEEM: An edge between two users exists if at least one follows the other. A user is labeled as racist, if he or she has at least one tweet labeled as racist; same for sexist; as none, if he or she is neither racist nor sexist.
- WICH: An edge between two users exists if at least one has retweeted the other. A user is labeled as offensive, if he or she has at least three offensive tweets.

Users without network connections in their respective dataset, so-called solitary users, do not receive a GraphSAGE embedding; their embedding vector only contains zeros.

The output of the three models is concatenated to a 534 or 535 respectively dimensional vector (DistilBERT: 2 or 3 dimensions depending on the output speech classes; GraphSAGE: 32 dimensions; bag-of-words: 500 dimensions) and fed into a hidden linear layer. This final layer with softmax activation reduces the output to the number of classes according to the selected dataset.

**Training** Several challenges have to be faced when it comes to training the model. In terms of sampling, we cannot randomly split the dataset: We have to ensure that tweets of any user do not appear in the train and test set; otherwise, we would have a data leakage. Therefore, sampling is done on the user level. Users are categorized into groups based on their class and the existence of a network. We gather six different categories for WASEEM and DAVIDSON and four categories for WICH. The train, validation, and test set all contain users from different classes by sampling these categories to prevent bias toward certain user groups. Due to the different tweet counts per user, the train set size varies between 60-70% depending on the dataset.

We under- and oversample the classes during training since all datasets are unbalanced. Moreover, we have to train the three submodels separately because the unsupervised training process of GraphSAGE cannot be combined with the supervised training of DistilBERT. DistilBERT is fine-tuned for two epochs with a batch size of 64 and an Adam optimizer (initial learning rate of  $5 \times 10^{-5}$  and a weight decay of 0.01). We train our GraphSAGE model, consisting of three hidden layers with 32 channels each, for 50 epochs with an Adam optimizer (initial learning rate of  $5 \times 10^{-3}$ ). The bag-of-words model does not require training. After training the submodels, we freeze them and train the hidden layer (10 epochs; Adam optimizer with an initial learning rate of  $1 \times 10^{-3}$ ).

## 4.2 Explainable AI Technique

We set model interpretability as a core objective of our work. To this end, we produce Shapley-values-based explanations at different levels of granularity. Shapley values are an established technique to estimate the contribution of input features w.r.t. the model’s output [25, 13]. Their suitability for this task has been proven both on a theoretical as well as on an empirical level [13].

As computing exact Shapley values is exponentially complex w.r.t. the input size and hence not feasible, accurate approximations are fundamental for their estimation [13]. As shown in Algorithm 1, we compute them by iteratively averaging each feature’s marginal contribution to a specific output class. We find that 15 iterations are sufficient for Shapley values to converge. A random sampling of features was used for reasons of simplicity. Finally, we can assign each feature a Shapley value, representing its relative impact score. A similar approximation approach has been used in [26].

There are two different granularity levels in terms of features: For instance, we can treat each model component (tweet, network, history) as a single feature and derive impact scores (Shapley values) for these components. Alternatively, each model component input or feature (e.g., each token of a tweet) can be treated separately on a more fine-grained level. As Shapley values are additive, they can be aggregated to represent component-level Shapley values. The way feature and components are excluded in order to compute their respective Shapley value changes based on these two levels listed in Table 2. Thus, our multimodal model can be explained on a single instance, and the role played by each model can always be retrieved.

Additionally, we partition the network graph into communities using the Louvain algorithm to derive Shapley values for individual network connections [1]. All user edges in that community with the target user are disabled to obtain the impact of a specific community, resulting in a new GraphSAGE generated user embedding as input for the multimodal model. The embedding vectors of solitary users that only contain zeros result in Shapley values equal to zero for the network component of all these users.

**Result:** Shapley value  $\{\phi_t\}_{t=1}^M$  for every feature  $\{x_t\}_{t=1}^M$   
**Input:**  $p$  sample probability,  $x$  instance,  $f$  model,  $I$  number of iterations  
**for**  $i = 0, \dots, I$  **do**  
  **for**  $t = 1, \dots, M$  **do**  
    sample a Bernoulli vector  $P = \{0, 1\}^M$  with probability  $p$   
    pick  $S$  a subset of the features  $\{x_t\}_{t=1}^M \setminus \{x_t\}$  according to  $P$   
    build  $x_S$  alteration of  $x$  with only features in  $S$   
     $\phi_t \leftarrow \phi_t \frac{i-1}{i} + \frac{f(x_{S \cup \{x_t\}}) - f(x_S)}{i}$   
  **end**  
**end**

**Algorithm 1:** Shapley value approximation algorithm. In our experiments,  $p = 0.7$  and  $I = 15$  were used as parameters.

**Table 2.** Masking strategies for SHAP on component and feature level

	Text	Network	History
<b>Component wise</b>	Masking BERT output with 0s	Setting GraphSAGE embedding to 0	Setting all vocabulary counts to 0
<b>Feature wise</b>	Masking each token individually	Disabling edges to user based on community and generating new embedding	Setting each vocabulary token count to 0 individually

## 5 Results

In the first subsection, we deal with answering RQ1 based on the classification performance of our architecture. The second subsection addresses the explainability of the models and related findings to answer RQ2.

**Table 3.** Classification models’ performance by different architectures and datasets

Model	Davidson			Waseem			Wich		
	P	R	F1	P	R	F1	P	R	F1
Text	75.3	77.1	76.1	77.5	84.1	80.3	89.8	91.7	90.7
Text + History	73.7	77.8	75.5	79.3	87.8	<b>82.7</b>	89.8	91.7	90.7
Text + Network	75.3	77.2	76.2	77.5	84.4	80.4	89.9	91.7	<b>90.8</b>
All	74.5	78.9	<b>76.5</b>	79.2	88.1	<b>82.7</b>	90.0	91.7	<b>90.8</b>

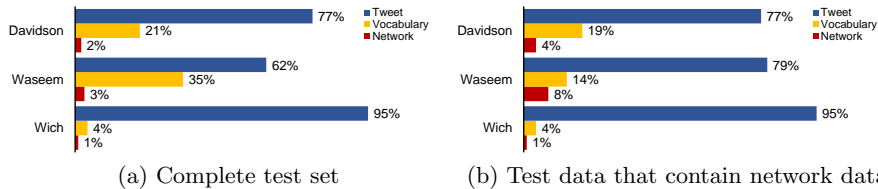
### 5.1 Classification Performance

Table 3 displays the different model architecture performance metrics for the three datasets. We find that combining text, history, and network increases the macro F1 score of WASEEM by 2.4 pp and of DAVIDSON by 0.4 pp. In the case of WICH, we observe only a minor increase of the precision by 0.1 pp. We ascribe these diverging increases to two aspects: Firstly, the network of WASEEM is the



densest one of all three, followed by DAVIDSON and WICH, as depicted in Table 1. Secondly, WICH’s text model has a high F1 score, meaning that this submodel presumably drives the predictions of the multimodal model. Our impact analysis using SHAP to identify each submodel’s relevance confirms this hypothesis, as depicted in Figure 2. It shows that the network and history data are less relevant for WICH’s multimodal model than for the other two models.

In order to answer RQ1, these results signify that leveraging a user’s previous posts and their social network data does improve abusive language classification. Additionally, the improvement of the F1 score is proportional to the network’s density – the higher the density, the higher the improvement.



**Fig. 2.** Avg. impact of each classifier’s submodels on the respective test set based on Shapley values

## 5.2 Explainability

In this subsection, we present the results of the XAI technique, SHAP, that we applied to our multimodal model. Firstly, we further investigate the impact of the network and history data added to the text model. Secondly, we show the explanation of a single tweet.

**Impact Analysis of the Submodels** Figure 2 visualizes the impact of the submodels on the multimodal model. We calculate the impact by aggregating the Shapley values for each submodel based on the tweets in the test set. Figure 2a displays the impact on the complete test set of each dataset, while Figure 2b shows the impact on test data that contains network data<sup>4</sup>.

Our first observation is that all classifiers are mainly driven by the text model, followed by the history and network model. Comparing Figure 2a and 2b, we see that network data, if available, contributes to the predictions of WASEEM’s and DAVIDSON’s multimodal models. If we compare the network model’s impact of both datasets in the context of network density (DAVIDSON:  $5 \times 10^{-4}$ ; WASEEM:  $3.4 \times 10^{-3}$ ), we can conclude that the denser the network is, the more relevant it is for the classification. These findings confirm our answer to RQ1.

In the case of WASEEM, we observe a large contribution of the history model (35%) for the complete test set. We can trace it back to four users that produced a

<sup>4</sup> Network data is not available for all users.

large portion of the dataset and mainly produced all abusive tweets. In general, the number of tweets in the user’s history correlates positively with the Shapley value for the history model, reflecting the impact of the history model on the prediction. While the correlation within WICH’s dataset is only weak ( $r_{Wich} = 0.172$ ), we observe a moderate correlation for the other two datasets ( $r_{Davidson} = 0.500$  and  $r_{Waseem} = 0.501$ ).

Regarding WICH’s dataset, the Shapley values indicate that the text model dominates (95%) the multimodal model’s prediction, while the other two (4% and 1%) play only a minor role. There are two reasons for this: First, the tweets are pseudo-labeled by a BERT model. Since we use a DistilBERT model similar to BERT, we achieve an outstanding F1 score of the text model (90.7%). The downside of such a good classification performance is that the multimodal model relies mainly on the text model’s output. Therefore, the history and network model are less relevant. Furthermore, the dataset’s network is characterized by a low degree of interconnectivity compared to the networks of the other two datasets (cf. Table 1).

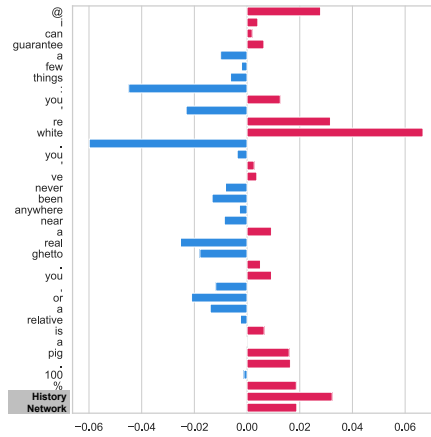
We established that aggregating the Shapley values of the test set with respect to RQ2 helps us better understand the relevance of each submodel. The insights gained by the applied XAI technique also confirmed our answer to RQ1 that user’s network and history data contribute to abusive language detection.

**Explaining a Single Tweet Classification** After investigating the model on an aggregated level, we focus on explaining the prediction of a single tweet. To do so, we select the following tweet from the DAVIDSON dataset that is labeled and correctly predicted as hateful by our multimodal model:

*@user i can guarantee a few things: you’re white. you’ve never been anywhere NEAR a real ghetto. you, or a relative is a pig. 100%*

In the following, we demonstrate the explainable capabilities of our multimodal model based on the selected tweet. Figure 3 plots the Shapley values of the tweet’s tokens and the user’s history and network (last two rows). These Shapley values indicate the relevance of the feature on the multimodal model’s prediction as hateful. A positive value (red-colored) represents a contribution favoring the classification as hateful, a negative value (blue-colored) that favors the classification as non-hateful.

We see that the most relevant word for the classification as hateful is ”white”, which should not be surprising because of the racist context. Furthermore, the @-symbol (representing a user mention) and ”you(’)re” are relevant for the classification model, indicating that directly addressing someone is recognized as a sign of hate for the classifier. In contrast, the punctuation of the tweet negatively influences the classification as hateful. A possible explanation is that correct spelling and punctuation are often disregarded in the context of abusive language. Beyond the textual perspective, we observe that the history and network submodels favor the classification as hateful. These inputs are relevant for our multimodal model to classify the tweet correctly. Considering Figure 4a (an

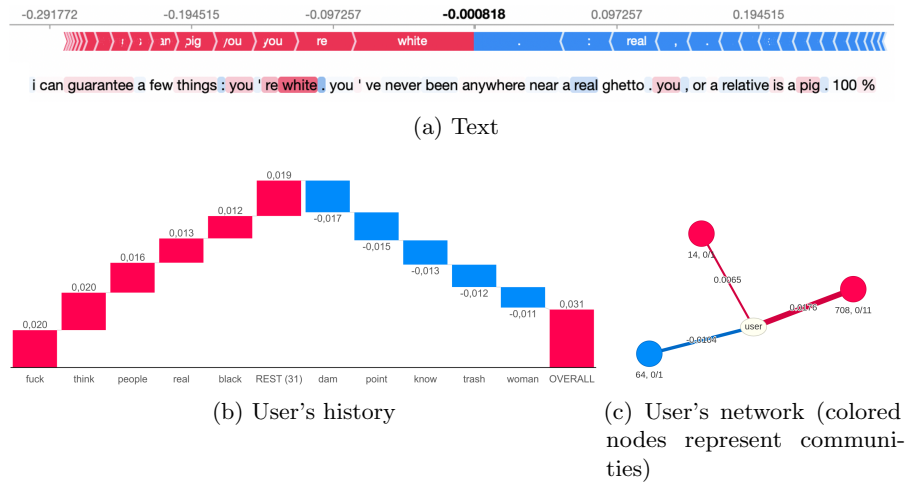


**Fig. 3.** Relevance of the different features in the form of Shapely values; positive, red values represent favoring a classification as hateful; negative, blue ones the opposite; Shapely values for history and network submodel are aggregated

alternative visualization of the Shapely values), we see that the text model slightly favors the classification as non-hateful, represented by the negative sum of Shapely values. Due to the input from the other two submodel, however, the multimodal model classifies the tweet correctly, making this an excellent example of how abusive language detection can profit from additional data.

Figures 4b and 4c break down the contribution of the history and network model, where Figure 4b is a waterfall chart displaying the most relevant terms that the user used in their previous posts—less relevant terms are summarized in the column named REST. As in the previous charts, red represents a positive contribution to the classification as hateful and blue vice versa. The last column, called OVERALL, is the sum of all terms’ Shapely values. In this case, the previous tweets of the user contain words that are primarily associated with hateful tweets; consequently, the history model favors a classification as hateful. Figure 4c shows the user’s ego network and its impact on the classification. The nodes connected to the user represent communities identified by the Louvain algorithm. The first number of a node’s label is an identifier; the second number is the number of haters in the community; the third number is the community’s total number of users. The color of the nodes and edges have the same meaning as in the other visualizations. In our case, two connected communities contribute to a hateful classification, while the left-pointing community counteracts this.

The presented explanations of the complete model and its submodels provide meaningful and reasonable information to understand better how the model decides to make predictions. These findings extend our answer to RQ2 from the previous section. Our explainable model provides explanations on an aggregated level and a single prediction level to make the classification more understandable.



**Fig. 4.** Explanations for predictions of test, history, and network submodel in the form of Shapely values (red, positive values favor a classification as hateful; blue, negative values favor a classification as non-hateful)

## 6 Discussion

We demonstrated that leveraging a user's history and ego network can improve abusive language detection regarding RQ1, consistent with the findings from other researchers [15, 22, 20]. Our multimodal approach is novel because we combine text, users' previous tweets, and their social relations in one model. The additional data sources provide further indications for the classification model to detect abusive language better. That can be helpful, especially when the classifier struggles with a precise prediction, as in our example in Section 5.2. Other examples are implicit language, irony, or sarcasm, which are hard to detect from a textual perspective. The improvement, however, varies between the datasets. We trace this back to the network density of the available data. WASEEM has the network with the highest density and exhibits the best improvement if we integrate history and network data. In contrast, the classification model based on WICH, the dataset with the least dense network, could be improved only slightly. A further difficulty concerning WICH's dataset is that the tweets are pseudo-labeled with a BERT model, and our text submodel uses DistilBERT. Hence, our text submodel performs so well that the multimodal model nearly ignores the outputs of the history and network submodels. Therefore, it was hard to identify any improvement. Relating to DAVIDSON, we had the problem of data degradation. Since the dataset does not contain any user or network data, we used the Twitter API to obtain them. But not all tweets were still available, causing us to use only 60% of the original dataset for our experiment. We require more appropriate datasets to investigate the integration of additional data sources in abusive language detection and refine this approach. For example,

Riberio et al. [22] have released a comprehensive dataset containing 4,972 labeled users. Unfortunately, they have not published the tweets of the users. We are aware that releasing a dataset containing social relations and text might violate the users' privacy. Therefore, we suggest anonymizing the data by replacing all user names with anonymous identifiers.

We proved that our multimodal model combined with the SHAP framework provides reasonable and meaningful explanations of its predictions associated with RQ2. These explanations allow us to gain a better understanding with respect of the models in two different ways: (1) the influence of the different submodels on the final predictions on an aggregated level; (2) the relevance of individual features (e.g., word, social relationship) for a single prediction. These explainable capabilities of our multimodal model are a further novelty. To our best knowledge, no one has developed such an explainable model for abusive language detection.

Even though the SHAP explanations are only an approximation, they are necessary for the reliable application of a hate speech detection model, as we have developed. It should be humanly interpretable how each of the three models influences predictions since we combine various data sources, which is especially true when one data source, such as the social network, is not fully transparent for the user. The reason for the missing transparency is that our network submodel learns patterns from social relations, which are more challenging to understand without any additional information than the ones from the text model. Therefore, these explainable capabilities are indispensable for such a system to provide a certain degree of transparency and build trustworthiness.

After focusing on the individual research questions, we have to add an ethical consideration regarding our developed model for various reasons. One may criticize that we integrate social network data, which is personal data, into our model and that the benefit gained by it bears no relation to the invasion of the user's privacy. However, we argue against it based on the following reasons: (1) We use social network data to train embeddings and identify patterns that do not contain any personal data. (2) The user's history and network are shown to enhance the detection rate, even if the used datasets are not the most appropriate ones for this experiment because of the limited density. Furthermore, detecting abusive language can be challenging if the author uses irony, sarcasm, or implicit wording. Therefore, context information (e.g., user's history or network) should be included because its benefit outweighs the damage caused by abusive language.

Another point of criticism could be the possible vulnerability to bias and systematic discrimination of users. In general, DL models are vulnerable to bias due to their black-box nature. In the case of a multimodal model, however, the issue is more aggravated because one submodel can dominate the prediction without any transparency for the user. For example, a model that classifies a user's tweet only because of their social relations discriminates the user with a high probability. We address this challenge by adding explainable capabilities with SHAP. Therefore, we claim that our multimodal model is less vulnerable to

bias than classical abusive language detection models applying DL techniques without XAI integration.

## 7 Conclusion & Outlook

This paper investigated whether users' previous posts and social network data can be leveraged to achieve good, humanly interpretable classification results in the context of abusive language. Concerning the classification performance (RQ1), we showed that the additional data improves the performance depending on the dataset and its network density. For WASEEM, we increased the macro F1 score by 2.4 pp, for DAVIDSON by 0.4 pp, and WICH by 0.1 pp. We found that the denser the network, the higher the gain. Nevertheless, the availability of appropriate datasets is a remaining challenge.

The model's interpretability (RQ2) demonstrated that our multimodal model using the SHAP framework produces meaningful and understandable explanations for its predictions. The explanations are provided both on a word level and connections to social communities in the user's ego network. The explanations help better understand a single prediction and the complete model if relevance scores are aggregated on a submodel level. Furthermore, explainability is a necessary feature of such a multimodal model to prevent bias and discrimination.

Integrating a user's previous posts and social network to enhance abusive language detection produced promising results. Therefore, the research community should continue exploring this approach because it might be a feasible way to address the challenge of detecting implicit hate, irony, or sarcasm. Concrete aspects that have to be addressed by future work are the following: (1) collecting appropriate data (in terms of size and network density) to refine our approach, (2) improving our model's architecture.

## Acknowledgments

We would like to thank Anika Apel and Mariam Khuchua for their contribution to this project. The research has been partially funded by a scholarship from the Hanns Seidel Foundation financed by the German Federal Ministry of Education and Research.

## References

1. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**(10), P10008 (Oct 2008)
2. Campbell, W., Baseman, E., Greenfield, K.: Content + context networks for user classification in twitter. In: *Frontiers of Network Analysis, NIPS Workshop*, 9 December 2013 (2013)

3. Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A.: Mean birds: Detecting aggression and bullying on twitter. In: WebSci. pp. 13–22 (2017)
4. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proc. 11th ICWSM Conf. (2017)
5. Fehn Unsvåg, E., Gambäck, B.: The effects of user features on Twitter hate speech detection. In: Proc. 2nd Workshop on Abusive Language Online (ALW2). pp. 75–85. ACL (2018)
6. Founta, A.M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., Leontiadis, I.: A unified deep learning architecture for abuse detection. In: WebSci. pp. 105–114. ACM (2019)
7. Friedman, B., Nissenbaum, H.: Bias in computer systems. *ACM Transactions on Information Systems* pp. 330–347 (1996)
8. Garland, J., Ghazi-Zahedi, K., Young, J.G., Hébert-Dufresne, L., Galesic, M.: Countering hate on social media: Large scale classification of hate and counter speech. In: Proc. 4th Workshop on Online Abuse and Harms. pp. 102–112 (2020)
9. Hamilton, W.L., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: NIPS. pp. 1024–1034 (2017)
10. Hennig, M., Brandes, U., Pfeffer, J., Mergel, I.: *Studying Social Networks. A Guide to Empirical Research*. Campus Verlag (2012)
11. Kreißel, P., Ebner, J., Urban, A., Guhl, J.: *Hass auf Knopfdruck. Rechtsextreme Trollfabriken und das Ökosystem koordinierter Hasskampagnen im Netz*. Institute for Strategic Dialogue (2018)
12. Li, S., Zaidi, N., Liu, Q., Li, G.: Neighbours and kinsmen: Hateful users detection with graph neural network. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer (2021)
13. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: NeurIPS (2017)
14. Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A.: Hatexplain: A benchmark dataset for explainable hate speech detection. arXiv preprint arXiv:2012.10289 (2020)
15. Mishra, P., Del Tredici, M., Yannakoudakis, H., Shutova, E.: Author profiling for abuse detection. In: COLING. pp. 1088–1098. ACL (2018)
16. Mishra, P., Yannakoudakis, H., Shutova, E.: Tackling online abuse: A survey of automated abuse detection methods. arXiv preprint arXiv:1908.06024 (2019)
17. Molnar, C.: *Interpretable Machine Learning* (2019), <https://christophm.github.io/interpretable-ml-book/>
18. Mosca, E., Wich, M., Groh, G.: Understanding and interpreting the impact of user context in hate speech detection. In: Proc. 9th Int. Workshop on Natural Language Processing for Social Media. pp. 91–102. ACL (2021)
19. Papagnies, E., Labatut, V., Dufour, R., Linares, G.: Graph-based features for automatic online abuse detection. In: SLSP. pp. 70–81. Springer (2017)
20. Qian, J., ElSherief, M., Belding, E., Wang, W.Y.: Leveraging intra-user and inter-user representation learning for automated hate speech detection. In: NAACL 2018 (Short Papers). pp. 118–123. ACL (2018)
21. Raisi, E., Huang, B.: Cyberbullying detection with weakly supervised machine learning. pp. 409–416. ASONAM '17, Association for Computing Machinery, New York, NY, USA (2017)
22. Ribeiro, M., Calais, P., Santos, Y., Almeida, V., Meira Jr, W.: Characterizing and detecting hateful users on twitter. In: Proc. Int. AAAI Conf. on Web and Social Media. vol. 12 (2018)

23. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In: 2019 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019 (2019)
24. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proc. 5th Int. Workshop on Natural Language Processing for Social Media. pp. 1–10. ACL (2017)
25. Shapley, L.: Quota solutions of n-person games. *Contributions to the Theory of Games* **2**, 343–359 (1953)
26. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* **41**(3), 647–665 (2014)
27. Švec, A., Pikuliak, M., Šimko, M., Bieliková, M.: Improving moderation of online discussions via interpretable neural models. In: Proc. 2nd Workshop on Abusive Language Online (ALW2). pp. 60–65. ACL (2018)
28. Vidgen, B., Hale, S., Guest, E., Margetts, H., Broniatowski, D., Waseem, Z., Botelho, A., Hall, M., Tromble, R.: Detecting East Asian prejudice on social media. In: Proc. 4th Workshop on Online Abuse and Harms. pp. 162–172. ACL (2020)
29. Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., Margetts, H.: Challenges and frontiers in abusive content detection. In: Proc. 3rd Workshop on Abusive Language Online. pp. 80–93. ACL (2019)
30. Vijayaraghavan, P., Larochelle, H., Roy, D.: Interpretable multi-modal hate speech detection. In: Proc. Int. Conf. on Machine Learning AI for Social Good Workshop (2019)
31. Wang, C.: Interpreting neural network hate speech classifiers. In: Proc. 2nd Workshop on Abusive Language Online (ALW2). pp. 86–92. ACL (2018)
32. Waseem, Z., Davidson, T., Warmlesley, D., Weber, I.: Understanding abuse: A typology of abusive language detection subtasks. In: Proc. 1st Workshop on Abusive Language Online. pp. 78–84. ACL (2017)
33. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In: Proc. NAACL Student Research Workshop. pp. 88–93. ACL (2016)
34. Wich, M., Bauer, J., Groh, G.: Impact of politically biased data on hate speech classification. In: Proc. 4th Workshop on Online Abuse and Harms. pp. 54–64. ACL (2020)
35. Wich, M., Breiting, M., Strathern, W., Naimarevic, M., Groh, G., Pfeffer, J.: Are your friends also haters? identification of hater networks on social media: Data paper. In: Companion Proc. Web Conference 2021. ACM (2021)
36. Williams, M.L., Burnap, P., Javed, A., Liu, H., Ozalp, S.: Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology* pp. 93–117 (2020)
37. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. ACL (2020)
38. Xu, J., Lu, T.C., et al.: Automated classification of extremist twitter accounts using content-based and network-based features. In: 2016 IEEE Int. Conf. on Big Data. pp. 2545–2549. IEEE (2016)



## A.6 STUDY XI

©2021 The Author(s).

Maximilian Wich, Svenja Räther, and Georg Groh (Sept. 2021). "German Abusive Language Dataset with Focus on COVID-19." In: *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*. Düsseldorf, Germany: KONVENS 2021 Organizers, pp. 247–252. ISBN: 978-1-954085-83-1. URL: <https://aclanthology.org/2021.konvens-1.26>

### *Publication Summary*

"The COVID-19 pandemic has had a significant impact on human lives globally. As a result, it is unsurprising that it has influenced hate speech and other sorts of abusive language on social media. Machine learning models have been designed to automatically detect such posts and messages, which necessitate a significant amount of labeled data. Despite the relevance of the COVID-19 topic in the context of abusive language, no annotated datasets with this focus are available. To solve these shortfalls, we target to create such a dataset. Our contributions are as follows: (1) a methodology for collecting abusive language data from Twitter with a substantial amount of abusive and hateful content, and (2) a German abusive language dataset with 4,960 annotated tweets centered on COVID-19. Both the methodology and the dataset are intended to aid researchers in improving abusive language detection." (Wich, Räther, and Groh, 2021, p. 1)

### *Author Contributions*

Maximilian Wich headed the research project. He developed the initial idea, the concept, and the methodology of the study. Furthermore, he trained the classification models for the different datasets. Additionally, he wrote most of the manuscript. Svenja Räther collected the data, headed the annotation process, and annotated data as part of her master's thesis supervised by Maximilian Wich. Furthermore, she provided feedback on the manuscript. Svenja Räther and Maximilian Wich share the first authorship. Georg Groh regularly discussed the ideas and concepts with the team and provided feedback on the study.

# German Abusive Language Dataset with Focus on COVID-19

Maximilian Wich\*, Svenja Räther\*, and Georg Groh

Technical University of Munich, Department of Informatics, Germany

maximilian.wich@tum.de, svenja.raether@tum.de, grohg@in.tum.de

## Abstract

The COVID-19 pandemic has had a significant impact on human lives globally. As a result, it is unsurprising that it has influenced hate speech and other sorts of abusive language on social media. Machine learning models have been designed to automatically detect such posts and messages, which necessitate a significant amount of labeled data. Despite the relevance of the COVID-19 topic in the context of abusive language, no annotated datasets with this focus are available. To solve these shortfalls, we target to create such a dataset. Our contributions are as follows: (1) a methodology for collecting abusive language data from Twitter with a substantial amount of abusive and hateful content, and (2) a German abusive language dataset with 4,960 annotated tweets centered on COVID-19. Both the methodology and the dataset are intended to aid researchers in improving abusive language detection.

## 1 Introduction

Hate speech is a serious challenge that social media platforms are currently confronting (Duggan, 2017). However, it is not limited to the online world. According to a study, there is a link between online hate and physical crime (Williams et al., 2020). As a result, it is critical to combat hate speech and other forms of abusive language on social media platforms to improve the conversation atmosphere and prevent spillover.

Owed to the large amounts of content created by billions of users, it is inefficient to detect this phenomenon manually. Therefore, its automatic detection is an essential part of the fight against this. Machine learning is a promising technology that aids in the training of classification models for detecting hate speech.

The success of a classification model depends largely on its training data. It requires data to learn patterns that can be used for solving the task. Large amounts of labeled data are required in the context of hate speech because hate speech is multifaceted and diversified (e.g., misogyny, racism, anti-Semitism) (Rieger et al., 2021). As a result, researchers have published many abusive language datasets in recent years (Vidgen and Derczynski, 2020; Wich et al., 2021b; Schmidt and Wiegand, 2017). The majority of the datasets are in English, and only a small portion is in German. Another shortcoming of the existing datasets is that, with some exceptions, they do not cover COVID-19-related hate (Vidgen et al., 2020; Alshalan et al., 2020; Ziems et al., 2020). COVID-19, however, has become a popular topic in the hate and extremist communities (Guhl and Gerster, 2020; Velásquez et al., 2020), making it a suitable topic in the hate speech and abusive language detection community as well. Our research goal is to develop a German abusive language dataset with an emphasis on COVID-19 to solve both shortcomings.

Contribution:

- With a topical focus, we present a methodology for collecting abusive language from Twitter.
- We report a 4,960-tweet German abusive language dataset with an emphasis on COVID-19. The labeling schema comprises two classes: *ABUSIVE* (22%) and *NEUTRAL* (78%).

## 2 Related Work

German abusive language datasets can be found in the literature. Ross et al. (2016) published a 469 tweets dataset on anti-refugee sentiment. Bretschneider and Peters (2017) published a dataset

---

\*These authors contributed equally to this work.

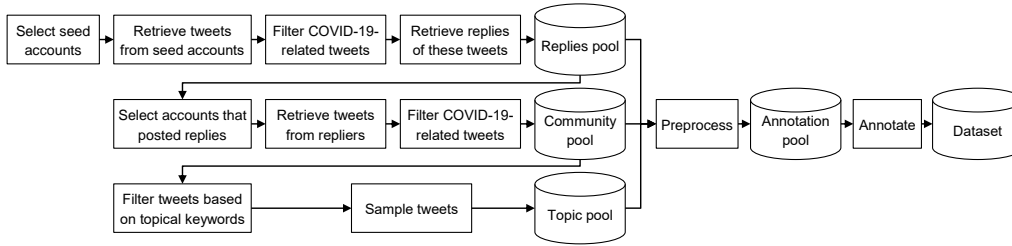


Figure 1: Dataset creation process adapted from R  ther (2021)

of 5,836 Facebook posts on anti-foreigner prejudices. Two abusive language datasets have been reported as part of GermEval, a series of shared tasks focusing on the German language (Wiegand et al., 2018; Stru   et al., 2019). The dataset from 2018 contains 8,541 tweets and the one from 2019 7,025. Both utilized the same labeling schema. Based on the interpretation of the data collection, the tweets do not seem to have a topical focus (Wiegand et al., 2018; Stru   et al., 2019). Two additional German datasets were reported as part of the multilingual shared task series "Hate Speech and Offensive Content Identification in Indo-European Languages" (HASOC) (Mandl et al., 2019, 2020, p. 14). The German dataset from the shared task contained 4,669 posts from Twitter and Facebook in 2019 (Mandl et al., 2019); 3,425 posts from YouTube and Twitter in 2020 (Mandl et al., 2020). The only German dataset that comprises posts from the COVID-19 period is from Wich et al. (2021a). However, the authors did not concentrate on COVID-19 content.

Several researchers have published abusive language datasets that directly tackle the COVID-19 topics, nevertheless, they are small in number. Vidgen et al. (2020) published an English Twitter dataset about East Asian prejudice from 20,000 posts collected during the pandemic. Ziems et al. (2020) collected tweets related to anti-Asian hate speech and counter hate. They annotated 2,400 tweets and utilized these tweets to train a classifier and detected "891,204 hate and 200,198 counter hate tweets" (Ziems et al., 2020, p.2). However, to the best of our knowledge, no one has reported a German abusive language or hate speech dataset with attention on COVID-19.

### 3 Methodology

The dataset creation process comprised three parts. The first one dealt with the data gathering and selection approach we employed to retrieve data from Twitter with a high portion of abusive content. Con-

sequently, the selected data is annotated by three annotators. Finally, we assessed the newly developed dataset based on dataset metrics and compared it with other German abusive language datasets.

#### 3.1 Collecting Data

Figure 1 demonstrates the data collection process that we report in the following. The tweets to be annotated are sampled from the *annotation pool* equally fed by three other pools—*replies pool*, *community pool*, and *topic pool*. Ensuring a topical concentration on COVID-19 and a high portion of hateful content is the reason for this approach.

The starting point of the data collection for all pools was a set of three seed accounts. These accounts originate from a study conducted by Richter et al. (2020), in which the authors have described influential Twitter accounts sharing misinformation about COVID-19. The accounts were selected by the authors based on the following criteria (Richter et al., 2020): (1) At least 20,000 accounts follow the account. (2) The account has shared or reported misinformation about COVID-19. (3) The account was active as of May 20, 2020. These accounts were chosen as seeds because hateful content often coincides with misinformation (Guhl and Gerster, 2020).

From these accounts, we retrieved the tweets that they published between 01.01.2020 and 20.02.2021 through the Twitter API. Subsequently, we filtered out the tweets that are related to COVID-19. We used a list of 65 keywords for this purpose (see Table 1). It comprised stemmed terms from a glossary about the current pandemic<sup>1</sup> and some additions. Next, we retrieved the replies to these tweets through the Twitter API—a reply is a tweet that refers to another tweet. These replies were stored in the *replies pool*. To ensure the quantity and quality of hateful content, two annotators analyzed a sample of 100 tweets.

<sup>1</sup> [www.dwds.de/themenglossar/Corona](http://www.dwds.de/themenglossar/Corona)

The *community pool* comprised COVID-19-related tweets from the accounts that replied to the seed accounts' tweets. We utilized a similar approach as in the previous phases. We retrieved the tweets from the accounts, limiting the maximum number of tweets per account to 500 and considering only tweets posted beyond 01.01.2020. The retrieved tweets were then filtered based on the 65 COVID-19-keywords. A sample of 100 tweets undergoes the same quality inspection as in the previous phase.

The third and last pool was the *topic pool*, whose purpose was to increase the prevalence of hateful content and topical diversity. It consists of tweets related to topics that coincide in the context of COVID-19 and hate speech (sCAN, 2020). Table 2 illustrates the topics provided by sCAN (2020) and the associated keywords that we employed for filtering the tweets. To balance the different topics, we limited the number of filtered tweets per keyword to 1,000.

After filling the data pools, we applied two pre-processing phases to the data. First, all tweets holding less than two textual tokens were removed. Second, close and exact duplicates were removed by using locality-sensitive hashing with Jaccard similarity (Leskovec et al., 2020). Third, account names appearing in the tweets are masked to reduce annotator bias created by account names recognition. The *annotation pool* was then created by sampling the pools equally.

### 3.2 Annotating Data

The annotation schema for the sampled tweets comprised two classes:

- **ABUSIVE**: The tweet comprised "any form of insult, harassment, hate, degradation, identity attack, and the threat of violence targeting an individual or a group" (Räther, 2021, p. 36).
- **NEUTRAL**: The tweet did "not fall into the **ABUSIVE** class" (Räther, 2021, p. 36).

The data is annotated by three non-experts (two female, one male; all between 20 and 30 years old).

To prepare them for the annotation process, they received training that contained a presentation of the annotation guidelines and a discussion among all annotators to define the task. Since the annotators are non-experts, we permitted them to skip tweets if they are indifferent (e.g., due to unclear cases or missing context information). This is to prevent the impairment of the quality of labels. The label indifference was handled as a missing label in the further course. Owing to limited resources, 275 tweets were annotated by two or three annotators to assess the inter-rater reliability with Krippendorff's alpha (Krippendorff, 2004). All other tweets received only one annotation from any of the annotators. We employed `doccano` as an annotation tool (Nakayama et al., 2018).

### 3.3 Evaluating Dataset

We compared our dataset with the GermEval and HASOC datasets by investigating the cross-dataset classification performance. For this purpose, we trained each dataset on a binary classification model for abusive language and assessed the models on all test sets. This is possible because the binary labels of all datasets are compatible. The objective of this assessment is to investigate how well our dataset generalizes and how well classifiers that were trained on a dataset without any COVID-19 content performed on our dataset. The classification model employed the German pre-trained BERT base model `deepset/gbert-base` as a basis (Chan et al., 2020). Before training the model, we removed all user names and URLs. The models were trained for 6 epochs with a learning rate of  $5 \times 10^{-5}$ . Evaluation was conducted after each epoch and the model with the highest macro F1 was selected. The validation set is 15% of the training set.

## 4 Results

At the end of the data collection process, we obtained 768,419 unique tweets from 7,629 users in our overlapping pools. The final dataset sampled from these pools without duplication, and anno-

Table 1: COVID-19-related keywords for filtering (table from Räther (2021, p. 84))

covid, corona, wuhan, biontech, pfizer, moderna, astra, zeneca, sputnik, abstandsregel, aluhut, antikörpertest, ansteck, asymptomatisch, ausgangssperre, ausgehverbot, ausreisesperre, balkonien, beatmungsgerät, besuchsverbot, desinf, durchsuchung, einreisesperre, einreiseverbot, epidemi, existenzangst, fallzahl, gesichtsvisier, gesundheitsamt, grundrechte, hygienedemo, hygienemaßnahme, immun, impf, infekt, influenza, inkubationszeit, intensivbett, inzidenz, kontaktbeschränkung, kontaktverbot, lockdown, lockerungen, mundschtutz, mutation, maske, pandemie, pcr, pharmaunternehmen, präventionsmaßnahme, plandemie, querdenk, quarantäne, reproduktionszahl, risikogruppe, sars-cov, shutdown, sicherheitsabstand, superspreader, systemrelevant, tracing-app, tröpfcheninfektion, übersterblichkeit, vakzin, virolog, virus

Table 2: Hate- and COVID-19-related topics and keywords (column Topic taken over word for word from sCAN (2020); entire table from R  ther (2021, p. 84))

Topic (sCAN, 2020)	Keywords
"Anti-Asian racism"	asiat, chines, ccp, wuhan, chinavirus
"Misinformation and geopolitical strategy"	amerika, milit��r, biowaffe
"Resurgence of old antisemitic stereotypes"	jude, j��disch, pest, schwarze tod
"New world order, «anti-elites» speech and traditional conspiracy theories"	elite, #nwo, weltordnung, deepstate, plandemie
"Fear of the «internal enemy», exclusion of the foreigner and scapegoating mechanisms"	greatreset, muslim, illegal, migrant

tations by our three annotators comprised 4,960 tweets. 22% of the tweets were labeled as *ABUSIVE* by our annotators, whereas 78% were labeled as *NEUTRAL*. The annotated tweets were created by 2,662 accounts—on average 1.86 tweets per account (min: 1; max: 41). All tweets were posted between January 2020 and February 2021.

Krippendorff’s alpha of the three annotators is 91.5%, which is a good score for inter-rater reliability. Only 275 tweets were annotated by two or three annotators owing to limited resources.

Table 3 demonstrates the classification metrics of the classifier trained and assessed on our COVID-19 dataset. The train set contained 3,485 tweets, the validation set 735, and the test set 740. We ensured that an author appeared only in one of the three sets. Without any architecture optimization or hyperparameter search, we obtained a macro F1 score of 82.9%. Considering the metrics for the *ABUSIVE* class, we can see that there is still room for improvement. However, this study does not aim to develop the latest state-of-the-art model. This classifier is intended to serve as a baseline for future studies utilizing our new COVID-19 dataset.

To compare our dataset with another German abusive language dataset, we investigated the cross-dataset classification performance. As indicated in Table 4, the rows correspond to the classifiers, whereas the columns to the test sets. We observed that the model trained on the COVID-19 dataset demonstrated similar performance as the ones from the GermEval datasets. Its macro F1 score is in the same range as the ones from GermEval and it performed similarly on the other test sets. The

Table 3: Classification metrics of COVID-19 classifier on its test set in percent

Class	Precision	Recall	F1
NEUTRAL	92.4	93.7	93.1
ABUSIVE	74.7	70.8	72.7
Macro avg.	83.5	82.2	82.9

Table 4: Cross-dataset classification performance (macro F1 in percent) – CD = COVID, GE = GermEval, HC = HASOC

	CD-19	GE 18	GE 19	HC 19	HC 20
CD-19	82.9	72.8	76.7	67.8	68.0
GE 18	73.4	76.9	74.6	65.4	65.4
GE 19	73.3	75.2	75.3	62.5	73.0
HC 19	60.8	63.4	63.9	66.4	64.6
HC 20	54.0	59.9	53.1	48.6	80.5

classifiers from the HASOC datasets step out of line. The HASOC 2020 classifier seemed to concentrate on a different type of abusive language. It performed quite well on its dataset but scored lower on all other test sets. Even if the GermEval classifiers scored higher results on the COVID-19 test set, they did not achieve the same F1 score as the COVID-19 classifier. This indicates that abusive language in the domain of COVID-19 varies from what it was before the pandemic.

## 5 Conclusion

We created a German abusive language dataset that focuses on COVID-19. It contains 4,960 annotated tweets from 2,662 accounts. 22% of the tweets are labeled as *ABUSIVE*, 78% as *NEUTRAL*. Due to limited resources, not all documents were annotated by two or more annotators. We prioritized holding a variety of tweets equivalent to the size of related German datasets. Furthermore, the high inter-rater reliability for the overlapping annotations indicates that the annotation behavior of the three annotators was well aligned. Also, the generalizability of the dataset demonstrates that our COVID-19 dataset has an equivalent cross-dataset classification performance.

Our second contribution is a dataset creation methodology for abusive language. We indicated that it aids in the creation of a dataset with a significant portion of abusive language.

We consider both our dataset and the dataset creation methodology noteworthy contributions to the hate speech detection community.

## Resources

Code and data are available under [github.com/mawic/german-abusive-language-covid-19](https://github.com/mawic/german-abusive-language-covid-19).

## Acknowledgments

This paper is based on a joined work in the context of Svenja Räther's master's thesis (Räther, 2021). The research has been partially funded by a scholarship from the Hanns Seidel Foundation financed by the German Federal Ministry of Education and Research. The paper is based upon work partially supported by Google Cloud Platform research credits.

## References

- Raghad Alshalan, Hend Al-Khalifa, Duaa Alsaeed, Heyam Al-Baity, and Shahad Alshalan. 2020. *Detection of Hate Speech in COVID-19-Related Tweets in the Arab Region: Deep Learning and Topic Modeling Approach*. *J Med Internet Res*, 22(12):e22609.
- Uwe Bretschneider and Ralf Peters. 2017. Detecting offensive statements towards foreigners in social media. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.
- Branden Chan, Stefan Schweter, and Timo Möller. 2020. *German's next language model*. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Maeve Duggan. 2017. *Online harassment 2017*. Pew Research Center.
- Jakob Guhl and Lea Gerster. 2020. *Krise und Kontrollverlust - Digitaler Extremismus im Kontext der Corona-Pandemie*. Institute for Strategic Dialogue.
- K. Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Content Analysis: An Introduction to Its Methodology. Sage.
- Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2020. *Finding Similar Items*, 3 edition, pages 78–137. Cambridge University Press.
- Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. *Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German*. In *Forum for Information Retrieval Evaluation*, FIRE 2020, pages 29–32, New York, NY, USA. Association for Computing Machinery.
- Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. *Overview of the HASOC Track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages*. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, pages 14–17, New York, NY, USA. Association for Computing Machinery.
- Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. *doccano: Text Annotation Tool for Human*. Software available from <https://github.com/doccano/doccano>.
- Svenja Räther. 2021. *Investigating Techniques for Learning with Limited Labeled Data for Hate Speech Classification*. Master's thesis, Technical University of Munich. Advised and supervised by Maximilian Wich and Georg Groh.
- Marie Richter, Chine Labbé, Virginia Padovese, and Kendrick McDonald. 2020. *Twitter: Superspreader von Corona-Falschinformationen*.
- Diana Rieger, Anna Sophie Kuempel, Maximilian Wich, T. Kiening, and Georg Groh. 2021. *Assessing the Prevalence and Contexts of Hate Speech in Fringe Communities: A Case Study of Alt-Right Communities on 8chan, 4chan, and Reddit*. In *Proceedings of 71st Annual ICA Conference*.
- Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. *Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis*. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, volume 17 of *Bochumer Linguistische Arbeitsberichte*, pages 6–9, Bochum.
- sCAN. 2020. *Hate speech trends during the Covid-19 pandemic in a digital and globalised age*. SCAN project – Platforms, Experts, Tools: Specialised Cyber-Activists Network. [scan-project.eu/wp-content/uploads/sCAN-Analytical-Paper-Hate-speech-trends-during-the-Covid-19-pandemic-in-a-digital-and-globalised-age.pdf](https://scan-project.eu/wp-content/uploads/sCAN-Analytical-Paper-Hate-speech-trends-during-the-Covid-19-pandemic-in-a-digital-and-globalised-age.pdf).
- Anna Schmidt and Michael Wiegand. 2017. *A survey on hate speech detection using natural language processing*. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.
- Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, Manfred Klenner, et al. 2019. *Overview of GermEval Task 2, 2019 shared task on the identification of offensive language*. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 354–365.

- Nicolás Velásquez, R Leahy, N Johnson Restrepo, Yonatan Lupu, R Sear, N Gabriel, Omkant Jha, B Goldberg, and NF Johnson. 2020. Hate multi-verse spreads malicious COVID-19 content online beyond individual platform control. *arXiv preprint arXiv:2004.00673*.
- Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.
- Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. [Detecting East Asian prejudice on social media](#). In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.
- Maximilian Wich, Melissa Breitingner, Wienke Strathern, Marlena Naimarevic, Georg Groh, and Jürgen Pfeffer. 2021a. Are your Friends also Haters? Identification of Hater Networks on Social Media: Data Paper. In *Companion Proceedings of the Web Conference 2021 (WWW'21 Companion)*.
- Maximilian Wich, Tobias Eder, Hala Al Kuwatly, and Georg Groh. 2021b. [Bias and comparison framework for abusive language datasets](#). *AI and Ethics*.
- Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)*.
- Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2020. Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. *The British Journal of Criminology*, 60(1):93–117.
- Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. 2020. Racism is a Virus: Anti-Asian Hate and Counterhate in Social Media during the COVID-19 Crisis. *arXiv preprint arXiv:2005.12423*.



PUBLICATIONS <sup>†</sup>

---

Publications in Appendix B are not formally relevant for examination in accordance with Exhibit 6 of the regulations for the award of doctoral degree.

## B.1 STUDY I

©2021 The Author(s), published under Creative Commons CC-BY 4.0 License<sup>1</sup>.

Diana Rieger, Anna Sophie Kümpel, Maximilian Wich, Toni Kiening, and Georg Groh (Oct. 2021). "Assessing the Extent and Types of Hate Speech in Fringe Communities: A Case Study of Alt-Right Communities on 8chan, 4chan, and Reddit." In: *Social Media + Society* 7.4. DOI: [10.1177/205630512111052906](https://doi.org/10.1177/205630512111052906)

---

<sup>1</sup> <https://creativecommons.org/licenses/by/4.0/>

### *Publication Summary*

"Recent right-wing extremist terrorists were active in online fringe communities connected to the alt-right movement. Although these are commonly considered as distinctly hateful, racist, and misogynistic, the prevalence of hate speech in these communities has not been comprehensively investigated yet, particularly regarding more implicit and covert forms of hate. This study exploratively investigates the extent, nature, and clusters of different forms of hate speech in political fringe communities on Reddit, 4chan, and 8chan. To do so, a manual quantitative content analysis of user comments (N = 6,000) was combined with an automated topic modeling approach. The findings of the study not only show that hate is prevalent in all three communities (24% of comments contained explicit or implicit hate speech), but also provide insights into common types of hate speech expression, targets, and differences between the studied communities." (Rieger et al., 2021, p. 1)

### *Author Contributions*

Diana Rieger was responsible for developing the research question, writing the manuscript, and analyzing the data for the manual content analysis. Anna Sophie Kümpel wrote the paper together with Diana Rieger and co-developed the research questions. Maximilian Wich was responsible for the third research question, developed the topic models, and analyzed the data. Regarding the writing of the paper, he wrote the corresponding parts and provided overall feedback. Toni Kiening was responsible for data collection, was one of the coders for the manual content analysis, and took part in the data analysis. Georg Groh regularly discussed the ideas and concepts and provided feedback on the study.

# Assessing the Extent and Types of Hate Speech in Fringe Communities: A Case Study of Alt-Right Communities on 8chan, 4chan, and Reddit

Diana Rieger<sup>1</sup> , Anna Sophie Kümpel<sup>2</sup> , Maximilian Wich<sup>3</sup>, Toni Kiening<sup>1</sup>, and Georg Groh<sup>3</sup>

Social Media + Society  
October-December 2021: 1–14  
© The Author(s) 2021  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/20563051211052906  
journals.sagepub.com/home/sms  


## Abstract

Recent right-wing extremist terrorists were active in online fringe communities connected to the alt-right movement. Although these are commonly considered as distinctly hateful, racist, and misogynistic, the prevalence of hate speech in these communities has not been comprehensively investigated yet, particularly regarding more implicit and covert forms of hate. This study exploratively investigates the extent, nature, and clusters of different forms of hate speech in political fringe communities on *Reddit*, *4chan*, and *8chan*. To do so, a manual quantitative content analysis of user comments ( $N = 6,000$ ) was combined with an automated topic modeling approach. The findings of the study not only show that hate is prevalent in all three communities (24% of comments contained explicit or implicit hate speech), but also provide insights into common types of hate speech expression, targets, and differences between the studied communities.

## Keywords

hate speech, alt-right, fringe communities, Reddit, 4chan, 8chan, content analysis, topic modeling

On 15 March 2019, a right-wing extremist terrorist killed more than 50 people in mosques in Christchurch, New Zealand, and wounded numerous others—livestreaming his crimes on Facebook. Only 6 weeks later, on 27 April, another right-wing extremist attack occurred in a synagogue in Poway near San Diego, in which one person was killed and three more injured. The perpetrators were active in an online community within the imageboard *8chan*, which is considered as particularly hateful and rife with right-wing extremist, misanthropic, and White-supremacist ideas. Moreover, both the San Diego and Christchurch shooters used *8chan* to post their manifestos, providing insights into their White nationalist hatred (Stewart, 2019). Following the attack in New Zealand, Internet service providers in Australia and New Zealand have temporarily blocked access to *8chan* and the similar—albeit less extreme—imageboard *4chan* (Brodtkin, 2019). After yet another shooting in El Paso was linked to activities on *8chan*, the platform was removed<sup>1</sup> from the Clearnet entirely, with one of *8chan*'s network infrastructure providers claiming the unique lawlessness of the site that “has contributed to multiple horrific tragedies” as the main reason for this decision (Prince, 2019).

Whether the perpetrators' activities on *8chan* and *4chan* actually contributed to their radicalization or motivation can hardly be determined. However, especially the platforms' politics boards (*8chan/pol/* and *4chan/pol/*, respectively) have repeatedly been linked to the so-called alt-right movement, “exhibiting characteristics of xenophobia, social conservatism, racism, and, generally speaking, hate” (Hine et al., 2017, p. 92; see also Hawley, 2017; Tuters & Hagen, 2020). *4chan/pol/*, in particular, has attracted the broader public's attention during Donald Trump's 2016 presidential campaign, often being the birthplace of conservative or even outright hateful and racist memes that circulated during the campaign. In addition to the mentioned communities on *4chan* and *8chan*, the controversial subreddit “The\_Donald” is often referenced as a popular and more

<sup>1</sup>LMU Munich, Germany

<sup>2</sup>TU Dresden, Germany

<sup>3</sup>Technical University of Munich (TUM), Germany

### Corresponding Author:

Diana Rieger, Department of Media and Communication, LMU Munich, Oettingenstrasse 67, 80538 Munich, Germany.  
Email: diana.rieger@ifkw.lmu.de



“mainstreamy” outlet for alt-right ideas as well (e.g., Heikkilä, 2017).

Although these political fringe communities are considered as particularly hateful in the public debate, only few studies (Hine et al., 2017; Mittos, Zannettou, Blackburn, & De Cristofaro, 2019) have investigated these communities with regard to the extent of hate speech. Moreover, the mentioned studies are exclusively built on automated dictionary-based approaches focusing on explicit “hate terms,” thus being unable to account for more subtle or covert forms of hate. To better understand the different types of hate speech in these communities, it also seems advisable to cluster comments in which hate speech occurs.

Addressing these research gaps, we (a) provide a systematic investigation of the extent and nature of hate speech in alt-right fringe communities, (b) examine both explicit and implicit forms of hate speech, and (c) merge manual coding of hate speech with automated approaches. By combining a manual quantitative content analysis of user comments ( $N=6,000$ ) and unsupervised machine learning in the form of topic modeling, this study aims at understanding the extent and nature of different types of hate speech as well as the thematic clusters these occur in. We first investigate the extent and target groups of different forms of hate speech in the three mentioned alt-right fringe communities on Reddit (r/The\_Donald), 4chan (4chan/pol/), and 8chan (8chan/pol/). Subsequently, by means of a topic modeling approach, the clusters in which hate speech occurs are analyzed in more detail.

## Hate Speech in Online Environments

Hate speech was certainly not invented with the Internet. Being situated “in a complex nexus with freedom of expression, individual, group, and minority rights, as well as concepts of dignity, liberty, and equality” (Gagliardone, Gal, Alves, & Martínez, 2015, p. 10), it has been in the center of legislative discussion in many countries for many years. Hate speech is considered to be an elusive term, with extant definitions oscillating between strictly legal rationales and generic understandings that include almost all instances of incivility or expressions of anger (Gagliardone et al., 2015). For the context of this study, we deem both the content and the targets as crucial for conceptualizing hate speech. Accordingly, hate speech is defined here as the expression of “hatred or degrading attitudes toward a *collective*” (Hawdon, Oksanen, & Räsänen, 2017, p. 254), with people being devalued not based on individual traits, but on account of their race, ethnicity, religion, sexual orientation, or other group-defining characteristics (Hawdon et al., 2017, see also Kümpel & Rieger, 2019).

There are a number of factors—resulting from the overarching characteristics of online information environments—suggesting that hate speech is particularly problematic on the Internet. First, there is the problem of permanence

(Gagliardone et al., 2015). Especially fringe communities are heavily centered on promoting users’ freedom of expression, making it unlikely that hate speech will be removed by moderators or platform operators. But even if hateful content is removed, it might have already been circulated to other platforms, or it could be reposted to the same site again shortly after deletion (Jardine, 2019). Second, the shareability and ease of disseminating content in online environments further facilitates the visibility of hate speech (Kümpel & Rieger, 2019). During the 2016 Trump campaign, hateful anti-immigration and anti-establishment memes were often spread beyond the borders of fringe communities, surfacing to mainstream social media and influencing discussions on these platforms (Heikkilä, 2017). Third, the (actual or perceived) anonymity in online environments can encourage people to “be more outrageous, obnoxious, or hateful in what they say” (Brown, 2018, p. 298), because they feel disinhibited and less accountable for their actions. Moreover, anonymity can also change the relative salience of one’s personal and social identity, thereby increasing conformity to perceived group norms (Reicher, Spears, & Postmes, 1995). Indeed, research has found that exposure to online comments with ethnic prejudices leads other users to post more prejudiced comments themselves (Hsueh, Yogeewaran, & Malinen, 2015), suggesting that the communication behavior of others also influences one’s own behavior. Fourth, and closely related to anonymity, there is the problem of the full or partial invisibility of other users (Brown, 2018; Lapidot-Lefler & Barak, 2012): The absence of facial expressions and other visibility originated interpersonal communication cues makes hate speech appear less hurtful or damaging in an online setting, thus increasing inhibitions to discriminate others. Last, one has to consider the community-building aspects that are particularly distinctive for online hate speech (Brown, 2018; McNamee, Peterson, & Peña, 2010). Not least in alt-right fringe communities, hate is often “memeified” and mixed with humor and domain-specific slang, creating a situation in which the use of hate speech can play a crucial role in strengthening bonds among members of the community and distinguishing one’s group from clueless outsiders (Tuters & Hagen, 2020). Taken together, the mentioned factors facilitate not only the creation and use of hate speech in online environments, but also its wider dissemination and visibility.

## Implicit Forms of Hate Speech

While many types of online hate speech are relatively straightforward and “in your face” (Borgeson & Valeri, 2004), hate can also be expressed in a more implicit or covert form (see Ben-David & Matamoros-Fernández, 2016 ; Benikova, Wojatzki, & Zesch, 2018; ElSherief, Kulkarni, Nguyen, Wang, & Belding, 2018; Magu & Luo, 2018; Matamoros-Fernández, 2017)—for example, by spreading negative stereotypes or strategically elevating one’s ingroup.

Implicit hate speech shares characteristics with what Buyse (2014, p. 785) has labeled fear speech, which is “aimed at instilling (existential) fear of another group” by highlighting harmful actions the target group has allegedly engaged in or speculations about their goals to “take over and dominate in the future” (Saha, Mathew, Garimella, & Mukherjee, 2021, p. 1111). Indeed, one variety of implicit hate speech can be seen in the intentional spreading of “fake news,” in which deliberate false statements or conspiracy theories about social groups are circulated to marginalize them (Hajok & Selg, 2018). This could be observed in connection with the European migrant crisis during which online disinformation often focused on the degradation of immigrants, for example, through associating them with crime and delinquency (Hajok & Selg, 2018, see also Humprecht, 2019).

Implicitness is a major problem for the automated detection of hate speech, as it “is invisible to automatic classifiers” (Benikova et al., 2018, p. 177). Using such implicit forms of hate speech is a common strategy to even avoid automatic detection systems and to cloak prejudices and resentments in “ordinary” statements (e.g., “My cleaning lady is really good, even though she is Turkish,” see Meibauer, 2013). Thus, implicit hate speech points to the importance of acknowledging the wider context of hate speech instead of just focusing on the occurrence of single (and often ambiguous) hate terms.

### *Extent of Hate Speech*

Considering the mentioned problems with the (automated) detection of hate speech, it is hard to determine the overall prevalence of hate speech in online environments. To account for individual experiences, extant studies have often relied on surveys to estimate hate speech exposure. Across different populations around the globe, such self-reported exposure to online hate speech ranges from about 28% (New Zealanders 18+, see Pacheco & Melhuish, 2018), to 64% (13- to 17-year-old US Americans, see Common Sense, 2018), and up to 85% (14- to 24-year-old Germans, see Landesanstalt für Medien NRW, 2018). In studies focusing both on younger and older online users (Landesanstalt für Medien NRW, 2018; Pacheco & Melhuish, 2018), exposure to online hate was more commonly reported by younger age groups, which might be explained by different usage patterns and/or perceptual differences. However, while these survey figures suggest that many online users seem to have been exposed to hateful comments, they tell us only little about the overall amount of hate speech in online environments. In fact, even a single highly visible hate comment could be responsible for survey participants responding affirmatively to questions about their exposure to online hate. Thus, to determine the actual extent of hate speech, content analyses are needed—although the results are equally hard to generalize. Indeed, the amount of content labeled as hate speech seems to differ considerably, depending on the studied

platforms and (sub-)communities, the topic of discussions, or the lexical resources and dictionaries used to determine what qualifies as hate speech (ElSherief et al., 2018; Hine et al., 2017; Meza, 2016). Considering our focus on alt-right fringe communities, we will thus aim our attention at the presumed and actual hatefulness of these discussion spaces.

## **The “Alt-Right” Movement and Fringe Communities**

### *What Is the Alt-Right?*

The alt-right (=abbreviated form of alternative right) is a rather loosely connected and largely online-based political movement, whose ideology centers around ideas of White supremacy, anti-establishmentarianism, and anti-immigration (see Hawley, 2017; Heikkilä, 2017; Nagle, 2017). Gaining momentum during Donald Trump’s 2016 presidential campaign, the alt-right “took an active role in cheerleading his candidacy and several of his controversial policy positions” (Forscher & Kteily, 2020, p. 90), particularly on the mentioned message boards on Reddit (r/The\_Donald), 4chan, and 8chan (/pol/ on both platforms). Similar to other online communities, the alt-right uses a distinct verbal and visual language that is characterized by the use of memes, subcultural terms, and references to the wider web culture (Hawley, 2017; Tuters & Hagen, 2020; Wendling, 2018). Another common theme is “the cultivation of a position that sees white male identity as threatened” (Heikkilä, 2017, p. 4), which is connected both to strongly opposing policies related to “political correctness” (e.g., affirmative action) and to condemning social groups that are perceived to be profiting from these policies (Phillips & Yi, 2018). Openly expressing these ideas often culminates in the use of hate speech, particularly against people of color and women. However, while discussion spaces linked to the alt-right are routinely described as hateful, there is little published data on the quantitative amount of hate speech in these fringe communities.

### *Hate Speech in Alt-Right Fringe Communities*

To our knowledge, empirical studies addressing the extent of hate speech in alt-right fringe communities have exclusively relied on automated dictionary-based approaches, estimating the amount of hate speech by identifying posts that contain hateful terms (Hine et al., 2017; Mittos et al., 2019). Focusing on 4chan/pol/, Hine and colleagues (2017) use the hatebase dictionary to assess the prevalence of hate speech in the “Politically Incorrect” board. They find that 12% of posts on 4chan/pol/ contain hateful terms, thus revealing a substantially higher share than the two examined “baseline” boards 4chan/sp/ (focusing on sports) with 6.3% and 4chan/int/ (focusing on international cultures/languages) with 7.3%. However, 4chan generally seems to be more hateful than

other social media platforms: Analyzing a sample of Twitter posts for comparison, the authors find that only 2.2% of the analyzed tweets contained hateful terms. Looking at the most “popular” hate terms used in 4chan/pol/, it is also possible to draw cautious conclusions about the (main) target groups of hate speech. The hate terms appearing most—“nigger,” “faggot,” and “retard”—are indicative of racist, homophobic, and ableist sentiments and suggest that people of color, the lesbian, gay, bisexual, transgender and queer or questioning (LGBTQ) community, and people with disabilities might be recurrent victims of hate speech.

Utilizing a similar analytical approach, but exclusively focusing on discussions about genetic testing, Mittos and colleagues (2019) investigate both Reddit and 4chan/pol/ with regard to their levels of hate. For Reddit, their analysis shows that the most hateful subreddits alluding to the topic of genetic testing are associated with the alt-right (e.g., r/altright, r/TheDonald, r/DebateAltRight), with posts displaying “clear racist connotations, and of groups of users using genetic testing to push racist agendas” (Mittos et al., 2019, p. 9). These tendencies are even more amplified on 4chan/pol/ where discussion about genetic testing are routinely combined with content exhibiting racial and anti-Semitic hate speech. Reflecting the findings of Hine and colleagues (2017), racial and ethnic slurs are prevalent and illustrate the boards’ close association with White-supremacist ideologies.

While these studies offer some valuable insights into the hatefulness of alt-right fringe communities, the dictionary-based approaches are unable to account for more veiled and implicit forms of hate speech. Moreover, although the most “popular” terms hint at the targets of hate speech, a systematic investigation of the addressed social groups is missing. Based on the literature review and theoretical considerations, our study thus sought to answer three overarching research questions:

*Research Question 1.* What percentage of user comments in the three fringe communities contains explicit or implicit hate speech?

*Research Question 2.* (a) In which way is hate speech expressed and (b) against which persons/groups is it directed?

*Research Question 3.* What is the topical structure of the coded user comments?

## Method

Our empirical analysis of alt-right fringe communities focuses on three discussion boards within the platforms Reddit (r/The\_Donald), 4chan (4chan/pol/), and 8chan (8chan/pol/), thus spanning from central and highly used to more peripheral and less frequented communities. While

Reddit, the self-proclaimed “front page of the Internet,” routinely ranks among the 20 most popular websites worldwide, 4chan and 8chan have (or had) considerably less reach. However, due to their connection with the perpetrators of Christchurch, Poway, and El Paso, 4chan and 8chan are nevertheless of high relevance for this investigation. All three platforms follow a similar structure and are divided into a number of different subforums (called “subreddits” on Reddit and “boards” on 4chan/8chan). While Reddit requires users to register to post or comment, both 4chan and 8chan do not have a registration system, thus allowing everyone to contribute anonymously. The specific discussion boards—r/The\_Donald, 4chan/pol/, and 8chan/pol/—were chosen due to their association with alt-right ideas as well as their relative centrality within the three platforms. Moreover, all three boards have previously been discussed as important outlets of right-wing extremists’ online activities (Conway, Macnair, & Scrivens, 2019).

In the following sections, we will first describe the data collection process and then outline the two methodological/analytical approaches used in this study: (a) a manual quantitative content analysis of user comments in the three discussion boards and (b) an automated topic modeling approach. While 4chan and 8chan are indeed imageboards, (textual) comments play an important role on these platforms as well. On Reddit, pictures can easily be incorporated in the original post that constitutes the beginning of a thread, but comments are by default bound to text. Due to our two-pronged strategy, the nature of these communities, and to ensure comparability between the discussion boards, we focused our analyses on the textual content of comments and did not consider (audio-)visual materials such as images or videos. However, we refer to their importance in the context of hate speech in the discussion.

## Data Collection

Since accessing and collecting content from the three discussion boards varies in complexity, we relied on different sampling strategies. Comments from r/The\_Donald were obtained by querying the Pushshift Reddit data set (Baumgartner, Zannettou, Keegan, Squire, & Blackburn, 2020) via *redditsearch.io*. Between 21 April and 27 April 2019, we downloaded a total of 70,000 comments, of which 66,617 could be kept in the data set after removing duplicates and deleted/removed comments. Comments from 4chan/pol/ were obtained by using the independent archive page *4plebs.org* and a web scraper. Between 14 April and 29 April 2019, a total of 16,000 comments were obtained, of which 15,407 remained after the cleaning process.<sup>2</sup> Finally, comments from 8chan/pol/ were obtained by directly scraping the platform: All comments in threads that were active on 24 April 2019 were downloaded, resulting in a data set of 63,504 comments for this community. For the manual quantitative content analysis, 2,000 comments were randomly sampled

from the data set of each of the three communities, thus leading to a combined sample size of 6,000 comments.

### **Approach I: Manual Quantitative Content Analysis**

As our first main category, we coded explicit hate speech in accordance with recurrent conceptualizations in the literature. Within this category, we defined insults (attacks to individuals/groups on the basis of their group-defining characteristics, e.g., Erjavec & Kovačič, 2012) as offensive, derogatory, or degrading expressions, including the use of ethnophaulisms (Kinney, 2008). Instead of coding insults in general, we distinguished between personal insults (i.e., attacks of a specific individual) and general insults (i.e., attacks of a collective), also coding the reference point of personal insults and the target of general insults. The specific reference points [(a) Ethnicity, (b) Religion, (c) Country of Origin, (d) Gender, (e) Gender Identity, (f) Sexual Orientation, (g) Disabilities, (h) Political Views/Attitudes] or targets [(a) Black People, (b) Muslims, (c) Jews, (d) LGBTQ, (e) Migrants, (f) People with Disabilities, (g) Social Elites/Media, (h) Political Opponents, (i) Latin Americans\*, (j) Women, (k) Criminals\*, (l) Asians) were compiled on the basis of research on frequently marginalized groups (Burn, Kadlec, & Rexer, 2005; Mondal, Silva, Correa, & Benevenuto, 2018), and inductively extended (targets marked with \*) during the coding process. Furthermore, we have coded violence threats as a form of explicit hate speech (Erjavec & Kovačič, 2012; Gagliardone et al., 2015), including both concrete threats of physical, psychological, or other types of violence and calls for violence to be inflicted on specific individuals or groups.

As our second main category, we coded implicit hate speech. To distinguish different subcategories of this type of hate speech, we relied more strongly on an explorative approach by focusing on communication forms that have been described in the literature as devices to cloak hate (see section “Implicit Forms of Hate Speech”). The first subcategory of implicit hate speech is labeled negative stereotyping and was coded when users expressed overly generalized and simplified beliefs about (negative) characteristics or behaviors of different target groups. The second subcategory—disinformation/conspiracy theories—reflects both “simple” disinformation and false statements about target groups and “advanced” conspiracy theories that represent target groups as maliciously working together toward greater ideological, political, or financial power (e.g., “the Jew media controls everything”). A third subcategory was labeled ingroup elevation and was coded when statements elevated or accentuated belonging to a certain (racial, demographic, etc.) group, oftentimes implicitly excluding and devaluing other groups. The last subcategory of implicit hate speech was labeled inhuman ideology. Here, it was coded whether a user comment supported or glorified hateful ideologies such as

National Socialism or White supremacy, including the worshipping of prominent representatives of such ideologies.

In addition, a category spam was added to exclude comments containing irrelevant content such as random character combinations or advertisements. The entire coding scheme as well as an overview of the main content categories described in the previous paragraphs can be accessed via an open science framework (OSF) repository<sup>3</sup>.

The manual quantitative content analysis was conducted by two independent coders. Both coders coded the same subsample of 10% from the full sample of comments to calculate inter-rater reliability with the help of the R package “tidycomm” (Unkel, 2021). Using both percent agreement and Brennan and Prediger’s Kappa, all reliability values were satisfactory ( $\kappa \geq 0.83$ , see also Table 1). Prior to the analyses, all comments coded as spam were removed, leading to a final sample size of 5,981 comments.

### **Approach II: Topic Modeling**

Topic modeling is an unsupervised machine learning approach to identify topics within a collection of documents and to classify these documents into distinct topics. Günther and Domahidi (2017) generally describe a topic as “what is being talked/written about” (p. 3057). Each topic would thus be represented in a cluster. Consequently, each cluster is assigned a set of words that are representative of the comments within the cluster. For our analysis, we first generated a topic model (TM<sub>1</sub>) for all 5,981 comments to gain an understanding of the topics within the entire data set. Combined with the manual coding, these results provide insights on which topics are more hateful than others. Second, another topic model (TM<sub>2</sub>) was created only for the comments identified as hateful ( $n=1,438$ ) to examine the clusters of the comments in which hate speech occurs. To do so, TM<sub>1</sub> and TM<sub>2</sub> were compared by investigating the transitions between the models. In addition, TM<sub>2</sub> was also combined with the manually coded data, allowing to establish a connection between the cluster, type, and targets of hate speech.

CluWords was selected as the topic model algorithm—a state-of-the-art short-text topic modeling technique (Viegas et al., 2019). The reason for not choosing a more conventional technique such as Latent Dirichlet Allocation (LDA) is that these do not perform well on shorter texts because they rely on word co-occurrences (Campbell, Hindle, & Stroulia, 2003; Cheng, Yan, Lan, & Guo, 2014; Quan, Kit, Ge, & Pan, 2015). CluWords overcomes this issue by combining non-probabilistic matrix factorization and pre-trained word-embeddings (Viegas et al., 2019). Especially the latter allows enriching the comments with “syntactic and semantic information” (Viegas et al., 2019, p. 754). For this article, the fast-Text word vectors pre-trained on the English Common Crawl dataset were used because it is trained on web data and thus an appropriate basis (Mikolov, Grave, Bojanowski, Puhersch, & Joulin, 2019).

**Table 1.** Inter-Rater Reliability for Coded Categories.

Category	Percentage agreement	Brennan and Prediger's kappa
Source of the comment	1	1
Spam	0.99	0.99
Personal insult	0.93	0.87
Target of the personal insult	0.92	0.9
Reference point of the personal insult	0.92	0.91
Second target of the personal insults <sup>a</sup>	0.99	0.98
Second reference point of the personal insult <sup>a</sup>	0.99	0.99
General insult to a group of people	0.92	0.84
Group reference of the insults	0.91	0.91
Second group reference of the insults <sup>a</sup>	0.98	0.98
Violence threats	0.96	0.94
Target of the violence threat(s)	0.94	0.94
Negative stereotyping	0.92	0.91
Second negative stereotyping <sup>a</sup>	0.97	0.97
Disinformation/Conspiracy theories	0.87	0.83
Reference point of the disinformation/conspiracy theory	0.87	0.86
Ingroup elevation	0.93	0.85
Inhumane ideology	0.96	0.94

Note.  $N=590$ , two coders, all categories were nominal.

<sup>a</sup>If present, more than one target (or group of targets) could be coded.

One challenge of topic modeling is to find a meaningful number of clusters. Since topic modeling is an unsupervised learning approach, there is no single right solution. To cope with this problem, the following five criteria have been used to determine an appropriate number of clusters: (a) the same number of topics for  $TM_1$  and  $TM_2$ , (b) a meaningful and manageable number of topics, (c) comprehensibility of the topics, (d) standard deviation of the topics' sizes, and (e) (normalized) pointwise mutual information.

## Results

### Results of Manual Quantitative Content Analysis

Addressing RQ1 (extent of explicit/implicit hate speech), we found that almost a quarter (24%,  $n=1,438$ ) of the analyzed 5,981 comments contained at least one instance of explicit or implicit hate speech (see Table 2). In 821 of the comments (13.7%), forms of explicit hate speech were identified (i.e., at least one of the categories personal insult, general insult, or violence threat was coded). Implicit hate speech (i.e., negative stereotyping, disinformation/conspiracy theories, ingroup elevation, and inhumane ideologies) occurred slightly more often and was observed in 928 comments (15.5%).

**Table 2.** Number of Comments Containing Hate Speech.

Comments contained . . .	Absolute	Relative <sup>a</sup> (%)
. . . no hate speech	4,543	76.0
. . . hate speech of at least one type <sup>b</sup>	1,438	24.0
. . . explicit hate speech	821	13.7
. . . implicit hate speech	928	15.5

<sup>a</sup> $n=5,981$ .

<sup>b</sup>Due to the fact that explicit and implicit hate speech can occur in the same comment, numbers of explicit and implicit hate speech do not add up to the overall numbers.

Focusing on RQ2a (forms of hate speech), general insults were the most common form of hate speech and observed in 570 comments: they were included in almost every 10th comment of the entire sample (9.5%) and in more than one-third of all identified hateful comments (39.6%). Disinformation and conspiracy theories followed next and made up 31.8% of all comments with hate speech ( $n=458$ ). Within this category, conspiracy theories ( $n=294$ ) were observed almost twice as often as mere disinformation ( $n=164$ ). In over a quarter of all hateful comments (25.7%), inhuman ideologies were referenced or expressed ( $n=369$ ), with 10.8% relating to National Socialism and 14.9% to White-supremacist ideologies. Violence threats were observed in 221 comments (3.7% total; 15.4% of hateful comments), negative stereotyping in 192 comments (3.2% total, 13.4% of hateful comments), and ingroup elevation was coded for 303 comments (5.1% total, 21.1% of hateful comments). Within our sample, personal insults emerged as the least common form of hate speech ( $n=139$ ), making up only 2.3% of all comments and 9.7% of all hateful comments.

Nevertheless, to answer RQ2b (reference points/targets of hate speech), we analyzed the reference points of these personal insults in more detail. Most personal insults attacked an individual's sexual orientation (32.1%), their ethnicity (27%), their political attitude (10.9%), or referred to an actual or alleged disability (10.2%). Personal insults referring to one's religion, country of origin, gender, or gender identity could only rarely be observed. For the categories general insults, violence threat, negative stereotyping, and disinformation/conspiracy theories, we further analyzed which groups were targeted with hateful sentiments (see Table 3). Jews were by far the most affected group and targets of explicit or implicit hate speech in 478 comments. When Jews were targeted, this happened most often in the context of disinformation/conspiracy theories and general insults. Black people were the second most targeted group in the sample (targeted in 277 comments), with attacks occurring primarily in the context of general insults. Other frequent targets were political opponents (targeted in 238 comments), Muslims (targeted in 148 comments), and the LGBTQ community (targeted in 127 comments).

To identify differences between the three fringe communities, we also conducted the analyses separately for  $r/$



**Table 3.** Targets of Hate Speech Across Different Types of Hate Speech.

Group	General insult	Violence threat	Negative stereotyping	Disinformation/conspiracies	Total
Black people	197	19	39	22	277
Muslims	42	26	34	46	148
Jews	182	41	44	211	478
LGBTQ	99	10	7	11	127
Migrants	7	5	3	4	19
People with disabilities	—	—	—	—	—
Social elites/media	8	3	4	35	50
Political opponents	38	38	52	110	238
Latin Americans	19	2	7	4	32
Women	21	10	18	6	55
Criminals	—	6	—	—	6
Asians	9	—	3	1	13
Rest/undefined	13	58	6	8	85
Total	635	218	217	458	1,528

LGBTQ: lesbian, gay, bisexual, transgender and queer or questioning.

**Table 4.** Amount of Hate Speech on the Studied Communities across Different Types of Hate Speech.

Extent of	r/The_Donald/ <i>n</i> = 1,998		4chan/pol/ <i>n</i> = 1,992		8chan/pol/ <i>n</i> = 1,991	
	Absolute	Relative (%)	Absolute	Relative (%)	Absolute	Relative (%)
	<b>Hate speech total</b>	275	<b>13.8</b>	478	<b>24.0</b>	685
<b>Explicit hate speech</b>	99	<b>5.0</b>	329	<b>16.5</b>	393	<b>19.7</b>
Personal insult	11	0.6	71	3.6	57	2.9
General insult	40	2.0	238	11.9	292	14.7
Violence threat	52	2.6	67	3.4	102	5.1
<b>Implicit hate speech</b>	207	<b>10.4</b>	247	<b>12.4</b>	474	<b>23.8</b>
Negative stereotyping	68	3.4	50	2.5	74	3.7
Disinformation/Conspiracy theory	114	5.7	125	6.3	219	11.0
Ingroup elevation	98	4.9	74	3.7	131	6.6
Inhumane ideology	12	0.6	98	4.9	259	13.0

The\_Donald, 4chan/pol/, and 8chan/pol/. Moving from the more “mainstreamy” r/The\_Donald to the outermost 8chan/pol/, the amount of hate speech increases steadily: While 13.8% of all analyzed comments on r/The\_Donald included at least one form of hate speech, we identified 24% of comments on 4chan/pol/ and even 34.4% of comments on 8chan/pol/ as containing hate speech. As can be inferred from Table 4, the amount of explicit and implicit hate speech also differed between the three communities: Particularly striking here is the low amount of explicit hate speech on r/The\_Donald, which is mainly due to the fact that general insults are much less common than on 4chan/pol/ and 8chan/pol/. Looking more closely at implicit hate speech, we see that 8chan/pol/ emerged as the community with the highest share of such indirect, more veiled forms of hate speech, resulting mainly from the relatively high amount of comments featuring disinformation/conspiracy theories and inhuman ideologies.

### Results of Topic Modeling

To answer RQ3 (topical structure of the coded comments), two topic models (TM<sub>1</sub> and TM<sub>2</sub>) were generated and combined with the results of the manual quantitative content analysis. TM<sub>1</sub> focuses of the entire data set, while TM<sub>2</sub> is restricted to the comments that were identified as containing hate speech. Table 5 shows the topics of TM<sub>1</sub>, their relative distribution between the sources, the absolute number of comments, and the proportion of hate speech. After the evaluation of different numbers of topics, 12 topics turned out to be most appropriate. Overall, the topics can be considered meaningful, and their content meets the expectations for these fringe communities (e.g., focus on political affairs, conspiracy theories, anti-Semitism)<sup>4</sup>. A2–A8 have a thematic focus, while A9, A11, and A12 bundle foreign-language comments. As A9–A12 are relatively small compared to the total number of comments (and

**Table 5.** Topics From TM<sub>1</sub> and Their Frequency Distribution.

Topics of TM <sub>1</sub>	r/The_Donald (%)	4chan/pol (%)	8chan/pol (%)	Absolute (hate speech share)
(A1) Really actually think know something never want certainly obviously though	37.9	30.1	32.0	1935 (14.7%)
(A2) Shit fucking damn dipshit asshole faggot bitch motherfucker dumbass goddamn	27.6	40.4	32.0	1368 (32.7%)
(A3) Government political society ideology people democratic nation economic citizens morality	47.3	28.4	24.4	603 (28.7%)
(A4) John Robert David James Michael Chris Richard Ryan Todd George	45.5	29.0	25.5	479 (17.1%)
(A5) Foods protein nutrient fats diet hormone cholesterol meat vitamins veggies	21.5	40.9	37.6	474 (7.8%)
(A6) Jews Muslims Zionists Arabs Judaism Christians Gentiles Kikes Semitic Goyim	22.4	30.8	46.9	429 (59.4%)
(A7) Poland Germany Europe France British Finland Sweden Russia Italy American	26.0	38.2	35.8	369 (32%)
(A8) Wikileaks FBI CIA FOIA Intel Mossad NVO files gov leaks	39.5	25.3	35.2	162 (10.5%)
(A9) ett drar och handlar speciellt samtliga framtida liknande tror sluta	20.0	29.1	50.9	55 (12.7%)
(A10) xt torrent urn magnet tn ut hd ui ii aws	26.5	20.4	53.1	49 (6.1%)
(A11) erfolg muessen vorausgesetzt betroffenen natuerlich dortigen verbreiten einzigen wahres skeptisch	0	10.3	89.7	29 (6.9%)
(A12) een voor wordt uit het niet gaat zijn krijg terugkeer	10.3	31.0	58.6	29 (44.8%)

consequently less meaningful), they will be excluded from the following analyses.

In general, each topic is equally distributed across the three sources with some noticeable exceptions: 47.3% and 45.5% of the comments from the political topics A3 and A4 originate from r/The\_Donald. Topic A2—consisting exclusively of swear words—can mostly be allocated to 4chan/pol (40.4%) and 8chan/pol (32.0%), which is in line with the results from the manual content analysis. The topic with a focus on anti-Semitism and Islam (A6) also exhibits an unequal distribution: r/The\_Donald’s share is only 22.4%, while 4chan/pol’s share is 30.8% and 8chan/pol’s is 46.9%. In light of the observed hatefulness of 4chan/pol and 8chan/pol, it is remarkable that both are the main origin of the identified topic focusing on nutrition (A5), which might be explained by their broader scope. Focusing on the occurrence of hate speech, the topics A2 (32.7%), A6 (59.4%), and A7 (32.0%) have to be highlighted due to their higher-than-average share of hate. This is not surprising, as the keywords from A2 only contain swear words, A6 covers (anti-)Semitic and Islamic comments, and A7 refers to foreign countries which are often the target of hate due to the alt-rights’ nationalist orientation.

To better understand the clusters/topics in which hate speech occurs, a second topic model (TM<sub>2</sub>) was generated

based on the 1,438 hateful comments only (see Tables 6 and 7). Both models show a similar topical structure and some topics from TM<sub>1</sub> are reflected in TM<sub>2</sub> as well: A1 is similar to H3 (generic topic), A2 to H1 (swear words), A6 to H2 (largely anti-Semitic), and A7 to H4 (foreign affairs). On the contrary, other topics emerged as more fine-grained when only considering hate speech-related comments (TM<sub>2</sub>). A good example is topic A3, which focuses on the government, politics, and society. Hateful comments from this topic can be found, among others, in the topics about US democrats and republicans (H5), political ideology (H9), and finances and taxes (H10).

Tables 6 and 7 depict the topics of TM<sub>2</sub> in combination with the manual analysis to get a deeper understanding of thematic clusters in which the different types of hate speech occur: The first one distinguishes between the different forms of explicit and implicit hate speech, the second one between the different targets of hate speech. Concerning the forms of hate speech, the comments from the topic with swear words (H1) tend to be explicit hate speech, particularly general insults (238 out of 398). In contrast to that, all other topics contain more implicit hate speech—a difference that should not be surprising due to the nature of the topics. What is interesting is the difference between the two (anti-)

**Table 6.** Topics of TM<sub>2</sub> Combined With Forms of Hate Speech From Manual Coding.

Topics of TM <sub>2</sub>	# Comments	Explicit			Implicit			Ingroup elevation
		Personal insult	General insult	Violence threat	Negative stereotyping	Disinformation/ Conspiracy theory	Inhumane ideology	
(H1) Shit fucking bitch asshole motherfucker faggot dipshit dumbass damn dick	396	88	238	54	38	58	50	53
(H2) Jews Goyim Kikes Semitic Ashkenazic Gentiles Zionists Arabs Yids Africans	265	12	138	37	39	110	87	45
(H3) Really actually think want know something going never obviously certainly	250	17	77	40	40	85	65	68
(H4) Poland Germany Europe Russian France British Austria Berlin Soviet American	96	4	29	13	6	27	51	8
(H5) Democrats republicans voters conservatives liberals people electorate government citizens socialists	86	1	21	10	18	40	19	51
(H6) David NWO Donald Hilary FBI CIA NBC James Clinton Kennedy	86	8	16	9	7	35	28	12
(H7) Islam Muslims Allah Quran Koran Christians Mohammedan Infidels Sunni Jihad	80	2	19	15	25	39	12	17
(H8) Murderers terror killing enemy innocents crimes violence deadly civilians horrific	79	2	17	30	11	25	24	22
(H9) Ideology worldview morality political societal dialectics nationalism liberalism dogma religion	43	1	6	3	7	20	18	15
(H10) Tax pay costs government price tariffs economic amount money considerable	41	4	7	8	2	13	8	9
(H11) muessen erfolg verbreiten natuerlich anfang wahres wieso vorausgesetzt irgend gebrauchen	11	0	0	1	0	4	6	3
(H12) een uit gaat voor wordt niet het krijg zijn deze	5	0	2	1	0	2	1	0
Total	1,438	139	570	221	193	458	369	303

**Table 7.** Topics of TM<sub>2</sub> Combined With Targets of Hate Speech From Manual Coding.

Topics of TM <sub>2</sub>	# Comments	Black people	Muslims	Jews	LGBTQ	Migrants	People with disabilities	Elites/ Media	Political opponents	Latin Americans	Women	Criminals	Asians	Rest
(H1) Shit fucking bitch asshole motherfucker faggot dipshit dumbass damn dick	396	104	16	68	87	9	—	11	49	6	27	1	4	28
(H2) Jews Goyim Kikes Semitic Ashkenazic Gentiles Zionists Arabs Yids Africans	265	60	22	230	3	1	—	4	18	8	5	—	2	4
(H3) Really actually think want know something going never obviously certainly	250	48	21	71	19	2	—	12	37	11	13	2	3	19
(H4) Poland Germany Europe Russian France British Austria Berlin Soviet American	96	18	10	22	3	3	—	1	11	3	—	—	—	6
(H5) Democrats republicans voters conservatives liberals people electorate government citizens socialists	86	13	2	7	—	1	—	8	56	2	3	—	—	1
(H6) David NWO Donald Hilary FBI CIA NBC James Clinton Kennedy	86	6	2	25	3	1	—	7	24	—	—	—	1	1
(H7) Islam Muslims Allah Quran Koran Christians Mohammedan Infidels Sunni Jihad	80	7	61	10	6	—	—	2	10	—	—	—	—	6
(H8) Murderers terror killing enemy innocents crimes violence deadly civilians horrific	79	14	6	14	5	—	—	2	20	2	3	2	3	14
(H9) Ideology worldview morality political societal dialectics nationalism liberalism dogma religion	43	2	4	11	1	—	—	3	9	—	4	—	—	2
(H10) Tax pay costs government price tariffs economic amount money considerable	41	3	3	15	—	2	—	—	2	—	—	1	—	4
(H11) muessen erfolg verbreiten natuerlich anfang wahres wieso vorausgesetzt irgend gebrauchen (H12) een uit gaat voor wordt niet het krijg zijn deze	11	—	1	3	—	—	—	—	1	—	—	—	—	—
Total	1,438	279	148	480	127	19	0	50	239	32	55	6	13	85

LGBTQ: lesbian, gay, bisexual, transgender and queer or questioning.

religious topics H2 ([anti-]Semitism) and H7 ([anti-]Islam). While the first one contains many explicit general insults (138 out of 265), the second one has a stronger focus on implicit hate speech, in particular on disinformation (39 out of 80) and negative stereotyping (25 out of 80). Beyond that, H4 and H5 have to be mentioned. H4, the topic about foreign affairs, has its maximum in the category inhuman ideologies (51 out of 96). The topic about US democrats and republicans (H5) exhibits a relatively large number of ingroup elevation (51 out of 86) and disinformation (40 out of 86).

Concerning the targets of hate speech, the automatically generated topics are in line with the manual coding, as shown in Table 7. The (anti-)Semitic and Islamic topic have their maximum in the respective target groups (230 out of 265; 61 out of 80). H4, the topic about US democrats and republicans, mainly contains comments targeting political opponents (56 out of 86). The two more generic topics (H1) and (H3) target a wider range of groups and their distribution is in line with the overall distribution of all topics.

## Discussion

Building on ongoing public debates about alt-right fringe communities—that have been described as “the home of some of the most vitriolic content on the Internet” (Stewart, 2019)—this study investigates whether these public perceptions withstand empirical scrutiny. Focusing on three central alt-right fringe communities on Reddit (r/The\_Donald), 4chan (4chan/pol/), and 8chan (8chan/pol/), we provide a systematic investigation of the extent and nature of both explicit and implicit hate speech in these communities. To do so, we combine a manual quantitative content analysis of user comments ( $N=6,000$ ) with an automated topic modeling approach that offers additional insights into the clusters in which hate speech occurs.

The most obvious finding to emerge from our analysis is that hate speech is prevalent in all three studied communities: In almost a quarter of the sample (24%), at least one instance of explicit or implicit hate speech could be observed. Reflecting results from an automated dictionary-based approach by Hine and colleagues (2017)—who identified 12% of comments on 4chan/pol/ to contain (explicitly) hateful terms—we found that 13.7% of all analyzed comments featured explicit hate speech. However, our manual quantitative content analysis allowed us to also examine the extent of more veiled, indirect forms of hate speech, which was found in 15.5% of all comments. Differences between platforms are in line with the expectations one might have when moving from the more moderate to the more extreme communities: Comparatively, r/The\_Donald featured the lowest amount of hate speech, followed by 4chan/pol/, and 8chan/pol/, suggesting that the “fringier” communities are distinctly more hateful.

Looking more closely at hate speech expression and common targets of hate speech, the results show that general

insults of groups, referencing, or spreading disinformation/conspiracy theories, as well as the expression or glorification of inhuman ideologies such as National Socialism or White supremacy occurred most frequently. The reason for the high incidence of general insults might partly result from including ethnophaulisms and other derogatory terms such as “newfag” and “oldfag” that are regularly used on 4chan and 8chan to refer to new versus experienced users. The observed prevalence of disinformation and conspiracy theories might thus be even more alarming than the use of “plain” insults.

With regard to the social groups affected by hate speech in alt-right fringe communities, our analysis shows that Jews were targeted most often, followed by Black people and political opponents. While Jews were similarly observed as being targets of general insults, they were most often referenced in the context of disinformation and conspiracy theories, which chimes in with the observed extent of National socialist and White-supremacist ideologies in the studied communities. Political opponents are most often referenced within disinformation and conspiracy theories as well, thus reflecting the communities’ close connection to populist attitudes that are associated with the demonization of institutions and political others (see Fawzi, 2019).

The topic models generated on the basis of the sampled user comments are in line with the results of the manual quantitative content analysis and provide additional insights into discussion topics that are likely to feature hate speech. They reflect the extent of (group-related) insults, anti-Semitic and anti-Islamic sentiments, and the strong nationalist orientation of the studied communities. Furthermore, the analysis shows that hate speech—although this might come as no surprise considering our focus on political fringe communities—often occurs in discussions about the government, the (US) political system, religious and political ideologies, or foreign affairs. Subsequent (computational) analyses could take these insights as a starting point to use specific contexts (=topics) for hate speech detection and artificial intelligence (AI) training sets.

Taking a look into potential directions for future studies, hate and antidemocratic content is not only conveyed through text: In an analysis of German hate memes, Schmitt and colleagues (2020) found that memes often display symbols, persons, or slogans known from National Socialism and the Nazi regime. Relatedly, Askanius (2021) traced an adaptation of stylistic strategies and visual aesthetics of the alt-right in the online communication of a Swedish militant neo-Nazi organization. Considering “that the visual form is increasingly used for strategically masking bigoted and problematic arguments and messages” (Lobinger, Krämer, Venema, & Benecchi, 2020, p. 347), and that images and videos tend to develop more virality than mere text (Ling et al., 2021), future studies should focus more strongly on such visual hate speech, which would also more adequately reflect the communication routines of the studied alt-right fringe communities.

Under the guise of “insider jokes,” humor, or memes, it is possible that hate speech is not recognized as such or is perceived as less harmful. Oftentimes, it cannot be judged as unequivocally criminal and is thus not deleted by platforms. Content that—due to this “milder” perception—also finds favor in groups that do not in principle share the hostile ideas behind it is thus increasingly becoming the norm (Fang & Woodhouse, 2017). Accordingly, it can be assumed that the frequent confrontation with hate speech is loosening the boundaries of what can be said and thought, even among initially uninvolved Internet users. This mainstreaming process is described, for example, by Whitney Phillips (2015), who notes the historical transition of hateful, racist memes from fringe communities on the Internet to an increasingly broader public. Sanchez (2020), therefore, warns against a normalization of the “dark humor” that occurs in viral hate memes and calls for critical consideration and research of a possible desensitization to hate and incitement as a consequence. This study adds to this body of literature by providing first evidence that implicit hate speech is as prevalent as explicit hate speech and should thus be considered when analyzing both the extent as well as the potential harm of online hate. In addition, future studies should emphasize the long-term perspective and potential dangers of this development in which mainstreaming would contribute to hate becoming more and more “normal.”

This work has limitations that warrant discussion. First, due to difficulties with the data collection, the initial number of comments on the analyzed communities varied, with 4chan/pol/ having a considerably smaller base of comments to sample from than r/The\_Donald and 8chan/pol/. Moreover, all comments were scraped in April 2019, which might have influenced the results due to specific (political) topics being more or less obtrusive during that time period, possibly also influencing the general amount of hate speech. Second, it should be noted that we did not explicitly exclude hate terms that are part of typical communication norms within the studied communities. Terms such as the mentioned “newfag” were coded as hate speech although they may simply reflect 4chan jargon and are not used with malicious intentions. Nevertheless, we intentionally decided to code it as hate speech as even “normalized” or unintended hate speech can have negative effects (e.g., Burn et al., 2005). Third, our methodology and analysis were focused on textual hate speech, which is why we are unable to account for the amount of hate speech that is transmitted via shared pictures, (visual) memes, or videos. As we have outlined above, it is nevertheless an important endeavor to include the analysis of visual hate speech for which the results of our study might provide a fruitful starting point.

Notwithstanding its limitations, this study provides a first systematic investigation of the extent and nature of hate speech in alt-right fringe communities and shows how widespread verbal hate is on these discussion boards. Further research is needed to confirm and validate our findings,

explore the effects of distinct forms of explicit and implicit hate speech on users, and assess the risks of virtual hate turning into real-life violence.

### Declaration of Conflicting Interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### ORCID iDs

Diana Rieger  <https://orcid.org/0000-0002-2417-0480>

Anna Sophie Kämpel  <https://orcid.org/0000-0001-7184-4057>

### Notes

1. In autumn 2019, 8chan was relaunched as *8kun*, which can be accessed from the Clearnet again. However, the original creator of 8chan, Fredrick Brennan, has not only publicly claimed to regret his creation but also vocally opposed the relaunch of 8chan (Roose, 2019).
2. Due to rate limits and technical hurdles, we were only able to scrape 1,000 comments per day from *4plebs.org*, which is why 4chan/pol/ has (a) overall the smallest initial data set and (b) the longest span of data collection.
3. <https://osf.io/yfxzw/>
4. Exceptions are A1 and A10. A1 is a generic topic containing comments that the algorithm could not assign to more meaningful classes. A10 is the result of comments containing links to file-sharing platforms.

### References

- Askanius, T. (2021). On frogs, monkeys, and execution memes: Exploring the humor-hate nexus at the intersection of neo-nazi and alt-right movements in Sweden. *Television & New Media*, 22(2), 147–165.
- Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit dataset. ArXiv:2001.08435 [Cs]. <http://arxiv.org/abs/2001.08435>
- Ben-David, A., & Matamoros-Fernández, A. (2016). Hate Speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication*, 10, 1167–1193.
- Benikova, D., Wojatzki, M., & Zesch, T. (2018). What does this imply? Examining the impact of implicitness on the perception of hate speech. In G. Rehm & T. Declerck (Eds.), *Language Technologies for the Challenges of the Digital Age* (pp. 171–179). Springer. [https://doi.org/10.1007/978-3-319-73706-5\\_14](https://doi.org/10.1007/978-3-319-73706-5_14)
- Borgeson, K., & Valeri, R. (2004). Faces of hate. *Journal of Applied Sociology*, 21(2), 99–111.
- Brodin, J. (2019, March 20). 4chan, 8chan blocked by Australian and NZ ISPs for hosting shooting video. *Ars Technica*. <https://arstechnica.com/tech-policy/2019/03/australian-and-nz-isps-blocked-dozens-of-sites-that-host-nz-shooting-video/>
- Brown, A. (2018). What is so special about online (as compared to offline) hate speech? *Ethnicities*, 18(3), 297–326.

- Burn, S. M., Kadlec, K., & Rexer, R. (2005). Effects of subtle heterosexism on gays, lesbians, and bisexuals. *Journal of Homosexuality, 49*(2), 23–38.
- Buyse, A. (2014). Words of violence: Fear speech, or how violent conflict escalation relates to the freedom of expression. *Human Rights Quarterly, 36*(4), 779–797.
- Campbell, J. C., Hindle, A., & Stroulia, E. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022.
- Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering, 26*(12), 2928–2941.
- Common Sense. (2018). *Social media, social life. Teens reveal their experiences*. [https://www.common-sense-media.org/sites/default/files/uploads/research/2018\\_cs\\_socialmediasociallife\\_fullreport-final-release\\_2\\_lowres.pdf](https://www.common-sense-media.org/sites/default/files/uploads/research/2018_cs_socialmediasociallife_fullreport-final-release_2_lowres.pdf)
- Conway, M., Macnair, L., & Scrivens, R. (2019). *Right-wing extremists' persistent online presence: History and contemporary trends* (pp. 1–24). International Centre for Counter-Terrorism (ICCT). <https://icct.nl/app/uploads/2019/11/RWEXOnline-1.pdf>
- ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Twelfth International AAAI Conference on Web and Social Media* (pp. 42–51). <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17910>
- Erjavec, K., & Kovačić, M. P. (2012). “You don’t understand, this is a new war!” Analysis of hate speech in news web sites’ comments. *Mass Communication and Society, 15*(6), 899–920.
- Fang, L., & Woodhouse, L. A. (2017, August 25). How white nationalism became normal online. *The Intercept*. <https://theintercept.com/2017/08/25/video-how-white-nationalism-became-normal-online/>
- Fawzi, N. (2019). Untrustworthy news and the media as “enemy of the people?” How a populist worldview shapes recipients’ attitudes toward the media. *The International Journal of Press/Politics, 24*(2), 146–164.
- Forscher, P. S., & Kteily, N. S. (2020). A psychological profile of the alt-right. *Perspectives on Psychological Science, 15*(1), 90–116.
- Gagliardone, I., Gal, D., Alves, T., & Martínez, G. (2015). *Countering online hate speech*. UNESCO.
- Günther, E., & Domahidi, E. (2017). What communication scholars write about: An analysis of 80 years of research in high-impact journals. *International Journal of Communication, 11*, 3051–3071.
- Hajok, D., & Selg, O. (2018). Kommunikation auf Abwegen? Fake news and hate speech in kritischer Betrachtung. *Jugend Medien Schutz-Report, 41*(4), 2–6.
- Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant Behavior, 38*(3), 254–266.
- Hawley, G. (2017). *Making sense of the alt-right*. Columbia University Press.
- Heikkilä, N. (2017). Online antagonism of the alt-right in the 2016 election. *European Journal of American Studies, 12*(2). <https://doi.org/10.4000/ejas.12140>
- Hine, G., Onalapo, J., Cristofaro, E. D., Kourtellis, N., Leontiadis, I., Samaras, R., Stringhini, G., & Blackburn, J. (2017). Kek, cucks, and god emperor Trump: A measurement study of 4chan’s politically incorrect forum and its effects on the web. In S. Gonzalez-Bailon, A. Marwick, & W. Mason (Eds.), *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)* (pp. 92–101). Association for the Advancement of Artificial Intelligence (AAAI).
- Hsueh, M., Yogeewaran, 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.
- Humphrecht, E. (2019). Where “fake news” flourishes: A comparison across four Western democracies. *Information, Communication & Society, 22*(13), 1973–1988.
- Jardine, E. (2019). Online content moderation and the Dark Web: Policy responses to radicalizing hate speech and malicious content on the Darknet. *First Monday, 24*(12). <https://doi.org/10.5210/fm.v24i12.10266>
- Kinney, T. A. (2008). Hate speech and ethno-phaulisms. In W. Donsbach (Ed.), *The international encyclopedia of communication*. Wiley. <https://doi.org/10.1002/9781405186407.wbiech004>
- Kümpel, A. S., & Rieger, D. (2019). *Wandel der Sprach- und Debattenkultur in sozialen Online-Medien. Ein Literaturüberblick zu Ursachen und Wirkungen von inziviler Kommunikation* [The Changing Culture of Language and Debate on Social Media: A Literature Review of the Causes and Effects of Incivil Communication]. Konrad-Adenauer-Stiftung, Landesanstalt für Medien NRW. (2018). *Hate speech und Diskussionsbeteiligung im Internet*.
- Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior, 28*(2), 434–443.
- Ling, C., AbuHilal, I., Blackburn, J., De Cristofaro, E., Zannettou, S., & Stringhini, G. (2021). Dissecting the meme magic: Understanding indicators of virality in image memes. ArXiv:2101.06535 [Cs]. <http://arxiv.org/abs/2101.06535>
- Lobinger, K., Krämer, B., Venema, R., & Benecchi, E. (2020). Pepe—just a funny frog? A visual meme caught between innocent humor, far-right ideology, and fandom. In B. Krämer & C. Holtz-Bacha (Eds.), *Perspectives on populism and the media* (pp. 333–352). Nomos. <https://doi.org/10.5771/9783845297392-333>
- Magu, R., & Luo, J. (2018). *Determining code words in euphemistic hate speech using word embedding networks* [Conference session]. Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Brussels, Belgium.
- Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society, 20*(6), 930–946.
- McNamee, L. G., Peterson, B. L., & Peña, J. (2010). A call to educate, participate, invoke and indict: Understanding the communication of online hate groups. *Communication Monographs, 77*(2), 257–280.
- Meibauer, J. (2013). Hassrede—Von der Sprache zur Politik [Hate Speech—From Language to Politics]. In J. Meibauer (Ed.), *Hassrede/Hate Speech. Interdisziplinäre Beiträge zu einer aktuellen Diskussion* (pp. 1–16). Gießener Elektronische Bibliothek.
- Meza, R. (2016). Hate-speech in the Romanian online media. *Journal of Media Research, 9*(26), 55–77.

- Mikolov, T., Grave, E., Bojanowski, P., Puhersch, C., & Joulin, A. (2019). *Advances in pre-training distributed word representations* [Conference session]. LREC 2018 - 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan.
- Mittos, A., Zannettou, S., Blackburn, J., & De Cristofaro, E. (2019, October 4). "And we will fight for our race!" A measurement study of genetic testing conversations on Reddit and 4chan [Conference session]. Proceedings of the 14th International AAAI Conference on Web and Social Media. ICWSM 2020, Atlanta, GA. <http://arxiv.org/abs/1901.09735>
- Mondal, M., Silva, L. A., Correa, D., & Benevenuto, F. (2018). Characterizing usage of explicit hate expressions in social media. *New Review of Hypermedia and Multimedia*, 24(2), 110–130.
- Nagle, A. (2017). *Kill all normies: The online culture wars from Tumblr and 4chan to the alt-right and Trump*. Zero Books.
- Pacheco, E., & Melhuish, N. (2018). *Online hate speech: A survey on personal experiences and exposure among adult New Zealanders*. Netsafe. <https://www.netsafe.org.nz/wp-content/uploads/2019/11/onlinehatespeech-survey-2018.pdf>
- Phillips, J., & Yi, J. (2018). Charlottesville paradox: The 'liberalizing' alt-right, 'authoritarian' left, and politics of dialogue. *Society*, 55(3), 221–228.
- Phillips, W. (2015). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. MIT Press.
- Prince, M. (2019, August 5). Terminating service for 8chan. *The Cloudflare Blog*. <https://blog.cloudflare.com/terminating-service-for-8chan/>
- Quan, X., Kit, C., Ge, Y., & Pan, S. J. (2015). *Short and sparse text topic modeling via self-aggregation* [Conference session]. IJCAI international joint conference on artificial intelligence, Palo Alto, CA, United States.
- Reicher, S. D., Spears, R., & Postmes, T. (1995). A social identity model of deindividuation phenomena. *European Review of Social Psychology*, 6(1), 161–198.
- Roose, K. (2019). 'Shut the site down,' says the creator of 8chan, a megaphone for gunmen. *The New York Times*. <https://www.nytimes.com/2019/08/04/technology/8chan-shooting-manifesto.html>
- Saha, P., Mathew, B., Garimella, K., & Mukherjee, A. (2021). "Short is the road that leads from fear to hate": Fear speech in Indian WhatsApp groups [Conference session]. Proceedings of the Web Conference 2021
- Sanchez, B. C. (2020). Internet memes and desensitization. *Pathways: A Journal of Humanistic and Social Inquiry*, 1(2), 1–11.
- Schmitt, J. B., Harles, D., & Rieger, D. (2020). Themen, motive und mainstreaming in rechtsextremen online-memes. *Medien & Kommunikationswissenschaft*, 68(1–2), 73–93.
- Stewart, E. (2019, May 3). 8chan, a nexus of radicalization, explained. *Vox*. <https://www.vox.com/recode/2019/5/3/18527214/8chan-walmart-el-paso-shooting-cloudflare-white-nationalism>
- Tuters, M., & Hagen, S. (2020). (((They))) rule: Memetic antagonism and nebulous othering on 4chan. *New Media & Society*, 22(12), 2218–2237.
- Unkel, J. (2021). *tidycomm: Data modification and analysis for communication research* (Version 0.2.1) [Computer software]. <https://CRAN.R-project.org/package=tidycomm>
- Viegas, F., Luiz, W., Canuto, S., Rosa, T., Gomes, C., Ribas, S., Rocha, L., & Gonçalves, M. A. (2019). *Cluwords: Exploiting semantic word clustering representation for enhanced topic modeling* [Conference session]. WSDM 2019—Proceedings of the 12th ACM international conference on web search and data mining, Melbourne, VIC, Australia.
- Wendling, M. (2018). *Alt-right: From 4chan to the White House*. Pluto Press.

### Author Biographies

Diana Rieger (PhD, University of Cologne) is a Professor of Communication Science at the Department of Media and Communication at LMU Munich. Her current work addresses characteristics and effects of hate speech, extremist online communication, and counter voices (e.g., counter narratives and counter speech).

Anna Sophie Kümpel (Dr. rer. soc., LMU Munich) is an Assistant Professor of communication at the Institute of Media and Communication at TU Dresden. Her research interests are focused on media uses and effects, particularly in the context of social media, (incidental exposure to) online news, and media entertainment.

Maximilian Wich (MSc, University of Mannheim) is a PhD student at the Technical University of Munich. His research interests include hate speech detection with machine learning (ML) and explainable artificial intelligence (XAI).

Toni Kiening (BA, LMU Munich) is a graduate from the communication science curriculum at the Department of Media and Communication at LMU Munich.

Georg Groh (Dr. rer. nat., TU Munich) heads the Social Computing research group at the Department of Informatics of the Technical University of Munich. His research focuses on modeling and inferring social context, for example, via ML-based natural language processing.



## B.2 STUDY III

©2020 Association for Computational Linguistics, published under Creative Commons CC-BY 4.0 License<sup>2</sup>.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh (Nov. 2020). “Identifying and Measuring Annotator Bias Based on Annotators’ Demographic Characteristics.” In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, pp. 184–190. DOI: [10.18653/v1/2020.alw-1.21](https://doi.org/10.18653/v1/2020.alw-1.21). URL: <https://www.aclweb.org/anthology/2020.alw-1.21>

---

<sup>2</sup> <https://creativecommons.org/licenses/by/4.0/>

### *Publication Summary*

"Machine learning is recently used to detect hate speech and other forms of abusive language in online platforms. However, a notable weakness of machine learning models is their vulnerability to bias, which can impair their performance and fairness. One type is annotator bias caused by the subjective perception of the annotators. In this work, we investigate annotator bias using classification models trained on data from demographically distinct annotator groups. To do so, we sample balanced subsets of data that are labeled by demographically distinct annotators. We then train classifiers on these subsets, analyze their performances on similarly grouped test sets, and compare them statistically. Our findings show that the proposed approach successfully identifies bias and that demographic features, such as first language, age, and education, correlate with significant performance differences." (Al Kuwatly, Wich, and Groh, 2020, p. 184)

### *Author Contributions*

Maximilian Wich headed the research project. He developed the initial idea, the concept, and the methodology of the study. Furthermore, he wrote a substantial part of the manuscript. Hala Al Kuwatly implemented the code to conduct the experiments as part of her Guided Research supervised by Maximilian Wich. In addition, she wrote a substantial part of the paper. Hala Al Kuwatly and Maximilian Wich share the first authorship. Georg Groh regularly discussed the ideas and concepts with the team and provided feedback on the study.

# Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics

**Hala Al Kuwatly\***  
TU Munich,  
Department of Informatics,  
Germany  
hala.kuwatly@tum.de

**Maximilian Wich\***  
TU Munich,  
Department of Informatics,  
Germany  
maximilian.wich@tum.de

**Georg Groh**  
TU Munich,  
Department of Informatics,  
Germany  
grohg@in.tum.de

## Abstract

Machine learning is recently used to detect hate speech and other forms of abusive language in online platforms. However, a notable weakness of machine learning models is their vulnerability to bias, which can impair their performance and fairness. One type is annotator bias caused by the subjective perception of the annotators. In this work, we investigate annotator bias using classification models trained on data from demographically distinct annotator groups. To do so, we sample balanced subsets of data that are labeled by demographically distinct annotators. We then train classifiers on these subsets, analyze their performances on similarly grouped test sets, and compare them statistically. Our findings show that the proposed approach successfully identifies bias and that demographic features, such as first language, age, and education, correlate with significant performance differences.

## 1 Introduction

According to the online harassment report published by Pew Research Center, "four-in-ten Americans have personally experienced online harassment, and 62% consider it a major issue." (Duggan, 2017, p.3). Online environments such as social media and discussion forums have created spaces for people to express their opinions and viewpoints, but this comes at the cost of hateful, offensive, and abusive content. Moderating this content manually requires a lot of staff and large amounts of hand-curated policies, which generated much interest in automatic content moderation systems that make use of recent advances in machine learning (Schmidt and Wiegand, 2017).

One challenge of training machine learning systems is the demand for large amounts of labeled

data. Hence, many researchers use crowdsourcing platforms to annotate their data sets (Davidson et al., 2017; Founta et al., 2018; Vidgen and Derczynski, 2020), although having expert annotators has proven to improve the quality of annotations (Waseem, 2016). Such crowdsourcing approaches, however, exposes hate speech detection systems to annotator bias. Hateful behavior can take many forms (Waseem et al., 2017), making it harder to obtain a clean, common definition of hate speech, and resulting in subjective and biased annotations. Biases in the annotations are then absorbed and reinforced by the machine learning models, causing systematically unfair systems (Bender and Friedman, 2018). Therefore, it is not surprising that a large body of work has identified and mitigated this bias (Bender and Friedman, 2018; Bountouridis et al., 2019; Dixon et al., 2018).

We already know that people with particular demographic characteristics (e.g., black, disabled, or younger people) become more frequently targets of hate (Vidgen et al., 2019b). An aspect that is sparsely investigated in this context is the relation between annotators' demographic features and a potential bias in the data set. We want to fill this gap by addressing the following research question:

How do annotators' demographic features such as gender, age, education and first language impact their annotations of hateful content?

To answer this question, we conduct the following exploratory study: We sample balanced subsets of data that are labeled by demographically distinct annotators. We then train classifiers on these subsets, analyze their performances on similarly split test sets, and compare them statistically.

## 2 Related work

Since unintended bias in hate speech datasets can impair the model's performance (Waseem, 2016)

---

\*These authors contributed equally to this work.

and fairness (Vidgen et al., 2019a; Dixon et al., 2018), a lot of recent work has been done to investigate this phenomenon (Wiegand et al., 2019; Kim et al., 2020).

Some work examined racial bias (Sap et al., 2019; Davidson et al., 2019; Xia et al., 2020), others explored gender bias (Gold and Zesch, 2018), aggregation bias (Balayn et al., 2018) and political bias (Wich et al., 2020b). The type of bias we are examining in this study is the annotator bias. Waseem (2016) studied the influence of annotator expertise on classification models and found that systems trained on expert annotations outperform those trained on amateur annotations, confirming and extending the results from Ross et al. (2017). Geva et al. (2019) showed that model performance improves when exposed to annotator identifiers, which suggests that annotator bias needs to be considered when creating hate speech models. Salminen et al. (2018) studied the difference between annotations of crowd workers from 50 countries and found those differences highly significant. Binns et al. (2017) examined the effect of the gender of the annotators on the performance of classifiers. Wich et al. (2020a) studied the similarities in the behaviour of the annotators to reveal biases that they bring into the data.

To the best of our knowledge, no one has developed a method to identify annotator bias based on multiple demographic characteristics of the annotators and measure its impact on the classification performance.

### 3 Data

We used the personal attack corpora from Wikipedia’s Detox project (Wulczyn et al., 2017), which contains 115,864 labeled comments from Wikipedia on whether the comment contains a form of personal attack. The labels are the following (Wikimedia, n.d.):

- Quoting attack: Indicator for whether the annotator thought the comment is quoting or reporting a personal attack that originated in a different comment.
- Recipient attack: Indicator for whether the annotator thought the comment contains a personal attack directed at the recipient of the comment.
- Third party attack: Indicator for whether the

Feature	Trainset size	Testset size	Total size
Gender	4,401	1,100	5,501
First language	2,038	509	2,547
Age group	6,782	1,696	8,478
Education	3,174	794	3,968

Table 1: Number of comments in each demographic feature’s datasets

annotator thought the comment contains a personal attack directed at a third party.

- Other attack: Indicator for whether the annotator thought the comment contains a personal attack but is not quoting attack, a recipient attack or third party attack.
- Attack: Indicator for whether the annotator thought the comment contains any form of personal attack. (Wikimedia, n.d.)

For our study, we used the attack label as the classification target label, not taking into consideration the other labels.

The comments were labeled by 4,053 crowdworkers. For 2,190 of them, we have the demographic information. For each of these annotators we have the following demographic features:

- Gender: ‘male’ or ‘female’
- English first language: ‘1’ or ‘0’; ‘1’ = annotator’s first language is English
- Age group: ‘Under 18’, ‘18-30’, ‘30-45’, ‘45-60’, ‘Over 60’. Since annotators are not equally distributed across age groups (see distribution plot in the appendix), we changed the grouping to ‘Under 30’ and ‘Over 30’.
- Education (highest obtained education level): ‘none’, ‘some’, ‘hs’, ‘bachelors’, ‘masters’, ‘doctorate’, ‘professional’. ‘hs’ is short for high school. Since annotators are not equally distributed across education levels (see distribution plot in the appendix), we changed the grouping to ‘Below hs’ (includes hs) and ‘Above hs’.

### 4 Methodology

We address the research question by training classification models on data from demographically distinct groups and comparing their performances<sup>1</sup>.

<sup>1</sup>Code available on GitHub: <https://github.com/mawic/annotator-bias-demographic-characteristics>

The hypothesis is that a statistically significant difference between the classifiers’ performances indicates an annotator bias related to the studied demographic feature.

In the first step, we group the annotators by their demographic features, such as gender, age, education level, and native language. For each of those features, we create  $m + 1$  datasets where  $m$  is the number of different values a demographic feature can take, e.g. for gender  $m$  could be equal to 2 if we only consider male and female annotators. All datasets have the same comments, but with different labels aggregated from annotators belonging to each different group. The additional dataset (+1) has labels aggregated from annotators belonging to all groups. It serves as a control group. We call this dataset the mixed dataset. We measured the inter-rater agreement within each group using Krippendorff’s alpha (Hayes and Krippendorff, 2007).

In the second step, we split the datasets into train and test sets, and train 20 classifiers for each group on the group’s training set and report F1 scores for all test sets. We train 20 classifiers to get multiple data points for each group’s classifier and then apply the Kolmogorov-Smirnov test to examine whether they are significantly different<sup>2</sup>. The null hypothesis in this context is that the two samples are drawn from the same distribution. If we can reject the null hypothesis ( $p < 0.05$ ) for a certain demographic feature, this will be evidence that annotators belonging to different groups of feature values hold different norms and are bringing in different biases into their annotations.

Concerning the classification model, we chose to make use of recent advancements in transfer learning and employ DistilBERT as a classifier due to the limited number of data points annotated by each group. DistilBERT (Sanh et al., 2019) is a smaller and faster distilled version of BERT (Devlin et al., 2018). In the context of abusive language detection, it provides a comparable performance (Vidgen et al., 2020). We used the base uncased version of DistilBERT (distilbert-base-uncased) with a maximum sequence length of 100, a learning rate of  $5 \times 10^{-6}$ , and 1cycle learning rate policy (Smith, 2018) and trained each classifier for 2 epochs.

---

<sup>2</sup>We trained 20 classifiers only for practical constraints.

## 4.1 Data split

To ensure the comparability of the classifiers, it is necessary to compile the training and test sets in the right way. Therefore, we define the following 2 conditions for selecting the comments: (1) All data sets of one feature contain the same comments. (2) At least 6 annotators from each demographic group annotated the comment. In the case of the gender group, that means a selected comment was annotated by at least 6 male and 6 female annotators.

For each demographic feature, we create 3 training and test set combinations. In the first one, the labels are taken from a random set of 6 annotators belonging to the first demographic group (e.g., males). In the second one, the labels of the comments are taken from a random set of 6 annotators belonging to the second demographic group (e.g., females). The third train and test sets are mixed: the labels of the comments are taken from a random set of 3 annotators belonging to the first demographic group and 3 annotators belonging to the second demographic group. While the subset of comments stays unchanged, for each of the 20 classifiers we sample the annotations of different random annotators. Data sets’ sizes can be found in Table 1.

We also performed the same experiments without the limitation of sharing the same comments in the data sets of each feature, in order to increase the size of comments in the splits. Results were very similar to our shared comments experiments.

## 5 Results

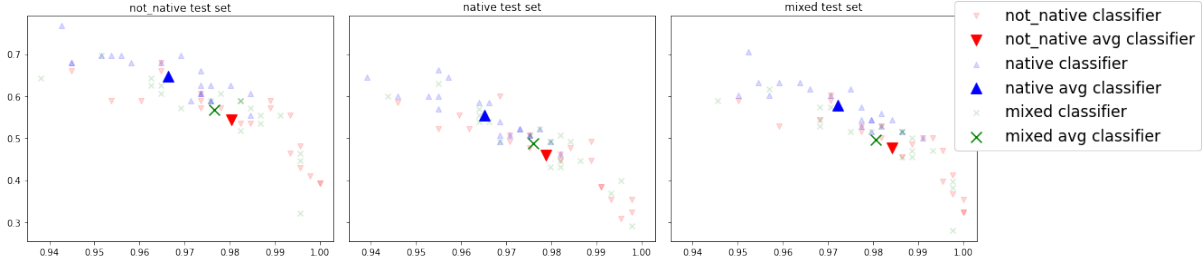
In this section, we report the results of our experiments for each demographic feature. The results comprise the inter-rater agreement of the annotators in the different groups, the averaged F1 scores of the trained classifiers, the sensitivity and specificity of the classifiers as charts, and the p-values generated by the Kolmogorov-Smirnov tests.

### 5.1 Gender

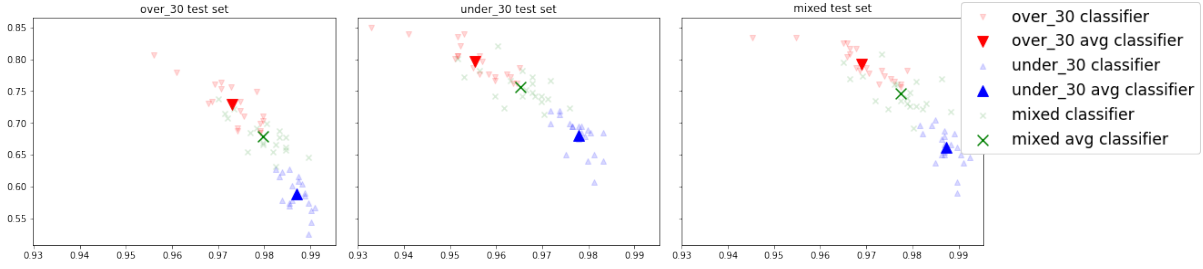
In regards to gender, we could not find evidence of any significant difference between male and female classifiers. Although the inter-rater agreement is significantly lower for females (0.45) than for males (0.51) (Table 4), the average F1 scores of the 20 classifiers trained for each group show no significant difference (Table 2). When analyzing the sensitivity and specificity graphs in Figure 1a,



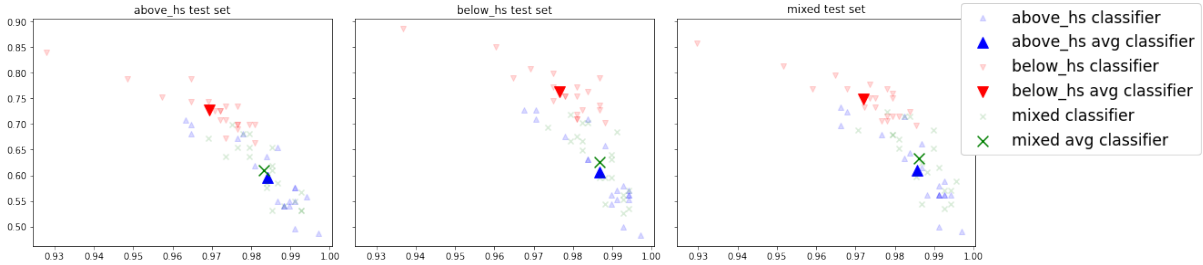
(a) Gender groups classifiers evaluated on gender groups test sets



(b) Language groups classifiers evaluated on language groups test sets



(c) Age groups classifiers evaluated on age groups test sets



(d) Education groups classifiers evaluated on education groups test sets

Figure 1: The x-axes are the specificity of the classifiers, and the y-axes are the sensitivity. Each transparent dot represents the specificity and sensitivity of each of the 20 classifiers trained for each group on the respective train set (dot marker) and evaluated on the respective test set (sub-figures). The opaque dots represent the average values.

one can also see no significant pattern or trend. The p-value resulting from the Kolmogorov-Smirnov test applied on the F1 scores of the 20 male classifiers and 20 female classifiers evaluated on the mixed test set is 0.83 (Table 3). Since it is larger than 0.05, we cannot conclude that a significant difference between the male and female classifier exists.

## 5.2 First Language

Our experiments on first language classifiers resulted in the following observations:

1. Classifiers trained on native-labeled data have a notably higher F1 score (Table 2) and are also more sensitive to all test sets (the blue triangles in Figure 1b), which suggests that they are particularly better at classifying comments

testset \ trainset	male	female	mixed
male	0.850	0.855	0.829
female	0.846	0.859	0.838
mixed	<b>0.856</b>	<b>0.862</b>	<b>0.848</b>
	native	not native	mixed
native	<b>0.814</b>	<b>0.818</b>	<b>0.816</b>
not native	0.768	0.786	0.764
mixed	0.783	0.778	0.772
	under 30	over 30	mixed
under 30	0.853	0.833	0.863
over 30	0.858	<b>0.870</b>	<b>0.883</b>
mixed	<b>0.860</b>	0.860	0.879
	below hs	above hs	mixed
below hs	<b>0.885</b>	<b>0.861</b>	<b>0.873</b>
above hs	0.839	0.830	0.839
mixed	0.847	0.836	0.850

Table 2: Average F1 scores of the classifiers.

Feature	p-value
Gender	$8.3 \times 10^{-1}$
First Language	$1.0 \times 10^{-3}$
Age group	$1.1 \times 10^{-8}$
Education	$1.4 \times 10^{-7}$

Table 3: Results of the Kolmogorov-Smirnov test, inputs to the tests are the F1 scores of the 20 classifiers evaluated on the mixed test set of each feature.

that contain personal attack.

- Classifiers trained on only non-native-labeled data perform almost as good as the baseline (classifier trained on mix-labeled data) (Table 2).
- We found very minor disparities in the specificity of both classifiers (Figure 1b).

The result of the Kolmogorov-Smirnov test on native and non-native classifiers is a p-value of  $1.0 \times 10^{-3}$  (Table 3), thus we can reject the null hypothesis and conclude that a significant difference does exist between them.

### 5.3 Age group

Our experiments resulted in the following observations:

- Classifiers trained on over-30-labeled data have higher F1 scores than classifiers trained on under-30 labeled data on all test sets. They are however comparable to the baseline (classifier trained on mix-labeled data) (Table 2).
- All classifiers are less sensitive to over-30-labeled test set (Figure 1c), which might suggest that it contains harder examples that all classifiers failed to correctly classify.

Feature	Group	Inter-rater Agreement
Gender	Male	0.51
	Female	0.45
	Mixed	0.48
English	Native	0.46
	Not native	0.50
	Mixed	0.48
Age group	Under 30	0.47
	Over 30	0.50
	Mixed	0.48
Education	Below hs	0.49
	Above hs	0.48
	Mixed	0.48

Table 4: Inter-rater agreement for all groups

The Kolmogorov-Smirnov test on the results of the two classifiers produces a p-value of  $1.1 \times 10^{-8}$  (Table 3), thus we can reject that they come from the same distribution and conclude that a significant difference does exist between them.

### 5.4 Education

Our experiments resulted in the following observations:

- The F1 scores of the classifiers trained on below-hs-labeled data are higher than scores of classifiers trained on above-hs-labeled data on all test sets (Table 2).
- Classifiers trained on below-hs-labeled data have a comparable specificity to the other classifiers but with a notably higher sensitivity on all test sets. (Figure 1d).

The Kolmogorov-Smirnov test with a p-value of  $1.4 \times 10^{-7}$  (Table 3) also shows that there exists a significant difference between the two groups.

## 6 Discussion

In light of our results, we can conclude that the gender of the annotator does not bring a significant bias in annotating personal attacks in the studied dataset. However, when Binns et al. (2017) explored the role of gender in *offensive* content annotations, they established a distinguishable difference between males and females. We think this is related to the nature of the annotation task itself. To investigate other tasks, our approach can further be applied in future work on the other data sets provided by Wikipedia’s Detox project (Wulczyn et al., 2017) such as aggressiveness and toxicity to investigate the effects of gender for those tasks.

When it comes to the first language of the annotators, it seems that native English speakers are gen-

erally better at identifying personal attacks in comments. The results also suggest that non-natives could not capture attack in comments that natives found to contain attack.

In addition, age groups and education levels of the annotators also seem to play a notable role in how attacks are perceived. Training a classifier on aggregated labels from all groups, even if the data is balanced between groups, does not seem to be fair to all groups involved.

Although we have only explored the demographic features provided by the data set and grouped some of them for reasons dictated by the data size, we think other features (e.g., race, ethnicity, and political orientation), different within feature groupings and feature intersections might produce new biases. While exploring all possible demographic features prior to building models is simply infeasible, the set of studied features can be determined per task.

Our approach demonstrated how particular training sets labeled by different groups of people can be used to identify and measure bias in data sets. These biases are never constant or static even within one group, for what counts as hateful is always subjective. In consequence, having only one version of ground truth is bound to produce biased systems. It is inevitable that training models on biased datasets produces systems that amplify those biases, whether these biases are exclusionary, prejudicial, or historical. Therefore and due to the conflicting and ever-changing definitions of hate speech among communities, we urge researchers in the hate speech domain to examine their data sets closely and thoroughly in order to understand their limitations and consequences.

## 7 Conclusion

This work explored bias in hate speech classification models where the task is inherently controversial and annotators' demographic data might influence the labels. We demonstrate how particular demographic features might bias the models in ways that are important to look into prior to using such models in production. We explored the performance of classification models trained and tested on different training and test data splits, in order to identify the fairness of these classifiers and the biases they absorb. We hope that our proposed method for identifying and measuring annotator bias based on annotators' demographic characteris-

tics will help to build fairer hate speech classifiers.

## Acknowledgments

This research has been partially funded by a scholarship from the Hanns Seidel Foundation financed by the German Federal Ministry of Education and Research.

## References

- Agathe Balayn, Panagiotis Mavridis, Alessandro Bozzon, Benjamin Timmermans, and Zoltán Szilávik. 2018. Characterising and mitigating aggregation-bias in crowdsourced toxicity annotations. In *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management*, volume 2276. CEUR.
- Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *International conference on social informatics*, pages 405–415. Springer.
- Dimitrios Bountouridis, Mykola Makhortykh, Emily Sullivan, Jaron Harambam, Nava Tintarev, and Claudia Hauff. 2019. Annotating credibility: Identifying and mitigating bias in credibility datasets.
- Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.
- Thomas Davidson, Dana Warmley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.
- Maeve Duggan. 2017. *Online Harassment 2017*. Pew Research Center.



- Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *arXiv preprint arXiv:1802.00393*.
- Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*.
- Michael Wojatzki Tobias Horsmann Darina Gold and Torsten Zesch. 2018. Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments.
- Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.
- Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets. *arXiv preprint arXiv:2005.05921*.
- Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.
- Joni Salminen, Fabio Veronesi, Hind Almerkhi, Soon-Gvo Jung, and Bernard J Jansen. 2018. Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 88–94. IEEE.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.
- Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International workshop on natural language processing for social media*, pages 1–10.
- Leslie N Smith. 2018. A disciplined approach to neural network hyper-parameters: Part 1—learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*.
- Bertie Vidgen, Austin Botelho, David Broniatowski, Ella Guest, Matthew Hall, Helen Margetts, Rebekah Tromble, Zeerak Waseem, and Scott Hale. 2020. [Detecting east asian prejudice on social media](#).
- Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *arXiv preprint arXiv:2004.01670*.
- Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019a. Challenges and frontiers in abusive content detection. Association for Computational Linguistics.
- Bertie Vidgen, Helen Margetts, and Alex Harris. 2019b. How much online abuse is there? a systematic review of evidence for the uk. *The Alan Turing Institute*.
- Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proc. 1st Workshop on NLP and Computational Social Science*, pages 138–142.
- Zeerak Waseem, Thomas Davidson, Dana Warmusley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.
- Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020a. Investigating annotator bias with a graph-based approach. In *Proc. 4th Workshop on Online Abuse and Harms*.
- Maximilian Wich, Jan Bauer, and Georg Groh. 2020b. Impact of politically biased data on hate speech classification. In *Proc. 4th Workshop on Online Abuse and Harms*.
- Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608.
- Wikimedia. n.d. Research:detox/data release. [https://meta.wikimedia.org/wiki/Research:Detox/Data\\_Release](https://meta.wikimedia.org/wiki/Research:Detox/Data_Release).
- Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399.
- Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. *arXiv preprint arXiv:2005.12246*.

## B.3 STUDY IV

©2020 Association for Computational Linguistics, published under Creative Commons CC-BY 4.0 License<sup>3</sup>.

Maximilian Wich, Hala Al Kuwatly, and Georg Groh (Nov. 2020). “Investigating Annotator Bias with a Graph-Based Approach.” In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, pp. 191–199. DOI: [10.18653/v1/2020.alw-1.22](https://doi.org/10.18653/v1/2020.alw-1.22). URL: <https://www.aclweb.org/anthology/2020.alw-1.22>

---

<sup>3</sup> <https://creativecommons.org/licenses/by/4.0/>

### *Publication Summary*

"A challenge that many online platforms face is hate speech or any other form of online abuse. To cope with this, hate speech detection systems are developed based on machine learning to reduce manual work for monitoring these platforms. Unfortunately, machine learning is vulnerable to unintended bias in training data, which could have severe consequences, such as a decrease in classification performance or unfair behavior (e.g., discriminating minorities). In the scope of this study, we want to investigate annotator bias—a form of bias that annotators cause due to different knowledge in regards to the task and their subjective perception. Our goal is to identify annotation bias based on similarities in the annotation behavior from annotators. To do so, we build a graph based on the annotations from the different annotators, apply a community detection algorithm to group the annotators, and train for each group classifiers whose performances we compare. By doing so, we are able to identify annotator bias within a data set. The proposed method and collected insights can contribute to developing fairer and more reliable hate speech classification models." (Wich, Al Kuwatly, and Groh, 2020, p. 191)

### *Author Contributions*

Maximilian Wich headed the research project. He developed the initial idea, the concept, and the methodology of the study. Furthermore, he implemented the experiments and wrote the paper. Hala Al Kuwatly provided feedback on the manuscript. Georg Groh regularly discussed the ideas and concepts with the team and provided feedback on the manuscript.

# Investigating Annotator Bias with a Graph-Based Approach

**Maximilian Wich**  
TU Munich,  
Department of Informatics,  
Germany  
maximilian.wich@tum.de

**Hala Al Kuwatly**  
TU Munich,  
Department of Informatics,  
Germany  
dehala.kuwatly@tum.de

**Georg Groh**  
TU Munich,  
Department of Informatics,  
Germany  
grohg@in.tum.de

## Abstract

A challenge that many online platforms face is hate speech or any other form of online abuse. To cope with this, hate speech detection systems are developed based on machine learning to reduce manual work for monitoring these platforms. Unfortunately, machine learning is vulnerable to unintended bias in training data, which could have severe consequences, such as a decrease in classification performance or unfair behavior (e.g., discriminating minorities). In the scope of this study, we want to investigate annotator bias — a form of bias that annotators cause due to different knowledge in regards to the task and their subjective perception. Our goal is to identify annotation bias based on similarities in the annotation behavior from annotators. To do so, we build a graph based on the annotations from the different annotators, apply a community detection algorithm to group the annotators, and train for each group classifiers whose performances we compare. By doing so, we are able to identify annotator bias within a data set. The proposed method and collected insights can contribute to developing fairer and more reliable hate speech classification models.

## 1 Introduction

A massive problem that online platforms face nowadays is online abuse (e.g., hate speech against women, Muslims, or African Americans). It is a severe issue for our society because it can cause more than poisoning the platform's atmosphere. For example, [Williams et al. \(2020\)](#) showed a relation between online hate and physical crime.

Therefore, people have started to develop systems to automatically detect hate speech or abusive language. The advances in machine learning and deep learning have improved these systems tremendously, but there is still much space for enhancements because it is a challenging and complex task

([Fortuna and Nunes, 2018](#); [Schmidt and Wiegand, 2017](#)).

A weakness of these systems is their vulnerability to unintended bias that can cause an unfair behavior of the systems (e.g., discrimination of minorities) ([Dixon et al., 2018](#); [Vidgen et al., 2019](#)). Researchers have identified different types and sources of bias that can influence the performance of hate speech detection models. [Davidson et al. \(2019\)](#), for example, investigated racial bias in hate speech data sets. [Wiegand et al. \(2019\)](#) showed that topic bias and author bias of data sets could impair the performance of hate speech classifiers. [Wich et al. \(2020\)](#) examined the impact of political bias within the data on the classifier's performance. To mitigate bias in training data, [Dixon et al. \(2018\)](#) and [Borkan et al. \(2019\)](#) developed an approach.

Another type of bias that caught researchers' attention is annotator bias. It is caused by the subjective perception and different knowledge levels of annotators regarding the annotation task ([Ross et al., 2017](#); [Waseem, 2016](#); [Geva et al., 2019](#)). Such a bias could harm the generalizability of classification models ([Geva et al., 2019](#)). Especially in the context of online abuse and hate speech, it can be a severe issue because annotating abusive language requires expert knowledge due to the vagueness of the task ([Ross et al., 2017](#); [Waseem, 2016](#)). Nevertheless, due to the limited resources and the demand for large datasets, annotating is often outsourced to crowdsourcing platforms ([Vidgen and Derczynski, 2020](#)). Therefore, we want to investigate this phenomenon in our paper. There is already research concerning annotator bias in hate speech and online abuse detection. [Ross et al. \(2017\)](#) examined the relevance of instructing annotators for hate speech annotations. [Waseem \(2016\)](#) compared the impact of amateur and expert annotators. One of their findings was that a system trained with data

labeled by experts outperforms one trained with data labeled by amateurs. Binns et al. (2017) investigated whether there is a performance difference between classifiers trained on data labeled by males and females. Al Kuwatly et al. (2020) extended this approach and investigated the relevance of annotators’ educational background, age, and mother tongue in the context of bias. Sap et al. (2019) examined racial bias in hate speech data sets and its impact on the classification performance. To the best of our knowledge, no one has investigated annotator bias by identifying patterns in the annotation behavior through an unsupervised approach. That is why we address the following research question in the paper: Is it possible to identify annotator bias purely on the annotation behavior using graphs and classification models?

Our contribution is the following:

- A novel approach for grouping annotators according to their annotations behavior through graphs and analyzing the different groups in order to identify annotator bias.
- A comparison of different weight functions for constructing the annotator graph modeling the annotator behavior.

## 2 Data

For our study, we use the Personal Attacks corpora from the Wikipedia Detox project (Wulczyn et al., 2017). It contains 115,864 comments from English Wikipedia that were labeled whether they comprise personal attack or not. In total, there are 1,365,217 annotations provided by 4,053 annotators from the crowdsourcing platform Crowdfunder — approximately 10 annotations for each comment. Each annotation consists of 5 categories distinguishing between different types of attack: *quoting\_attack*, *recipient\_attack*, *third\_party\_attack*, *other\_attack*, and *attack*. In our experiments, we only use the 5<sup>th</sup> category (*attack*) because it covers a broader range than the other labels. Its value is 1 if ”the comment contains any form of personal attack” (Wikimedia, n.d.). Otherwise it is 0. The corpora also contain demographic information (e.g., gender, age, and education) of 2,190 annotators. But this data is not relevant to our study.

## 3 Methodology

Our approach is to group annotators according to their annotation behavior and analyze perfor-

mance of classification models trained on annotations from these groups. To do so, we firstly group the annotators according to their annotation behavior using a graph. Secondly, we split the data set by the groups and their respective annotations. Thirdly, we train classifiers for each annotator group and then compare their performances. The reader can find a detailed description of the steps in the following<sup>1</sup>:

### Creating Annotator Graph

In the first step, we create an undirected unweighted graph to model the annotation behavior of the annotators (e.g., how similar the annotations of two annotators are). Each node represents an annotator. An edge between two nodes exists if both annotators annotate at least one same data record. Additionally, each edge has a weight that models the similarity between the annotations of the data records. To calculate the weight, we selected four functions that we will compare:

1. **Agreement Rate:** It is the percentage in which both annotators agree on the annotation for a data record:

$$a = \frac{n_{agree}}{n_{agree} + n_{disagree}}$$

where  $n_{agree}$  is the number of data records that both annotated and assigned the same labels to and  $n_{disagree}$  is the number of data records that both annotated and assigned different labels.

2. **Cohen’s kappa (Cohen, 1960):** It is often used as a measure for inter-rater reliability.

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

where  $p_0$  is the ”proportion of observed agreements” (Sim and Wright, 2005, p.258) among the data records annotated by both annotators and  $p_e$  is ”proportion of agreements expected by chance” (Sim and Wright, 2005, p.258) among the records. The range of  $\kappa$  is between  $-1$  and  $+1$ .  $+1$  corresponds perfect agreement;  $\leq 0$  means agreement at chance or no agreement (Cohen, 1960). If both annotators select the same label for all records,  $\kappa$  is not

<sup>1</sup>Code available on GitHub: <https://github.com/mawic/graph-based-method-annotator-bias>

defined. In this case, we remove the edge. An alternative would be to keep the edge and assign 1. But we rejected this idea because of the following consideration. Let us assume that we have 4 annotators (A,B,C, and D). A and B assigned the same label to the same comment. C and D assigned the same labels to the same 20 comments. In both cases,  $\kappa$  is not defined. Assigning the same value (e.g., 1) to both edges would weigh both equally. But the edge between C and D should receive a higher weight because the agreement between A and B could be a coincidence.

3. **Krippendorff's alpha** (Krippendorff, 2004): It is another inter-rater reliability measure, which is defined as follows:

$$\alpha = 1 - \frac{D_0}{D_e}$$

”where  $D_0$  is the observed disagreement among values assigned to units of analysis [...] and  $D_e$  is the disagreement one would expect when the coding of units is attributable to chance rather than to the properties of these units” (Krippendorff, 2011, p.1). Further details of the calculation are provided by Krippendorff (2011). Similar to  $\kappa$ ,  $\alpha$  is not defined if the annotators choose the same label for all records. We handle this case in the same way as above.

4. **Heuristic**: To overcome the undefined issue, we define a heuristic weight function taking the relative agreement rate and the number of commonly annotated data records (overlap) between two annotators into account. The function is defined by four boundary points:

- The maximum weight (1.0) is reached, if two annotators commonly annotated  $n$  data records and agree on all annotations.  $n$  is the maximal number of data records that is commonly annotated by two annotators and is defined by the data set.
- The minimum weight (0) is reached, if two annotators commonly annotated  $n$  data records and disagree on all annotations.
- A weight that is 20% larger than the minimum weight (0.2) is reached, if two an-

notators commonly annotated only one data record and disagree.

- A weight that is 60% larger than the minimum weight (0.6) is reached, if two annotators commonly annotated only one data record and agree.

The transition between the four boundary points is gradually calculated. The algorithm can be found in the appendix. The purpose of the approach is to consider the overlap besides the agreement rate because the larger the overlap the more reliable is the agreement rate. Cohen's alpha and Krippendorff's alpha provide this, but their weakness is the undefined issue, which is a realistic scenario for our annotation task.

All weight functions are normalized between 0 and 1 to make the results comparable, if they are not already in this range.

### Detecting Annotator Groups

The goal of the next step is to group the annotators according to their annotation behavior. For this purpose, we apply the Louvain method, an unsupervised algorithm for detecting communities in a graph (Blondel et al., 2008). After that, we filter the communities with at least 250 members. Otherwise, the groups do not comprise enough data records that were annotated by their members in order to train a classification model.

### Splitting Data According to Groups

After detecting the groups, we split the comments and annotations according to the groups. For each weight function and the corresponding graph, we do the following: We select those comments that were annotated by at least one member of every group. For each group, we create a data set containing these comments and the annotations from the group's members. The label for each comment is the majority vote of the group's annotators. In addition, we create a further data set that serves as a baseline and is called group 0 for all experiments. The data set contains the same comments, but the labels are the results of all 4,053 annotators. After that, all data sets for a weight function are split in a training and test set in the same manner to ensure the comparability of the data sets. This is done for each of the four weight functions.

Weight function	Agreement Rate	Cohen’s kappa	Krippendorff’s alpha	Heuristic Function
Number of nodes	4,053	4,053	4,053	4,053
Number of edges	1,560,078	691,229	691,229	1,560,078
Average degree	769.8	341.1	341.1	769.8
Density	0.190	0.084	0.084	0.190
Connected componets	1	5	5	1
Distribution of edge weights				

Table 1: Graph metrics

### Training Classification Models for Groups and Comparing Their Performances

For the classification model, we use a pre-trained DistilBERT that we fine-tune for our task (Sanh et al., 2019). It is smaller and faster to train than classical BERT, but it provides comparable performance (Sanh et al., 2019). In the context of abusive language detection, it shows a similar performance like larger BERT models (Vidgen et al., 2020). Since we need to train several models for different weight functions and groups, we choose the lighter model.

The basis of our classification model is the pre-trained `distilbert-base-uncased`, which is the distilled version of `bert-base-uncased`. It has 6 layers, a hidden size of 768, 12 self-attention heads, and 66M parameters. To fine-tune the model for our task, we apply the 1cycle learning rate policy suggested by Smith (2018) with a learning rate of  $5e-6$  for 2 epochs. The batch size is 64 and the size of the validation set is 10% of the training set. Furthermore, we limit the number of input tokens to 150. The task that DistilBERT is fine-tuned for is to distinguish between the labels "ATTACK" and "OTHER".

After training the models, we compare their performances (F1 macro). For this purpose, each model is evaluated on its own test set and the one from the other groups including group 0, which represents all annotators. Instead of reporting the

F1 score, we report them relatively to our baseline (group 0) because it allows a better comparison of the results. Additionally, the actual F1 score are not relevant for this analysis.

## 4 Results

The experiments show that our proposed method enables the grouping of annotators according to similar annotation behavior. Classifiers separately trained on data from the different groups and evaluated with the other groups’ test data exhibit noticeable differences in classification performance, which confirms our approach. The detailed results can be found in the following:

### Annotator Graph

We created one graph for each weight function. Table 1 provides the key metrics of the generated graphs. It is conspicuous that the graphs with Cohen’s kappa and Krippendorff’s alpha weight function have only 691,229 edges, while the other two have 1,560,078. This difference also causes the divergence of the average degree and density. The reason for the difference is that many relations between two annotators comprise only one comment. If both agree on an annotation, Cohen’s kappa and Krippendorff’s alpha are not defined; consequently, we do not have an edge. Therefore, graphs with these weight functions have fewer edges.

Weight function	Agreement Rate	Cohen's kappa	Krippendorff's alpha	Heuristic Function	Function
Number of identified groups	4	10	10	4	
Number of selected groups	3	4	4	3	
AVG(annotators/groups)	970.50	738.00	731.20	959.50	
SD(annotators/groups)	635.81	651.00	659.90	629.92	
Size of training set/test set	69,792 / 17,448	19,736 / 4,934	17,941 / 4,485	68,696 / 17,174	
Distribution of group sizes (selected groups)					

Table 2: Results of community detection

### Community Detection

Table 2 shows the results of community detection. While the Louvain algorithm split the graphs with the Agreement Rate and Heuristic Function as weight functions in 4 groups, 10 groups in the graph with Cohen's kappa and with Krippendorff's alpha were detected. An explanation for the divergence is the difference between the number of edges of the graphs. Since the groups have various numbers of members, we select only these with at least 250 annotators due to two reasons. By doing so, we ensure that we have enough annotated comments to train the classifiers. It may be noted at this juncture that only comments were selected for the training/test set if they were annotated by the group. Therefore, groups with a small number of annotators would have reduced the size of the training/test set. The distribution of the size of the training/test set is similar to the one of the numbers of identified groups. For Agreement Rate we have 69,792 annotated comments for the training set and 17,448 for the test set, for the Heuristic Function 69,696 and 17,174, for the Cohen's kappa 19,736 and 4,934, and for Krippendorff's alpha 17,941 and 4,485. The smaller data sizes for the last two are related to the smaller average size of groups.

To compare the different groups, we computed the inter-rater agreement for each group and between the groups by using Krippendorff's alpha. To calculate the rate between the groups, we compute Krippendorff's alpha using the union of all annotations from both groups. The inter-rater agreement scores (in percent, 100% means perfect agreement) for all four weight functions are depicted in Figure 1. The first column of each subfigure shows

the inter-rater agreement within each group. The 4/5 columns right to the line provide the inter-rater agreement between the groups, and the last column shows the average inter-rater agreement between the groups. Please note that the inter-rater agreement scores are hard to compare between the different weight functions/subfigures because the groups, the comments, and the annotations are different. To a certain degree, the results of the Agreement Rate and the Heuristic Function are comparable and the one of Cohen's kappa and Krippendorff's alpha because these pairs have the same number of groups and a similar number of comments. If we look at the inter-rater agreement within the groups (first column of each subfigure), we see that the groups exhibit varying scores and that the deviations to the baseline (group 0, data set average) also differ. If the score is higher than the baseline, the group is more coherent in regards to the annotations. If it is lower, the group is less coherent. Furthermore, the more scores are higher than the baseline, the better because it means that the algorithm is able to create more coherent groups.

Considering these aspects, we can say that Krippendorff's alpha and Heuristic Function produce better results than the other two. In the case of the Heuristic Function, the distance between the lowest and highest inter-rater reliability score (49.2% vs. 39.8%) is larger than the one of the Agreement Rate. In the case of Krippendorff's alpha, the distance between the lowest and highest score is the same as for Cohen's kappa. However, groups 3 and 4 of Krippendorff's alpha (49.8% and 50.0%) have higher scores than the two groups of Cohen's kappa with the highest inter-rater reliability (49.5% and



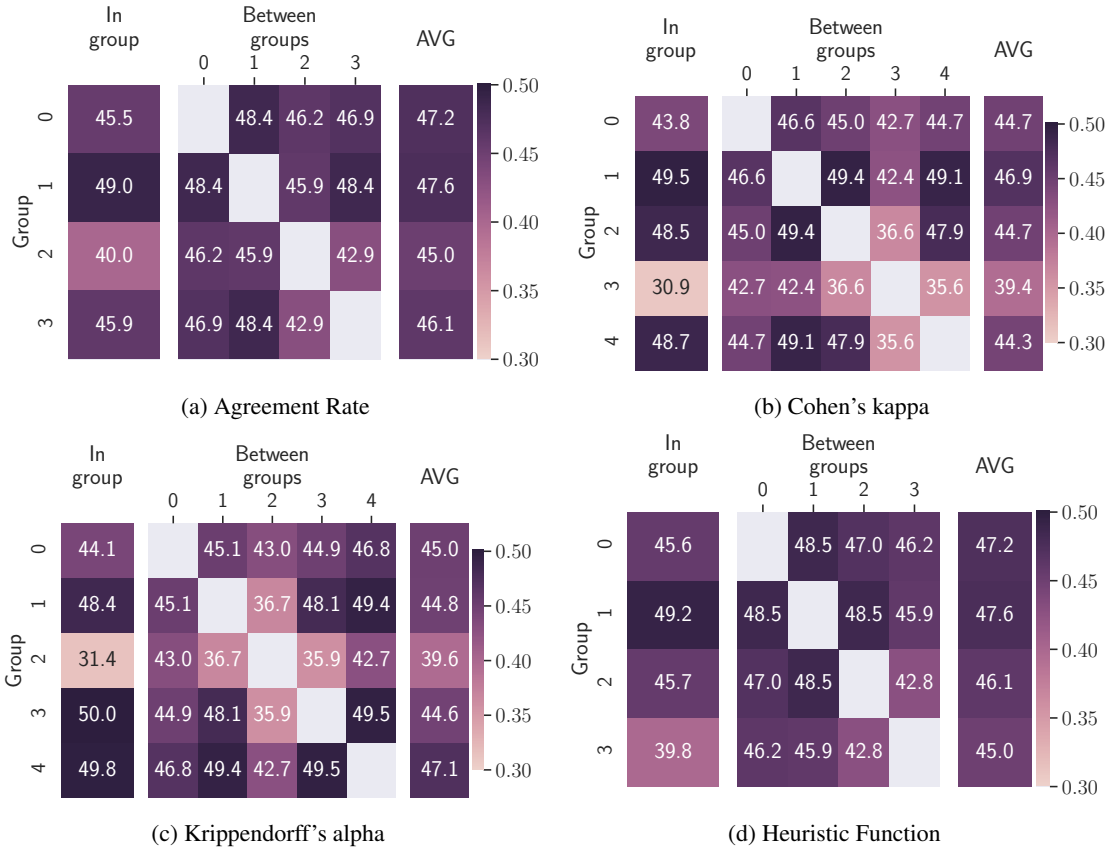


Figure 1: Inter-rater agreement within and between groups for different weight functions

48.7%). In both cases, the distance function is able to split the annotators into more coherent groups and a remainder than the other distance function. Since both distance functions and their results are hard to compare due to a different number of comments and groups, we choose both (Krippendorff's alpha and Heuristic Function) for the last part of the experiment.

### Classification Models and Their Performances

Instead of reporting the macro-F1 scores for the classifiers trained on the different group-specific training sets and tested on the all group-specific test sets, we report them relatively to the baseline (trained on group 0 and tested on group 0) for easier comparison, as depicted on Figure 2. The baseline for Krippendorff's alpha has a macro-F1 score of 85.27%, the one of Heuristic Function 88.57%. In addition to the relative scores, the figures contain an extra column and row with average values for better comparability.

It is conspicuous that the deviations reported in the first column of each matrix are lower than the rest. The reason is the following: These columns report the performances of the classifiers for the

different groups on the baseline test set. Since the baseline test set has the largest number of annotations, the labels are more coherent. Consequently, classifiers perform better on the baseline test set than on their own, less coherent test sets.

Figure 2a shows the results for Krippendorff's alpha distance function. The first observation is that the classifiers of groups 1 (+0.54) and 4 (+0.32) perform better on the baseline test set (group 0) than the baseline classifier. We can ascribe this to the fact that group 1 (48.4%) and group 4 (49.8%) have higher inter-rater reliability scores than the baseline (44.1%), meaning the annotations of groups 1 and 4 are more coherent. However, group 3 shows that higher inter-rater reliability does not directly imply a better performance on the baseline. It has a score of 50.0%, but it performs worse on the baseline test set (-0.23) and all classifiers perform poorly on the test set of group 3. A possible explanation can be that the annotations within the group are coherent but less coherent with respect to all other annotations. Group 2 exhibits the lowest performance on the baseline test set (-0.89) and all classifiers perform poorly on its test set. The reason is the noticeably low inter-rater reliability of 31.4% — the

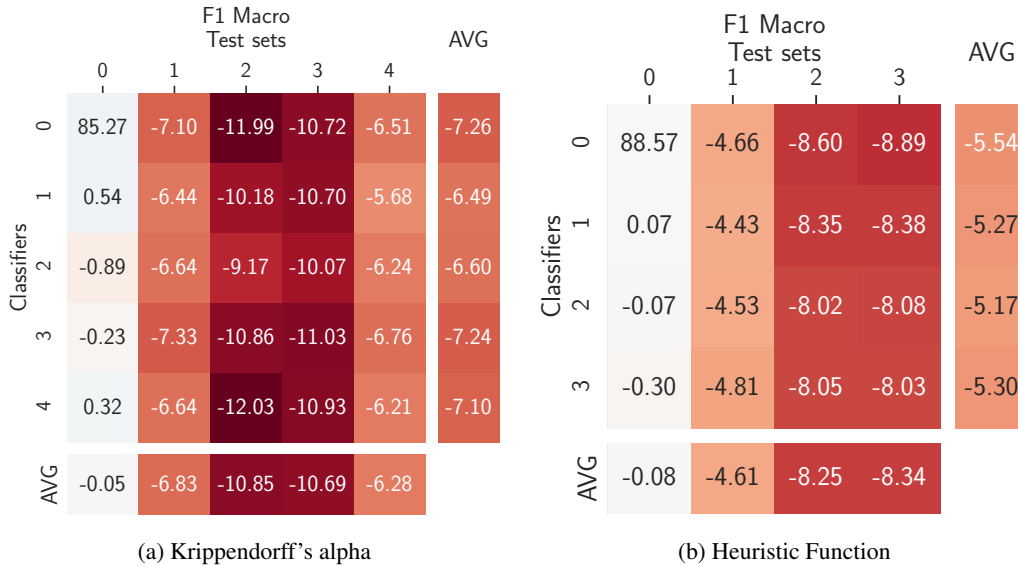


Figure 2: Macro F1 scores relative to the baseline (0,0)

lowest of all Krippendorff's alpha's groups. The low score indicates that the community detection algorithm grouped the annotators together whose annotation behavior is less compatible with the other's one.

In the case of the Heuristic Function (cf. Figure 2b), we can also find a group that performs better on the baseline test set than the baseline classifier and that has a high inter-rater reliability score (49.2%) — group 1 (+0.07). The explanation is the same as the one for groups 1 and 4 of Krippendorff's alpha. The classifier with the largest discrepancy is the one of group 3 (-0.30). This should not be surprising because group 3 has the lowest inter-rater reliability within the group (39.8%) and between the group and the baseline (46.2%). That is also the reason why all groups perform poorly on the test set of group 3. The group is comparable to group 2 of Krippendorff's alpha. Annotators that have an annotation behavior different from the rest are grouped together.

## 5 Discussion

The results show that the proposed method is suitable for identifying annotator groups purely based on annotation behavior. The deviations in inter-rater agreement rates of the groups and in the classifiers' performances prove this.

In regards to the weight functions, we found that both Krippendorff's alpha and Heuristic Function are more suitable than the other functions. Both are able to separate the annotators into different

groups based on their annotation behavior. However, it is difficult to choose a winner between both because of missing comparability. An advantage of our Heuristic Function in regards to Krippendorff's alpha as weight functions is that it does not have the undefined issue if two annotators assign only one type of label to the comments to be labeled. A potential improvement could be to combine Krippendorff's alpha weight function with the Heuristic Function.

The results of our method can be linked to annotator bias in the following manner: An identified annotator group that has a high inter-rater agreement within the group, but poor classification performance on the other test sets indicates that it has a certain degree of bias as the group's annotation behavior differs from the rest. For such insights, we see currently two possible use cases:

- The insights can be used to mitigate annotator bias. The annotations of these groups can either be weighted differently or deleted to avoid transferring the bias to the classification model.
- The insights can be used to build classification models that model the annotator bias. This can be helpful for tasks that do not have one truth but rather multiple perspectives. In the case of online abuse, it is possible that one group is more tolerant towards abusive language and another one less tolerant.

The novelty of our approach is that it is unsu-

pervised and does not require any stipulation of bias that you want to detect in advance. Existing approaches, such as [Binns et al. \(2017\)](#), who investigated gender bias, or [Sap et al. \(2019\)](#) and [Davidson et al. \(2019\)](#), who examined racial bias, defined in their hypothesis which kind of bias they want to uncover. Our method, however, does not require any pre-defined categories to detect bias.

## 6 Conclusion

In this paper, we proposed a novel graph-based method for identifying annotator bias through grouping similar annotation behavior. It differs from existing approaches by its unsupervised nature. But the method requires further research and refinement. To address our limitations, we propose the following future work:

Firstly, we used only one data set for our study. The approach, however, should be also tested and refined with other data sets. The Wikipedia Detox project, for example, provides two more data sets with the same structure, but with different tasks (toxicity and aggression). In general, data availability is a challenge of this kind of research because hate speech data sets mostly contain aggregated annotations. Therefore, we urge researchers releasing data sets to provide the unaggregated annotations as well.

Secondly, other approaches for grouping the annotators should be investigated. We used only one community detection method, the Louvain algorithm. But there are many more methods, such as the Girvan-Newman algorithm ([Girvan and Newman, 2002](#)) and the Clauset-Newman-Moore algorithm ([Clauset et al., 2004](#)).

Thirdly, our methods should be extended so that it can handle smaller groups. Our current approach requires at least 250 annotators in a group to ensure that we have enough training data. But it would be interesting to investigate smaller groups in the hope that these groups are more coherent in regards to their annotation behavior.

## Acknowledgments

This research has been partially funded by a scholarship from the Hanns Seidel Foundation financed by the German Federal Ministry of Education and Research.

## References

- Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators’ demographic characteristics. In *Proc. 4th Workshop on Online Abuse and Harms*.
- Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *International conference on social informatics*, pages 405–415. Springer.
- Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.
- Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Proc. 28th WWW Conf.*, pages 491–500.
- Aaron Clauset, Mark EJ Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Physical review E*, 70(6):066111.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.
- Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.
- Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).
- Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *2019 Conference on Empirical Methods in Natural Language Processing*, pages 1161–1166.
- Michelle Girvan and Mark EJ Newman. 2002. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826.
- K. Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Content Analysis: An Introduction to Its Methodology. Sage.
- Klaus Krippendorff. 2011. Computing krippendorff’s alpha-reliability.

- Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wotzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proc. 57th ACL Conf.*, pages 1668–1678.
- Anna Schmidt and Michael Wiegand. 2017. [A survey on hate speech detection using natural language processing](#). In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.
- Julius Sim and Chris C Wright. 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268.
- Leslie N Smith. 2018. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*.
- Bertie Vidgen, Austin Botelho, David Broniatowski, Ella Guest, Matthew Hall, Helen Margetts, Rebekah Tromble, Zeerak Waseem, and Scott Hale. 2020. [Detecting east asian prejudice on social media](#).
- Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *arXiv preprint arXiv:2004.01670*.
- Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. [Challenges and frontiers in abusive content detection](#). In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.
- Zeerak Waseem. 2016. [Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter](#). In *Proc. First Workshop on NLP and Computational Social Science*, pages 138–142.
- Maximilian Wich, Jan Bauer, and Georg Groh. 2020. Impact of politically biased data on hate speech classification. In *Proc. 4th Workshop on Online Abuse and Harms*.
- Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. [Detection of Abusive Language: the Problem of Biased Datasets](#). *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608.
- Wikimedia. n.d. Research:detox/data release. [https://meta.wikimedia.org/wiki/Research:Detox/Data\\_Release](https://meta.wikimedia.org/wiki/Research:Detox/Data_Release).
- Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2020. Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, 60(1):93–117.
- Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399.

## B.4 STUDY VII

©2021/2022 Springer

Reprinted with permission from:

Laurence Lerch, Maximilian Wich, Tobias Eder, and Georg Groh (2022). "Mediale Hasssprache und technologische Entscheidbarkeit: Zur ethischen Bedeutung subjektiv-perzeptiver Datenannotationen in der Hate Speech Detection." In: *Medien – Demokratie – Bildung: Normative Vermittlungsprozesse und Diversität in mediatisierten Gesellschaften*. Ed. by Gudrun Marci-Boehncke, Matthias Rath, Malte Delere, and Hanna Höfer. Wiesbaden: Springer Fachmedien Wiesbaden, pp. 295–310. ISBN: 978-3-658-36446-5. DOI: [10.1007/978-3-658-36446-5\\_17](https://doi.org/10.1007/978-3-658-36446-5_17)

This thesis includes the accepted version of publication and not the final published version.

### *Publication Summary*

Social media offers an easy way to communicate with thousands or millions of people. But this ease of use and reach also facilitated the spread of hate speech. Therefore, researchers started to develop algorithms to detect hate speech and other forms of abusive language automatically. However, these studies mostly focused on the technological aspect and neglected the ethical component. In this study, we addressed this gap and connected the technological methodology with the ethical conception. The study is structured as follows: firstly, the technological aspects of hate speech detection were outlined. Secondly, the concept of hate speech was examined from an ethical perspective based on the theories from John Stuart Mill, John Rawls, and Jeremy Waldron. Thirdly, we combined both perspectives and derived from them the implications for implementing the hate speech detection model. Based on these implications, we proposed in the fourth part a solution to how an annotation process should be designed to meet the ethical requirements.

### *Author Contributions*

Laurence Lerch headed the research project that is based on his bachelor's thesis supervised by Maximilian Wich. He substantially contributed to the conception of the study and wrote the largest part of the paper. Laurence Lerch and Maximilian Wich co-created the idea and concept of the study. Maximilian Wich was responsible for the technological aspects. Furthermore, he wrote a part of the paper and provided feedback. Tobias Eder contributed to the conception of the paper, wrote parts of the paper, and provided feedback. Georg Groh regularly discussed the ideas and concepts with the team and provided feedback on the manuscript.

# Mediale Hasssprache und technologische Entscheidbarkeit

## Zur ethischen Bedeutung subjektiv-perzeptiver Datenannotationen in der Hate Speech Detection

*Laurence Lerch, Maximilian Wich, Tobias Eder, Georg Groh*

**Zusammenfassung:** Mit den kommunikativen Möglichkeiten in den Sozialen Medien verbreitet sich auch *Hate Speech*. Die Technikwissenschaften haben es sich daher zum Ziel gesetzt, solche Inhalte algorithmisch detektierbar zu machen. In dieser Untersuchung wird gezeigt, dass jedoch gerade die Verbindung der ethischen Konzeptionen zu Hate Speech und der technologischen Methodik zur Detektion zu vertiefen ist. Erstens ist das Spezifikum von Hate Speech herauszuarbeiten, wobei insbesondere eine Abgrenzung zu *offense* und *hateful* erfolgen muss. Zweitens ist aufbauend auf ethischen Theorien die Methodik des maschinellen Lernens kritisch zu begutachten, da sich Limitationen aus deontologischer wie auch konsequentialistischer Sicht zeigen lassen und die Forderung nach Transparenz deutlich werden muss. Drittens sind Kompromisslösungen nur dann sinnvoll, wenn sie die Problematik subjektiver Annotationen in den Blick nehmen und sich der Kritik definitorischer, methodischer und ethischer Perspektiven bewusst sind.

**Keywords:** *Hate Speech, Hate Speech Detection, Machine Learning, Technikethik, Äußerungsfreiheit, Social Media, Hasskommentare, Transparenz, Annotationen, Künstliche Intelligenz*

### 1. Einleitung

Mediale Kommunikation in der Gesellschaft ermöglicht nicht nur eine Vernetzung von Menschen unterschiedlichster Nationen, Sprachen und Wertesystemen. Sie schafft gleichermaßen Möglichkeiten, Worte und Aussagen von Hate Speech zu empfangen, zu rezipieren und zu verbreiten. Unternehmen und Soziale Netzwerke wie Twitter, Facebook, Instagram oder TikTok müssen sich damit vor der Herausforderung gestellt sehen, jene Beiträge und Kommentare zu moderieren und gegebenenfalls zu regulieren. Aus diesem Grund hat sich in den letzten Jahren ein Forschungsfeld im Bereich der Informatik und Technologieentwicklung etabliert, das sich zum Ziel gesetzt hat, quantitative Methoden des maschinellen Lernens in der automatisierten Hate Speech Detection nutzbar zu machen. Dieser Beitrag versteht sich jedoch nicht aus einer ausschließlich technologischen Perspektive. Vielmehr zeigt sich vor dem Hintergrund aktueller Forschung, dass für die Hate Speech Detection fundamentale Fragen zur Definition von Hate Speech, zur Methodik selbst und zu subjektiven Annotationsprozessen bisher nicht oder nur unzureichend gestellt wurden. Gleichzeitig wird für eine ethische Perspektive deutlich, dass auch sie den Untersuchungsgegenstand der technologischen Hate Speech Detection nur unzureichend zu behandeln wusste.

In einer transdisziplinären Diskussion von Ethik und Informatik nähert sich der folgende Beitrag in Abschnitt 2 zunächst an eine Definition an, die für eine ethisch gestaltete Hate Speech Detection notwendig sein muss. Anschließend soll in Abschnitt 3 eine spezifisch technologische Perspektive zum Stand aktueller Forschung zu Wort kommen. In Abschnitt 4 folgt eine zunächst isoliert ethische Untersuchung zu möglichen Theorien der Äußerungsfreiheit und Hate Speech, wobei diese in Abschnitt 5 mit einer Kritik an der allgemeinen Methodik der Hate Speech Detection zu verbinden ist. In Abschnitt 6 werden zuletzt Kompromisslösungen zu Annotationen und Annotationsprozessen vorgeschlagen, die sich eingebettet in die Limitationen technologischer Möglichkeiten zu sehen haben.

## 2. Annäherungsversuche an eine Definition

*Hate Speech* als Begriff wird in der Forschung keineswegs eindeutig verwendet und zugleich nicht selten mit konzeptuell verschiedenen Terminologien verbunden. Für Schmidt und Wiegand (2017, S. 1) ist Hate Speech daher "a broad umbrella term for numerous kinds of insulting user-created content".

Angrenzende Terme an den Begriff Hate Speech waren in den letzten Jahren zum Beispiel *offensive language* (Razavi et al. 2010) oder auch *othering language* (Burnap und Williams 2016; Alorainy et al. 2019). Auf jene Bezeichnungen wird in Abschnitt 5 näher eingegangen werden.

Über vage Definitionen hinausgehend, versteht Liriam Sponholz (2018, S. 48) Hate Speech als „öffentliche Kommunikation bewusster und/oder intentionaler Botschaften mit diskriminierenden Inhalten“ und verdeutlicht im gleichen Zuge einen interessanten Aspekt: Hate Speech sei weder eine Frage von Hass noch von Sprache. Mit Blick auf dessen Konsequenzen sollte Hate Speech jedoch etwas mit Hass zu tun haben können, nachdem Hass bei Rezipienten erzeugt oder gefördert werden kann. Da im Folgenden außerdem Hasskommentare als spezifische Ausprägung von Hate Speech untersucht werden, wie sie auf Twitter oder Facebook vorkommen könnten, ist für die vorliegende Untersuchung die deutsche Übersetzung *Hasssprache* nicht notwendigerweise zu verwerfen.

Weiterhin unterscheidet Sponholz (2018, S. 62ff.) in Anlehnung an John Austin und John Searle bei Hate Speech drei Elemente der Sprechakttheorie, die auch für eine subjektiv aufgeladene Detektion wichtig werden:

- Im *Lokutionären Modell* kann Hate Speech anhand dessen identifiziert werden, was sprachlich tatsächlich dargestellt ist – bei einem Kommentar also unter anderem durch den verwendeten Wortschatz. Die nachfolgend beschriebenen, bisherigen Methoden der Hate Speech Detection bauen zu großen Teilen auf einem solchen Modell auf.
- Im *Illokutionären Modell* zeigt sich eine Differenz von Hate Speech und anderen Formen symbolischer bzw. nicht-intentionaler Diskriminierung, da es Ziel von Hate Speech ist, "Handlungen auf Basis der kommunikativen Herstellung von Ungleichheit zwischen Menschen zu vollziehen, sei es behaupten, mitteilen, zu versprechen, zu drohen, zu klagen oder zu entlassen" (Sponholz 2018, S. 71).
- Im *Perlokutionären Modell* wird Hate Speech an seinen Folgen gemessen, insbesondere mit offener Gewalt. Damit zeigt sich also eine stärker konsequenzorientierte Definition, die unter den Begriff *dangerous speech* fallen kann.

Eine solche Einteilung ist vor dem Hintergrund unterschiedlicher Methodiken zur Detektion geradezu notwendig, um das Spezifikum von Hate Speech herausarbeiten zu können. Als Basis hierfür lohnt sich ein Blick in die aktuelle Forschung technologischer Hate Speech Detection.

## 3. Technologische Methodik zur Detektion von Hate Speech

Mit dem Ziel einer *algorithmic content moderation* (Gorwa et al. 2020) hat sich in den letzten Jahren ein Forschungsfeld im Bereich des Natural Language Processing (NLP) gebildet, das sich vor allem den Methoden des maschinellen Lernens und des Supervised Learning bedient. Vereinfacht wird hierzu ein Datensatz, bestehend aus einer großen Anzahl von Einträgen respektive Kommentaren, annotiert. Das bedeutet, ausgewählte Personengruppen klassifizieren, ob ein Eintrag als Hate Speech zu verstehen sei oder nicht. Anschließend wird jener annotierte Datensatz genutzt, um ein Modell zu trainieren, woraufhin ein solches Modell die aus den Trainingsdaten erlernten Schemata zur Klassifizierung unbekannter Daten nutzen kann.

Im konkreten Fall der Hate Speech Detection identifizieren Schmidt und Wiegand (2017) verschiedene Ansätze zur Umsetzung, zum Beispiel *Wortgeneralisierungen* und *Lexikale Ressourcen*. Diese sind maßgeblich in einem lokutionären Modell zu verorten, sehen sich aber vor der Limitation kontext- und



diskursgebundener Aussagen wie Ironie oder Sarkasmus. Ein besonders interessanter Ansatz sind miteinbezogene *Meta-* beziehungsweise *Hintergrundinformationen*. Damit können zum Beispiel Kommentare, die kontextgebunden als Ironie zu verstehen sind, durch eine Analyse des bisherigen Tweet-Verlaufs des Users oder durch Assoziation mit bestimmten sozialen Gruppen tatsächlich als solches identifiziert werden.

Kritisierend ist in den letzten Jahren häufig auf die Probleme eines *Data Bias* hingewiesen worden, sodass diese Thematik inzwischen gar Einzug in die Populärwissenschaft, Gesellschaft und Literatur finden konnte, wie Werke von O’Neil (2017) oder Criado-Perez (2020) zeigen. Gerade für den vorliegenden Untersuchungsgegenstand subjektiv beeinflusster Hate Speech Detection sind zwei Arten von Bias im Supervised Learning zu unterscheiden: Ein *Dataset Bias* und ein *Annotator Bias*. Beide Arten können zu Diskriminierung führen – gerade das, was mit der Hate Speech Detection ursprünglich verhindert werden soll.

Im *Dataset Bias* ist die Auswahl der zu annotierenden Daten in solch einer Weise verzerrt, dass die Generalisierbarkeit der Modelle signifikant beeinträchtigt werden kann. Wiegand et al. (2019) zeigen in ihrer Untersuchung, dass ein in der Forschung viel genutzter Datensatz einen thematischen Bias bezogen auf Sportnachrichten aufzeigt. Dadurch lernt ein Modell möglicherweise, anhand von Begriffen aus einem Sportkontext Hasskommentare von normalen Kommentaren zu unterscheiden. Und nicht etwa anhand von Hate Speech selbst. Andererseits kann ein solcher Bias durch eine mangelnde Diversität in der Auswahl der User entstehen. Ein Beispiel hierfür ist der weit verbreitete Datensatz von Waseem und Hovy (2016), der zwar tausende, bereits annotierte Tweets von Twitter enthält, die meisten davon jedoch lediglich von wenigen Usern stammen (Arango et al. 2019).

Ein *Annotator Bias* dagegen kann unter anderem dann entstehen, wenn subjektive Wahrnehmung von annotierenden Personengruppen in den Datensatz einfließen, insofern nach Waseem (2016) die Qualität der annotierten Datensätze maßgeblich davon abhängt. Dabei kommt er in seiner Untersuchung zu dem Schluss, dass Laien zwar geneigter sind, Kommentare als Hate Speech zu klassifizieren. In der Qualität jedoch übertreffen expertenbasiert annotierte Datensätze die von Laien. Grundsätzlich ist dabei auch die Frage nach einer genauen und eindeutigen Definition von Hate Speech zu verorten, nachdem “results imply that hate speech is a vague concept that requires significantly better definitions and guidelines in order to be annotated reliably” (Ross et al. 2016, S. 9). Eine solche Sicht ist kohärent mit Waseem (2016, S. 141), nachdem “hate speech is hard to annotate without intimate knowledge of hate speech” – und *knowledge* eine substantielle und objektive Definition benötigt.

Solche Erkenntnisse stehen jedoch im häufigen Gegensatz zur aktuellen Praxis: Wie Vidgen und Derczynski (2020, S. 15) in ihrer umfangreichen Untersuchung feststellen, ist Crowdsourcing “widespread in NLP research because it is relatively cheap and easy to implement”. Ein weiteres Problem, das mit der Praxis des Crowdsourcings und der Beauftragung unabhängiger und gegebenenfalls unbekannter Personen und Personengruppen geschaffen wird, ist eine Reduktion der Transparenz im gesamten Annotationsprozess, welche eine Überprüfbarkeit vor dem breiten Einsatz in gesellschaftlicher Praxis erschwert. Wie Al Kuwatly et al. (2020) beschreiben, können durch demographische Faktoren der annotierenden Personen (zum Beispiel Geschlecht oder Alter) Unterschiede in deren Annotationsverhalten gemessen werden. Sap et al. (2019) zeigen in diesem Zusammenhang auf, wie Insensibilität von annotierenden Personen gegenüber Dialekten des African American English zu rassistischem Bias führen können. Aufgrund jener unbefriedigenden Situation wird im Folgenden eine zunächst isoliert ethische Perspektive zur Konzeption von Hate Speech besprochen, welche die Grundlage für ganzheitliche, praxisrelevante (Kompromiss-)Lösungen bilden soll.

#### 4. Ethische Grundfragen zum Konzept *Hate Speech*

Wie in Abschnitt 2 bereits angeklungen, ist mit *Hate Speech* ein Konzept verbunden, das zunächst diffus und uneindeutig verwendet und diskutiert wird. Um weitere Klärung auch aus ethischer Sicht zu ermöglichen, legt dieser Abschnitt drei verschiedene Zugänge dar: John Stuart Mill (2011) und sein Konsequentialismus; John Rawls (1999) mit Blick auf seine *Theory of Justice*; und Jeremy Waldron (2012), der es vermochte, mit seinem Werk *The Harm in Hate Speech* auf fundamentale Konzeptionen von *Hate Speech* aufmerksam zu machen. Zwar ist es nicht möglich, alle Theorien in ihrer Gänze auszulegen, doch werden pointierte Aspekte herausgegriffen, die eine Relevanz für den vorliegenden Untersuchungsgegenstand haben.

Nach John Stuart Mill (2011) ist mit seinem viel zitierten *Harm Principle* grundsätzlich all das geboten bzw. erlaubt, was keinen Schaden anrichtet. Eine weit verbreitete Kritik an einem solchen Ansatz kann jedoch sein, dass *harm* nicht immer eindeutig sein muss (Schefczyk und Schramme 2015). Mill zieht in Bezug zur Äußerungsfreiheit interessanterweise eine Differenzierung von *mere offense* und *genuin harm*, wobei das Schadensprinzip nur für die letztere Einstufung greife, während bei ersterer davon auszugehen ist, dass eine Regulierung nicht stattfinden sollte (Brink 2009, S. 55). Auch für *Hate Speech* selbst ist dies essentiell, als in einem solchen Modell von dem verkürzten Begriff *offense* und folglich *offensive* Abstand genommen werden muss. Die Schwelle, unter das *Harm Principle* zu fallen, ist damit qualitativ erhöht. Gleichzeitig wird durch eine solche strenge Konsequenzorientierung eine Nähe zum angesprochenen, perlokutionären Modell deutlich. Die grundsätzliche Möglichkeit, *offense* im Rahmen eines über Mill hinausgehenden *offense principle* (Feinberg 1985) zu regulieren und in den Sozialen Netzwerken zu moderieren, bleibt jedoch weiterer Diskussion würdig und kann gegebenenfalls ebenso zum Gegenstand maschineller Detektion werden.

Ein fundamental andere Annäherung an die Thematik ist mit John Rawls (1999) und seiner *Theory of Justice* möglich. Er entwickelte mit der *Original Position* und dem *Schleier des Nichtwissens* ein populäres Gedankenspiel, welches in abstrahierter Weise interessant für die *Hate Speech* Detection ist, obgleich er sich selbst nie speziell mit *Hate Speech* auseinandergesetzt hat (Waldron 2012, S. 69f.). Es stellt sich die Frage, ob nicht jenes Gedankenspiel auch für Datenannotationen eine – wenn auch idealisierte – Idee ist, frei von subjektiven Empfindungen *unbiased* Entscheidungen bei Hasskommentaren zu treffen. Dies würde gleichzeitig aber auch bedeuten, dass Annotationen durch intransparente Crowdsourcing Maßnahmen nicht *direkt* als deskriptive Untersuchung der Wahrnehmung von *Hate Speech* einen normativen Anspruch haben können, insofern sie subjektiver Entscheidbarkeit und Wahrnehmung ausgeliefert wären. Erst im Nachhinein müsste überprüft werden, ob die Annotationen und die Entscheidungen dieser deontologischen Kritik standhalten können. Ein solches Begründungsmodell, wie es John Rawls versteht, ist damit aus methodischer Perspektive höchst interessant und schafft einen Gegenpol zu einer rein konsequenzorientierten Argumentation.

Ein dritter, umfassender Ansatz ist mit Jeremy Waldron (2012) möglich, insofern er sich direkt mit *Hate Speech* auseinandergesetzt hat. Für seine Argumentation der Regulierung stellt er nämlich eine interessante, gesellschaftssoziale und demokratiebasierende Frage, die sich implizit in einer Definition von *Hate Speech* widerspiegelt: Wie sieht eine *well-ordered society* aus, die mit *Hate Speech* umgeht? Zwar sieht er sich selbst nicht in der Rawlsschen Tradition, greift jedoch diesen Aspekt der *well-ordered society* aus der *theory of justice* auf – nämlich, dass jede Person die Prinzipien akzeptiert und weiß, dass dies auch die anderen tun. Waldron kommt dabei zu zwei Konzeptionen, die sich gegenseitig bedingen und im *public good of inclusiveness* resümieren: *dignity* und *assurance*. Also die Vergewisserung des Schutzes der Würde eines jeden Teils der Gemeinschaft. Mit jener *dignity* ist jedoch „not just some Kantian aura“ (Waldron 2012, S. 5) gemeint, sondern „a person’s basic entitlement to be regarded as a member of society in good standing, as someone whose membership of a minority group does not disqualify him or her from ordinary social interaction“ (Waldron 2012, S. 105). Als Beispiele für eine darauf aufbauende *Hate Speech* Regulierung führt er Länder an wie Dänemark, Kanada, Neuseeland,

das Vereinigte Königreich und auch Deutschland. Im Falle Deutschlands bezieht er sich auf §130 StGB Abs. 2 und dabei explizit den Angriff auf die Würde des Menschen in Art. 1 GG.

In Bezug auf das bereits bei Mill angedeutete Verhältnis von *offense* und *hate speech* erkennt auch Waldron, dass *offense* zwar persönlich verletzend ist. Dieses Verletzt-*fühlen* darf jedoch im Vergleich zur definierten *dignity* und *assurance* kein Maßstab zur Regulierung von Hate Speech sein. Auch mit Waldron ist es somit für die Detektion von Hate Speech nur richtig, einen qualitativen Unterschied zu *offense* und *offensive* anzunehmen. Natürlich kann Hate Speech beleidigend, verletzend oder angreifend sein, aber so ist doch nicht jede Beleidigung oder Verletzung zwangsläufig Hate Speech. Exemplarisch wird seine Theorie deutlich an zwei Aussagen, die Hate Speech mit sich und in sich tragen und die hierbei die vorherig angedeutete Definition verdeutlichen können. Hate Speech vermittelt zunächst den Zielpersonen:

“Don’t be fooled into thinking you are welcome here. The society around you may seem hospitable and nondiscriminatory, but the truth is that you are not wanted, and you and your families will be shunned, excluded, beaten, and driven out, whenever we can get away with it. We may have to keep a low profile right now. But don’t get too comfortable. Remember what has happened to you and your kind in the past. Be afraid.” (Waldron 2012, S. 2)

Damit ist der diskriminierende Aspekt, auf den bereits in einer Begriffsbestimmung hingewiesen wurde, in literarisch fühlbarer Weise aufmerksam gemacht. Als nun Hate Speech deutlich über persönlichen Schaden hinausgeht, werden aber auch gesellschaftssoziale und demokratische Strukturen untergraben. Zentral ist vor dem Hintergrund seiner Argumentation, dass Waldron über diesen reinen Aspekt des Angriffs auf Personen als Teil von Personengruppen hinausgeht und explizit eine Wir-Die-Dichotomie beschreibt. Dies macht er in einer weitere Botschaft von Hate Speech deutlich, wobei insbesondere jener Aspekt in den folgenden Abschnitten als Kritik an der Hate Speech Detection verstanden werden kann:

“We know some of you agree that these people are not wanted here. We know that some of you feel that they are dirty (or dangerous or criminal or terrorist). Know now that you are not alone. Whatever the government says, there are enough of us around to make sure these people are not welcome. There are enough of us around to draw attention to what these people are really like. Talk to your neighbors, talk to your customers. And above all, don’t let any more of them in.” (Waldron 2012, S. 2f.)

## 5. Methodische Kritik und Lösungsansätze der technologischen Hate Speech Detection

Trotz des Versuchs, *harm* im Gegensatz zu *offense* objektiver auszurichten, bleibt die grundsätzliche Frage nach subjektivem und wahrgenommenen Empfinden von Leid oder Schaden bestehen. Diese Frage diskutiert zum einen die Möglichkeiten einer Methodik des maschinellen Lernens, zum anderen aber auch die darauf aufbauende Notwendigkeit der Annotation von Daten, wodurch negative Effekte resultieren können. Eine zunächst allgemeinere, jedoch erhellende Analyse für das Verhältnis von Ethik und Artificial Intelligence bezogen auf unterschiedliche, normative Theorien findet sich bei Virginia Dignum (2019). Dabei greift sie metaethischen Fragestellungen auf und weitet sie anschließend auf die konkrete Umsetzbarkeit von Artificial Intelligence aus.

Sie führt hierfür zwei Modelle an: *Top-Down* und *Bottom-Up*. Wohingegen ein Top-Down Ansatz, unter anderem auch bekannt als *Symbolische AI*, regelbasiert logische Entscheidungen treffen soll, wird mit einem Bottom-Up Ansatz versucht, kognitive Prozesse mit erlerntem Wissen respektive Erfahrung zu modellieren. Eine hier vorgestellte Theorie der Gerechtigkeit sieht sich in einer solchen Klassifikation sicherlich mit einem Top-Down Ansatz vertreten, da Regeln die Grundlage wären. Ein Konsequentialismus ist weniger eindeutig zu verorten. Eine Möglichkeit von Bottom-Up Ansätzen ist

jedoch nicht zwangsläufig abzulehnen, nachdem einzelne Erfahrungen von Konsequenzen eine Generalisierung auf neue Situationen induzieren könnten.

Mit Blick auf die Ansätze selbst verfolgt Bottom-Up jedoch eine höchst diskussionswürdige Gleichstellung von “social acceptability with ethical acceptance” (Dignum 2019, S. 76). Im konkreten Fall der Hate Speech Detection im Sinne maschinellen Lernens bedeutet dies, dass eine solche Methodik nur dann zu rechtfertigen sei, wenn mit einer deskriptiven Untersuchung ein Normativitätsanspruch verbunden wird. Sofern annotierende Personengruppen ein getreues und diverses Abbild einer Gesellschaft darstellen, könnte man verleitet sein, dass tatsächlich subjektive Wahrnehmung von Hate Speech eine Rolle spielen darf oder gar soll. Doch gibt es auch hier eine probabilistische Problematik: Da algorithmische Entscheidungen maßgeblich einer statistischer Basis und damit folglich den Wahrnehmungen der Mehrheit entsprechen, besteht eine immanente Gefahr, Minderheiten in der Entscheidung per Mehrheitsprinzip zu überstimmen. Selbst wenn Kommentare mehrfach annotiert werden, bleiben solche Sonderfälle der Uneinigkeit ohne Möglichkeit zum Diskurs unbeachtet.

Wie Dignum (2019) andeutet, sollen hybride Modelle eine Verbindung von Top-Down und Bottom-Up Ansätzen schaffen. Doch muss hierbei auf eine weitere Problemstellung hingewiesen werden, die für die Hate Speech Detection zentral ist: Die Nachteile der einzelnen Ansätze werden nicht *per se* in einer hybriden Anwendung umgangen. Mit dem Wissen um jene komplexe Konzeption von Hate Speech besteht die Gefahr, in einer solchen Verbindung die Nachteile zu verknüpfen und die Vorteile zu übergehen. In der Praxis mag es solche Ansätze der Verbindung probabilistischer und regelbasierter Methoden vereinzelt geben (Burnap und Williams 2015), jedoch ist ein explizite Diskussion auf ethischer sowie quantitativer Basis und dessen Notwendigkeit bisweilen nicht erkennbar.

Zudem müssen strikt regelbasierte Ansätze wie lexikalisch-lokutionäre Methoden sowohl aus Perspektive der Erfolgsaussichten als auch vor einem ethischen Hintergrund kritisch betrachtet werden. Diese bewegen sich entweder im Bereich von *offense* und nehmen das Spezifikum einer Wir-Die-Dichotomie respektive einer ethischen Konsequenzorientierung nicht auf. Oder aber sie erhöhen die Rate von falsch-positiven Klassifizierungen, da Kontext, Ironie, Sarkasmus und ähnliches weiterhin regelbasiert herausfordernd sind. Eine *direkte* Umsetzung von Regeln, wie sie in Social Media Richtlinien statuiert sind, ist damit aufgrund mangelnder Diskursfähigkeit und fehlenden Kontextwissens im Sinne eines deontologischen Ansatzes zum heutigen Zeitpunkt außer Reichweite. Davon unberührt bleibt die Möglichkeit, mit jenen Bottom-Up Ansätze regelbasierte Entscheidungen zu simulieren. Dazu muss versucht werden, die Annotation von Hate Speech Kommentaren frei von subjektiver Wahrnehmung zu ermöglichen, sodass letztlich jene Regeln bei wenigen Annotationen auf eine Gesamtheit von Kommentaren generalisiert werden. Deutlich zu machen ist: Dabei handelt es sich tatsächlich lediglich um eine Simulation, einen *indirekten* Versuch, regelbasierte Entscheidungen zu implementieren. Aus streng methodischer Sicht bleibt ein solcher Ansatz ein Kompromissversuch, der als Annäherung nur durch die erfasste Komplexität von Hate Speech zu rechtfertigen ist.

Zwar sind selbst direkte menschliche Urteile zur Moderation von Hate Speech stets Entscheidungen, die trotz der Vorgabe von Richtlinien mehr oder minder beeinflusst sein können von subjektiver Wahrnehmung. Der Schritt einer Generalisierung von einzelnen Entscheidungen auf unbekannte Kommentare bleibt aber ein kritischer Punkt, bei dem zahlreiche, auch technische Faktoren zu diskutieren sind. Der Hauptfokus der disziplin-internen Diskurse in der Informatik liegt dabei auf dem Faktor der Transparenz, der die Datengrundlage, Annotationsmechanismen, technische Details verwendeter Algorithmen und die Evaluierungsmethodik einschließt.

Ein inzwischen verbreiteter Lösungsansatz für Transparenz sind Ansätze zur Verbesserung der Dokumentation der konkreten Methodik, wie beispielsweise bei Gebru et al. (2018) oder Bender und Friedman (2018) ausgeführt wird. Da sich jene Ansätze nicht allein auf das Feld Hate Speech Detection beziehen, sind sie auf alle datenbasierenden Modelle anwendbar. So schlagen Bender und Friedman vor, im Fall von natürlichsprachlichen Daten auf die sozio-demografischen Kontexte der aufgenommenen

Personen, also die Urheber der Texte, sowie der annotierenden Personengruppen einzugehen. Darüber hinaus appellieren sie für die genaue Dokumentation der enthaltenen Varietäten einer Sprache, der Sprechsituation, sowie den detaillierten Gründen für die Aufnahme eben genau dieser verwendeten Daten in den Datensatz (Bender und Friedman 2018).

Die explizite Transparenz soll dabei in zweifacher Weise der Methodik dienlich sein: (1) Zum einen, um insbesondere bei Wiederverwendung einen Einblick in die genauen Hintergründe und Inhalte von Datensätzen zu ermöglichen. (2) Zum anderen, um einen Katalog relevanter Kriterien öffentlich zu machen, wodurch eine kritische Auseinandersetzung bereits während des Selektionsprozesses der Daten gefördert wird und eine Datenbasis sowie jene Prozesse der Annotation qualitativ verbessert werden können.

Auf einer technisch-pragmatischen Ebene zielen Faktoren der kritischen Auseinandersetzung insbesondere auf das bereits angesprochene Problem von Bias und Ungleichbehandlung. Dazu zählen Ansätze zur technischen Bereinigung von Bias in natürlichsprachlichen Modellen (Bolukbasi 2016) oder zur Herstellung von Fairness unter Berücksichtigung unterschiedlicher soziokultureller Gruppen (Johndrow und Lum 2019). Diesen ist gemein, dass sie sich mit existierenden Datensätzen und Modellen beschäftigen, die nachgewiesenermaßen Schwächen in den jeweiligen Feldern aufweisen. Nur eine nachträgliche Milderung der durch bestimmte Daten auftretenden Probleme kann somit das Ziel sein.

Im Allgemeinen beschäftigen sich die Ansätze aus dem Feld der Informatik jedoch primär mit konkreten Problemen während des Trainings und Einsatzes von Hate Speech Detection Modellen. Was größtenteils noch unbeachtet bleibt ist eine Einordnung der epistemischen Voraussetzungen solcher Systeme sowie einer Auseinandersetzung mit der Praxis derzeitiger Annotationsformen. Selbst in den Fällen, in denen die Annotationen als entscheidender Problemfaktor identifiziert werden, fehlt eine Bewertung, die über die Zusammensetzung der annotierenden Gruppen hinausgeht. Mit Blick auf Annotationsformen selbst sollen daher im letzten Abschnitt perspektivisch und exemplarisch konkrete Problematiken der Begriffsnutzung im Kontext von Hate Speech in der jüngeren Forschung aufgezeigt und untersucht werden.

## **6. Perspektiven zur begrifflichen Umsetzung von Annotationen in der Hate Speech Detection**

Ungeachtet methodischer Kritik und Kompromissnotwendigkeit bei der Hate Speech Detection, ist in der technologischen Forschungslandschaft selbst die Praxis deskriptiver Annotationen keineswegs in aller Breite und in ihrer Umsetzung diskutiert worden. Das Ziel der folgenden Ausführungen ist dazu nicht, allumfassende Lösungen bereitzustellen. Doch sollen diese Aspekte in ihrer beispielhaften Natur als Beitrag zum notwendigen Diskurs zu verstehen sein, wodurch Verbesserungsvorschläge zu Annotationen möglich sein werden.

Mit Blick auf die vorherigen Überlegungen ist zunächst festzuhalten, dass auf Richtlinien basierende Annotationen zumindest zum jetzigen Zeitpunkt als geeignetste Möglichkeit zu betrachten sind, einen solch deskriptiven Prozess annäherungsweise ethisch zu gestalten. Unbedingt müssen hierbei Guidelines unterstützen, wobei Vidgen und Derczynski (2020) in diesem Zusammenhang erkennen, dass Annotationsguidelines vieler Datensätze nicht veröffentlicht oder geteilt werden. Zudem ist es dafür unerlässlich, diverse Gruppen mit der Annotation zu beauftragen, diese im Rahmen der Definition zu schulen und gleichzeitig eine Transparenz zur iterativen Überprüfung und Weiterentwicklung zu ermöglichen.

Gerade der Aspekt einer abgrenzenden Definition, auf der Guidelines aufgebaut werden können, ist jedoch in der technologischen Forschung bisher wenig beachtet. So mischen sich allerlei Begrifflichkeiten, die von *hasserfüllt/hateful* über *offensive*, *abusive*, *racist* oder *sexist* reichen und synonym verwendet werden. Burnap und Williams (2015, S. 227) fragen beispielsweise für ihre

Annotationen, ob “this text offensive or antagonistic in terms of race ethnicity or religion” sei. Dabei konnte Abschnitt 3 dieser Untersuchung zeigen, dass gerade *offense* und *offensive* in seiner Begrifflichkeit nicht deckungsgleich mit Hate Speech sind. Davidson et al. (2017) ziehen diese Unterscheidung in ihrem annotierten Datensatz richtigerweise.

Obgleich in zahlreicher Forschung als solches bezeichnet, kann auch *hasserfüllt* beziehungsweise *hateful* nicht als Alternative verstanden werden. Was oder wer ist mit *hateful* konkret gemeint? Geht es um “hateful users” (Waseem und Hovy 2016, S. 92), dann ist ein solcher Begriff abzulehnen, da ein emotionales Erfüllt-sein von Hass nicht konstitutiv notwendig für Hate Speech ist. Aber auch eine Spezifizierung auf “hateful tweets” (Davidson et al. 2017), “hateful comments” (Schmidt und Wiegand 2017, S. 6), “hateful content” (Ross et al. 2016, S. 6) oder “hateful responses” (Burnap und Williams 2016, S. 3) ist ungeeignet: Ein Kommentar selbst muss weder von Hass bestimmt sein, noch muss dieser zwingend Hass ausdrücken. Sponholz (2018) verortet daher Hateful Speech im Lokutionären Modell. Vor dem Hintergrund jener Kritik sind, abhängig von der jeweils angenommenen ethischen Konzeption, Begriffe hilfreich wie *hassschürend* und/oder *othering*. *Hassschürend* oder auch *hetzerisch* greifen einen stärker konsequentialistischen Ansatz auf, wobei ein perlokutionärer Sprechakt mit einbezogen wird. Jedoch erfordert damit die Bewertung von Folgen durch annotierende Personen weitreichende und über die Lokution hinausgehende Überlegungen. Aus der Perspektive individueller Wahrnehmung scheint *hassschürend* vor dem Hintergrund differenter Erfahrung von Hate Speech subjektiver zu bewerten.

*Othering*, im Deutschen im Sinne einer *Ausgrenzung* übersetzt, ist mit Blick auf eine Konzeption von Hate Speech bei Waldron am stimmigsten. Eine solche “us-them-dichotomy” (Schmidt und Wiegand 2017, S. 2) findet sich in Ansätzen beispielsweise bei Alorainy et al. (2019). Zur Simulation der Detektion illokutionärer Sprechakte können Nennungen abstrahierter Botschaften hilfreich sein, wie sie Waldron beschreibt, insofern sie über den rein lokutionären Akt von Hateful Speech und über eine starke Konsequenzorientierung hinausgehen. In Annotationsprozessen und Guidelines können solche Begrifflichkeiten und Zitate Missverständnisse ausräumen und gleichzeitig das Spezifikum der Hate Speech Detection im Vergleich zu Beleidigungen oder rein persönlichen Angriffen in den Vordergrund rücken.

Trotzdem kann auch ein solcher Vorschlag nur als Kompromiss verstanden werden, da er keine allgemeine Antwort auf die im vorherigen Abschnitt angedeutete Allgemeinkritik an der Methodik zulässt und breit angelegte Diskurse im Prozess der Annotation nicht möglich sind. Eine Nutzung in der Gesellschaft kann daher zum jetzigen Zeitpunkt weder sinnvoll noch zielführend sein. Jedoch bietet jener Kompromiss von *othering* und *hassschürend* in Verbindung mit dem Wissen um die Limitationen der Hate Speech Erkennbarkeit eine Basis, auf der Annäherungen an ethisch gestaltete und transparent erarbeitete Lösungen auch in Zukunft versucht werden sollten.

## 7. Zusammenfassung

Bewusst hat diese Arbeit Diskussionen ausgeklammert, die in aller Breite ein algorithmisches Verstehen, ein Bewusstsein, ein Weltwissen oder gar menschliche Intentionalität umfassen. Fokus und Ziel dieser Untersuchung war es dagegen, technologische Hate Speech Detection und ethische Kritik praxisrelevant sowie transdisziplinär zu verbinden. Dabei konnte auf Basis terminologischer, ethischer und methodischer Analysen gezeigt werden:

- Definitiv beschreibt Hate Speech je nach Konzeption entweder objektiv negative Folgen und/oder eine Wir-Die-Dichotomie, wobei Gefühle eine untergeordnete Rolle spielen. Hate Speech Detection hat sich vor dem Hintergrund dieser Konzeption und mit Blick auf ethische Theorien von *offense* sowie *hateful* abzugrenzen und muss dies auch in der Praxis der Datenannotation beachten.

- Aus technologischer Perspektive können aktuelle Methodiken keine deontologischen Theorien als direkten Top-Down Ansatz umsetzen. Subjektive Erfahrung spielt aber im Bottom-Up Modell eine entscheidende Rolle. Auch mit Blick auf Sprechakte wird eine Limitation der Hate Speech Detection deutlich, da sich diese maßgeblich im Bereich der Lokution bewegt.
- Als möglicher Kompromiss können Umsetzungen ethischer Theorien nur simuliert werden, wobei Annotationen expertenbasiert und divers sein müssen. Transparente Diskurse, Prozesse und Guidelines haben die Grundlage zu bilden, nach der ein subjektiver *Annotator Bias* minimiert wird. Fragen zur Begrifflichkeit sollen diskutiert werden, weil ansonsten die Natur von Hate Speech nicht aufgegriffen wird. Dabei erweisen sich letztlich *hassschürend* und *othering* als vielversprechend.

Für die Informatik ist es folglich notwendig, über eine reine Methodik der Hate Speech Detection hinauszugehen und die Frage nach einer Metakritik zur grundsätzlichen Möglichkeit zu stellen. Auf der anderen Seite liegt es in der Pflicht der Geisteswissenschaften und der Ethik, neben einer isolierten Konzeption von Hate Speech auch allgemein-technologische Erkennbarkeit zu diskutieren. Ob ein komplexes Konstrukt wie Hate Speech nämlich jemals erfolgreich maschinell detektiert werden kann und zugleich einer ethischen Kritik standhalten wird, bleibt zum jetzigen Zeitpunkt nicht vorherzusagen. Eine stetige Annäherung an jenes Ideal sollte jedoch auch in Zukunft Auftrag und Ziel einer solchen transdisziplinären Forschungen sein.

## Literatur

Al Kuwatly, Hala, Wich, Maximilian, und Groh, Georg (2020). Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics. In: *Proceedings of the Fourth Workshop on Online Abuse and Harms* (S. 184-190). doi:10.18653/v1/2020.alw-1.21

Alorainy, Wafa, Burnap, Pete, Liu, Han, und Williams, Matthew L. (2019). "The Enemy Among Us": Detecting Cyber Hate Speech with Threats-based Othering Language Embeddings. *ACM Transactions on the Web* 13(3), Artikel 14, 26 Seiten. doi:10.1145/3324997

Arango, Aymé, Pérez, Jorge, und Poblete, Barbara (2019). Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (S. 45-54). doi:10.1145/3331184.3331262

Bender, Emily M., und Friedman, Batya (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6, 587-604. doi:10.1162/tacl\_a\_00041

Bolukbasi, Tolga, Chang, Kai-Wei, Zou, James, Saligrama, Venkatesh, und Kalai, Adam (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems* (S. 4356-4364).

Burnap, Pete, und Williams, Matthew 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. doi:10.1002/poi3.85

Burnap, Pete, und Williams, Matthew L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science* 5, Artikel 11. doi:10.1140/epjds/s13688-016-0072-6

Brink, David O. (2009). Mill's liberal principles and freedom of expression. In: C. L. Ten (Hrsg.), *Mill's On Liberty: A Critical Guide* (S. 40-61). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511575181.003

Criado-Perez, Caroline (2020). *Unsichtbare Frauen: Wie eine von Daten beherrschte Welt die Hälfte der Bevölkerung ignoriert*. München: btb Verlag.

Davidson, Thomas, Warmley, Dana, Macy, Michael, und Weber, Ingmar (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In: *Proceedings of the 11th International AAAI Conference on Web and Social Media* (S. 512-515).

Dignum, Virginia (2019). *Responsible Artificial Intelligence. How to Develop and Use Artificial Intelligence in a Responsible Way*. Cham: Springer. doi:10.1007/978-3-030-30371-6

Feinberg, Joel (1985). *The Moral Limits of the Criminal Law: Offense to Others. Volume 2*. Oxford: Oxford University Press.

Gebru, Timnit, Morgenstern, Jamie, Vecchione, Briana, Vaughan, Jennifer Wortman, Wallach, Hanna, Daumé III, Hal, und Crawford, Kate (2018). Datasheets for Datasets. *arXiv:1803.09010*

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

Johndrow, James E., und Lum, Kristian (2019). An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *The Annals of Applied Statistics* 13(1), 189-220.

Mill, John S. (2011). *On Liberty*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139149785

O'Neil, Cathy (2017). *Angriff der Algorithmen: Wie sie Wahlen manipulieren, Berufschancen zerstören und unsere Gesundheit gefährden*. München: Carl Hanser.

Rawls, John (1999). *A Theory of Justice*. Cambridge, MA: Harvard University Press. doi:10.2307/j.ctvkjb25m

Razavi, Amir H., Inkpen, Diana, Uritsky, Sasha, und Matwin, Stan (2010). Offensive Language Detection Using Multi-level Classification. In: A. Farzindar, V. Kešelj (Hrsg.), *Advances in Artificial Intelligence. Canadian AI 2010. Lecture Notes in Computer Science*, Volume 6085 (S. 16-27). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-13059-5\_5

Ross, Björn, Rist, Michael, Carbonell, Guillermo, Cabrera, Benjamin, Kurowsky, Nils, und Wojatzki, Michael (2016). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In: *3rd Workshop on Natural Language Processing for Computer-Mediated Communication* (S. 6-9). doi:10.17185/dupublico/42132

Sap, Maarten, Card, Dallas, Gabriel, Saadia, Choi, Yejin, und Smith, Noah A. (2019). The Risk of Racial Bias in Hate Speech Detection. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (S. 1668-1678). doi:10.18653/v1/P19-1163

Schefczyk, Michael, und Schramme, Thomas (2015). Einleitung. In: M. Schefczyk, T. Schramme (Hrsg.), *John Stuart Mill: Über die Freiheit* (S. 1-10). Berlin, Boston: De Gruyter.

Schmidt, Anna, und Wiegand, Michael (2017). A Survey on Hate Speech Detection using Natural Language Processing. In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (S. 1-10). doi:10.18653/v1/W17-1101

Sponholz, Liriam (2018). *Hate Speech in den Massenmedien. Theoretische Grundlagen und empirische Umsetzung*. Wiesbaden: Springer VS. doi:10.1007/978-3-658-15077-8

Vidgen, Bertie, und Derczynski, Leon (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE* 15(12): e0243300. doi:10.1371/journal.pone.0243300

Waldron, Jeremy (2012). *The Harm in Hate Speech*. Cambridge, MA; London: Harvard University Press.

Waseem, Zeerak (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In: *Proceedings of 2016 EMNLP Workshop on NLP and Computational Social Science* (S. 138-142). doi:10.18653/v1/W16-5618



Waseem, Zeerak, und Hovy, Dirk (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In: *Proceedings of NAACL Student Research Workshop* (S. 88-93). doi:10.18653/v1/N16-2013

Wiegand, Michael, Ruppenhofer, Josef, und Kleinbauer, Thomas (2019). Detection of Abusive Language: the Problem of Biased Datasets. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (S. 602-608). doi:10.18653/v1/N19-1060

## B.5 STUDY IX

©2021 Association for Computational Linguistics, published under Creative Commons CC-BY 4.0 License<sup>4</sup>.

Edoardo Mosca, Maximilian Wich, and Georg Groh (June 2021). “Understanding and Interpreting the Impact of User Context in Hate Speech Detection.” In: *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. Online: Association for Computational Linguistics, pp. 91–102. DOI: [10 . 18653 / v1 / 2021 . socialnlp - 1 . 8](https://doi.org/10.18653/v1/2021.socialnlp-1.8). URL: [https : //aclanthology.org/2021.socialnlp-1.8](https://aclanthology.org/2021.socialnlp-1.8)

---

<sup>4</sup> <https://creativecommons.org/licenses/by/4.0/>

### *Publication Summary*

"As hate speech spreads on social media and online communities, research continues to work on its automatic detection. Recently, recognition performance has been increasing thanks to advances in deep learning and the integration of user features. This work investigates the effects that such features can have on a detection model. Unlike previous research, we show that simple performance comparison does not expose the full impact of including contextual- and user information. By leveraging explainability techniques, we show (1) that user features play a role in the model's decision and (2) how they affect the feature space learned by the model. Besides revealing that—and also illustrating *why*—user features are the reason for performance gains, we show how such techniques can be combined to better understand the model and to detect unintended bias." (Mosca, Wich, and Groh, 2021, p. 91)

### *Author Contributions*

Edoardo Mosca headed the research project, implemented the code, and wrote the paper. He and Maximilian Wich co-created the idea and conception of the research project. Furthermore, Maximilian Wich provided feedback. Georg Groh regularly discussed the ideas and concepts with the team and provided feedback on the study.

# Understanding and Interpreting the Impact of User Context in Hate Speech Detection

**Edoardo Mosca**

TU Munich,  
Department of Informatics,  
Germany

edoardo.mosca@tum.de

**Maximilian Wich**

TU Munich,  
Department of Informatics,  
Germany

maximilian.wich@tum.de

**Georg Groh**

TU Munich,  
Department of Informatics,  
Germany

grohg@in.tum.de

## Abstract

As hate speech spreads on social media and online communities, research continues to work on its automatic detection. Recently, recognition performance has been increasing thanks to advances in deep learning and the integration of user features. This work investigates the effects that such features can have on a detection model. Unlike previous research, we show that simple performance comparison does not expose the full impact of including contextual- and user information. By leveraging explainability techniques, we show (1) that user features play a role in the model’s decision and (2) how they affect the feature space learned by the model. Besides revealing that—and also illustrating *why*—user features are the reason for performance gains, we show how such techniques can be combined to better understand the model and to detect unintended bias.

## 1 Introduction

Communication and information exchange between people is taking place on online platforms at a continuously increasing rate. While these means allow everyone to express themselves freely at any time, they are massively contributing to the spread of negative phenomena such as online harassment and abusive behavior. Among those, which are all to discourage, online hate speech has attracted the attention of many researchers due to its deleterious effects (Munro, 2011; Williams et al., 2020; Duggan, 2017).

The extremely large volume of online content and the high speed at which new one is generated exclude immediately the chance of content moderation being done manually. This realization has naturally captured the attention of the *Machine Learning* (ML) field, seeking to craft automatic and scalable solutions (MacAvaney et al., 2019; Waseem et al., 2017; Davidson et al., 2017).

Methods for detecting hate speech and similar abusive behavior have been thus on the rise, consistently improving in terms of performance and generalization (Schmidt and Wiegand, 2017; Mishra et al., 2019b). However, even the current state of the art still faces limitations in accuracy and is yet not ready to be deployed in practice. Hate speech recognition remains an extremely difficult task (Waseem et al., 2017), in particular when the expression of hate is implicit and hidden behind figures of speech and sarcasm.

Alongside language features, recent works have considered utilizing user features as an additional source of knowledge to provide detection models with context information (Fehn Unsvåg and Gambäck, 2018; Ribeiro et al., 2018). As a general trend, models incorporating context exhibit improved performance compared to their pure text-based counterparts (Mishra et al., 2018, 2019a). Nevertheless, the effect, which these additional features have on the model, has not been interpreted or understood yet. So far, models have mostly been compared only in terms of performance metrics. The goal of this work is to shed light on the impact generated by including user features—or more in general context—into hate speech detection methods. Our methodology heavily relies on a combination of modern techniques coming from the field of *eXplainable Artificial Intelligence* (XAI).

We show that adding user and social context to models is the reason for performance gains. We also explore the model’s learned features space to understand how such features are leveraged for detection. At the same time, we discover that models incorporating user features suffer less from bias in the text. Unfortunately, those same models contain a new type of bias that originates from adding user information.

## 2 Related Work

### 2.1 Explainability for Recognition Models

A limited amount of research has focused on applying XAI techniques to the hate speech recognition case. For instance, Wang (2018) adapts a number of explainability techniques from the computer vision and applies them to a hate speech classifier trained on Davidson et al. (2017). Feature occlusion was used to highlight the most relevant words for the final classifier prediction and activation maximization selected the terms that the classifier captured and judged as relevant at a dataset-level. Vijayaraghavan et al. (2019) constructs an interpretable multi-modal detector that uses text alongside social and cultural context features. The authors leverage attention scores to quantify the relevance of different input features. Wich et al. (2020) applies post-hoc explainability on a custom dataset in German to expose and estimate the impact of political bias on hate speech classifiers. More in detail, left- and right-wing political bias within the training data is visualized via DeepSHAP-based explanations (Lundberg and Lee, 2017).

MacAvaney et al. (2019) combines together multiple simple classifiers to assemble a transparent model. Risch et al. (2020) reviews and compares several explainability techniques applied to hate speech classifiers. Their experimentation includes popular post-hoc approaches such as LIME (Ribeiro et al., 2016) and LRP (Bach et al., 2015) as well as self-explanatory detectors (Risch et al., 2020).

For our use case, we apply *post-hoc explainability* approaches (Lipton, 2018). We use external techniques to explain models that would otherwise be black-boxes (Arrieta et al., 2020). In contrast, *transparent models* are interpretable thanks to their intuitive and simple design.

### 2.2 Context Features for Hate Speech Detection

Models have been continuously improving since the first documented step towards automatic hate speech detection Spertus (1997). The evolution of recognition approaches has been favored by advances in *Natural Language Processing* (NLP) research (Mishra et al., 2019b). For instance, s.o.t.a detectors like Mozafari et al. (2020) exploit high-performing language models such as BERT (Devlin et al., 2019).

A different research branch took an alternative

path and explored the inclusion of social context alongside text. These additional features are usually referred to with the terms *user features*, *context features*, or *social features*. Some tried incorporating the gender (Waseem, 2016) and the profile’s geolocation and language (Galán-García et al., 2016). Others instead utilized the user’s number of followers or friends (Fehn Unsvåg and Gambäck, 2018).

Modeling users’ social and conversational interactions via their corresponding graph was also shown to be rewarding (Mishra et al., 2019b; Cécillon et al., 2019). Ribeiro et al. (2018) creates additional features by measuring properties like betweenness and eigenvector centrality. Mishra et al. (2018) and Mishra et al. (2019a) instead fed the graph directly to the model either embedded as matrix or via using graph convolutional neural network (Hamilton et al., 2017).

While previous work explored the usage of a wide range of context features (Fehn Unsvåg and Gambäck, 2018), detection models have only been compared in terms of performance metrics. Besides accuracy, researchers have not focused on other changes that such features could have on the model. Our work shows that indeed this addition entails a large impact on the recognition algorithm’s behavior and substantially changes its characteristics.

## 3 Experimental Setup

In this section, we describe in detail the different datasets and detection models that we include in our interpretability-driven analysis.

### 3.1 Data and Preprocessing

Previous research has produced several datasets to support further developments in the hate speech detection area (Founta et al., 2018; Warner and Hirschberg, 2012). Some became relatively popular to benchmark and test new ideas and improvements in recognition techniques. For our experimentation, we pick the DAVIDSON (Davidson et al., 2017) and the WASEEM (Waseem and Hovy, 2016) datasets. The choice was motivated by their variety of speech classes and popularity as detection benchmarks.

Both benchmarks consist of a collection of tweets coupled with classification tasks with three possible classes. DAVIDSON contains  $\sim 25,000$  tweets of which 1,430 are labeled as *hate*, 19,190 as *offensive*, and 4,163 as *neither* (Davidson et al., 2017). As classification outcomes in WASEEM in-

stead, we have *racism*, *sexism*, and *neither*. The three classes contain 3,378, 1,970, and 11,501 tweets respectively (Waseem and Hovy, 2016). We were not able to retrieve the remaining 65 of the original 16,914 samples.

We follow the same preprocessing steps for both datasets. First, terms belonging to categories like *url*, *email*, *percent*, *number*, *user*, and *time* are annotated via a category token. For instance, “341” is replaced by “<number>”. After that, we apply word segmentation and spell correction based on Twitter word statistics. Both methods and statistics were provided by the *ekphrasis*<sup>1</sup> text preprocessing tool (Baziotis et al., 2017).

In addition to the tweets that represent the text (or content) component of our input features, we also retrieve information about the tweet’s authors and their relationships. In a similar fashion as done in Mishra et al. (2018), we construct a *community graph*  $G = (V, E)$  where each node represents a user and two nodes are connected if at least one of the two users follows the other one. We were able to retrieve  $|V| = 6,725$  users and  $|E| = 19,597$  relationships for DAVIDSON, while for WASEEM we have  $|V| = 2,024$  and  $|E| = 9,955$ .

The respective average node degrees are 2,914 and 4,918 and the overall graphs’ densities:

$$D = \frac{2 \cdot |E|}{|V|(|V| - 1)}$$

are 0.00087 and 0.00486 respectively.

We immediately notice that both graphs are very sparse. In particular, we have 3,393 users not connected to anyone in DAVIDSON and 927 in WASEEM. For reference, Mishra et al. (2018) achieves a graph density of 0.0075 on WASEEM, with only  $\sim 400$  authors being solitary, i.e. with no connections. We assume the difference is reasonable as data availability considerably decreases over time.

### 3.2 Detection Models

Our experimentation and findings are based on the comparison of two detection models, one that solely relies on text features and one that instead incorporates context features. To better capture their behavioral differences, we build them to be relatively simple and also to not differ in the text-processing part.

<sup>1</sup><https://github.com/cbaziotis/ekphrasis>

The first model, shown in figure 1, computes the three classification probabilities only based on the tweets’ content. The input text is fed to the model as *Bag of Words* (BoW), which is then processed by two fully connected layers. We refer to this model as *text model*.

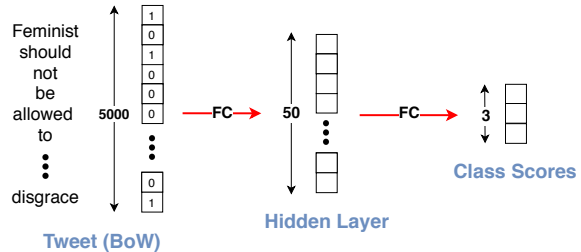


Figure 1: Architecture of the text model.

The second model instead leverages the information coming from three input sources: the tweet’s text, the user’s vocabulary, and the follower network. The first input is identical to what is fed to the text model. The second is constructed from all the tweets of the author in the dataset and aims to model their overall writing style. Concretely, we merge the tweets’ BoW representations, i.e. we apply a logical-OR to their corresponding vectors. The third is the author’s follower network and describes their online surrounding community. On a more technical note, this can be extracted as a row from the adjacency matrix of our community graph described in section 3.1. Note that s.o.t.a hate speech detector used similar context features (Mishra et al., 2018, 2019a). We refer to this model as *social model*.

As sketched in figure 2, the different input sources are initially processed separately in the model’s architecture. After the first layer, the intermediate representations from the different branches are concatenated together and fed to two more layers to compute the final output. Note that the text- and social models have the same dimensions for their final hidden layer and can be seen as equivalent networks working on different inputs.

## 4 Proposed Analysis

We now describe our methodology in detail. Recall that our models differ precisely on the usage of user features. As we will see shortly, their comparison beyond accuracy measurements sheds light on the different model properties and hence on the potential impact of incorporating context features.

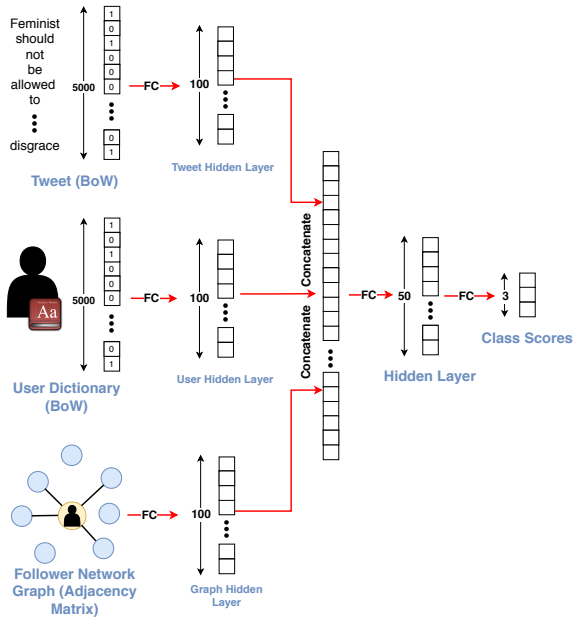


Figure 2: Architecture of the social model.

#### 4.1 Training and Performance

We apply the same training and testing procedure to all models and datasets. We keep the 60% of the data for training while splitting the remaining equally between validation and test set, i.e. 20% each.

Tables 1 and 2 report our results in terms of F1 scores for WASEEM (Waseem and Hovy, 2016) and DAVIDSON (Davidson et al., 2017) respectively. To increase our confidence in their validity, we average the performance over five runs with randomly picked train/validation/test sets. We observe different trends for the two datasets.

Speech Class	Text Model	Social Model
Racism	0.711	0.735
Sexism	0.703	0.832
Neither	0.881	0.907
Overall	0.829	0.872

Table 1: F1 Scores on Waseem and Hovy (2016).

On WASEEM, the social model considerably outperforms (by 4.3%) our text model. The performance gain is general and not restricted to any single class. Quite surprisingly, our text model performs better on racist tweets than sexist ones, although the sexism class is almost twice as big. This suggests that sexism is, at least in this case, somewhat harder to detect by just looking at the tweet content. On the contrary, our social model shows an impressive improvement in the sexism class (al-

most 13%), suggesting the presence of detectable patterns in sexist users and their social interactions.

Speech Class	Text Model	Social Model
Hate	0.154	0.347
Offensive	0.939	0.939
Neither	0.809	0.815
Overall	0.876	0.886

Table 2: F1 Scores on Davidson et al. (2017).

On DAVIDSON, we only observe a contained improvement (1%). Moreover, the jump in performance is restricted to the hate class, containing a tiny amount of samples. We believe the difference between the two datasets should be expected due to the lower amount of user data available for DAVIDSON. Considering these results, we focus on applying our technique on the WASEEM dataset in the remainder of this paper. Nevertheless, the respective results on DAVIDSON can be found in the appendix A. While on both datasets we do not outperform the current s.o.t.a—Mishra et al. (2019a) on WASEEM and Mozafari et al. (2020) on DAVIDSON—our results are comparable and thus satisfactory for our purposes.

#### 4.2 Shapley Values Estimation

We now apply a first post-hoc explainability method. For each feature we calculate its corresponding *Shapley value* (Shapley, 1953; Lundberg and Lee, 2017). That is, we quantify the relevance that each feature has for the prediction of a specific output. Shapley values have been shown—both theoretically and empirically—to be an ideal estimator for feature relevance (Lundberg and Lee, 2017).

As exact Shapley values are exponentially complex to determine, we use accurate approximation methods as done in (Lundberg and Lee, 2017; Štrumbelj and Kononenko, 2014). Figure 3 shows concrete examples in which Shapley values are calculated for both models on two test tweets from WASEEM.

For our social model, we consider the user vocabulary and the follower network as single features for simplicity. Notably, the context is used by the social model and can play a significant role in its prediction. Hence, we can confirm the context features to be the reason for the performance gains. We can empirically exclude that the differences between the text- and the social model architectures justify the jump in performance.

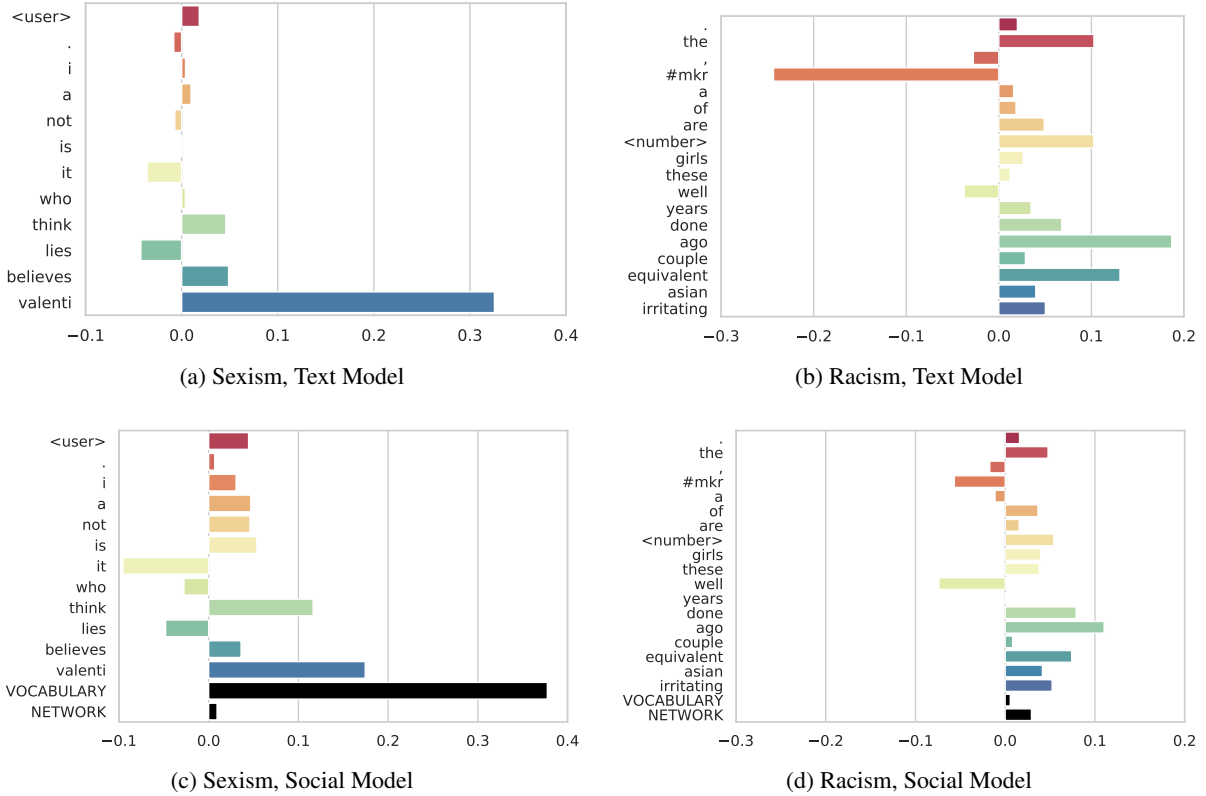


Figure 3: Example of features contribution, computed via Shapley value approximation, for our text and social models. In (a) and (c) we use as input the tweet “<user> I think Arquette is a dummy who believes it. Not a Valenti who knowingly lies.”. The sexist tweet refers to the actress Patricia Arquette, who spoke in favour of gender equality, and the feminist writer Jessica Valenti. Some words are missing in the plot as our BoW dimension is limited during preprocessing. In (b) and (d), we use the racist tweet “These girls are the equivalent of the irritating Asian girls a couple of years ago. Well done, 7. #MKR”. The hashtag refers to the Australian cooking show “My Kitchen Rules”.

### 4.3 Feature Space Exploration

We have seen that detection models can benefit from the inclusion of context features. We now focus on understanding *why* this is the case. Shapley values and more in general feature attribution methods can quantify *how much* single features contribute to the prediction. Yet, alone, they do not give us any intuition to answer our why-question.

We look at the feature space learned by our models, which can be considered a global explainability technique. For our text model, we remove the last layer and feed the tweets to the remaining architecture. The output is a 50-dimensional embedding for each tweet. We employ the *t-Distributed Stochastic Neighbor Embedding* (t-SNE) (Van der Maaten and Hinton, 2008) to reduce the embeddings to two dimensions for visualization purposes.

The resulting plot, in figure 4d, shows all the tweets in a single cluster. Racist tweets look more concentrated in one area than sexist ones, suggest-

ing that sexism is somewhat harder to detect for the model. This result is coherent with our per-class performance scores.

We apply the same procedure to the social model. In this case, we visualize the hidden layer of each separate branch as well as the final hidden layer analogous to the text model. Not surprisingly, the tweet branch (figure 4a) looks very similar to the feature space learned by our text model. The user’s vocabulary branch (figure 4b) instead shows the samples distributed in well-separated clusters. Notably, racist tweets have been restricted to one cluster and we can also observe pure-sexist and pure-neither clusters. The follower network branch (figure 4c) looks similar though cluster separation is not as strong. Once more, we notice racism more concentrated than sexism, which is considerably more mixed with regular tweets. To some extent, this result is in line with the notion of *homophily* among racist users (Mathew et al., 2019).



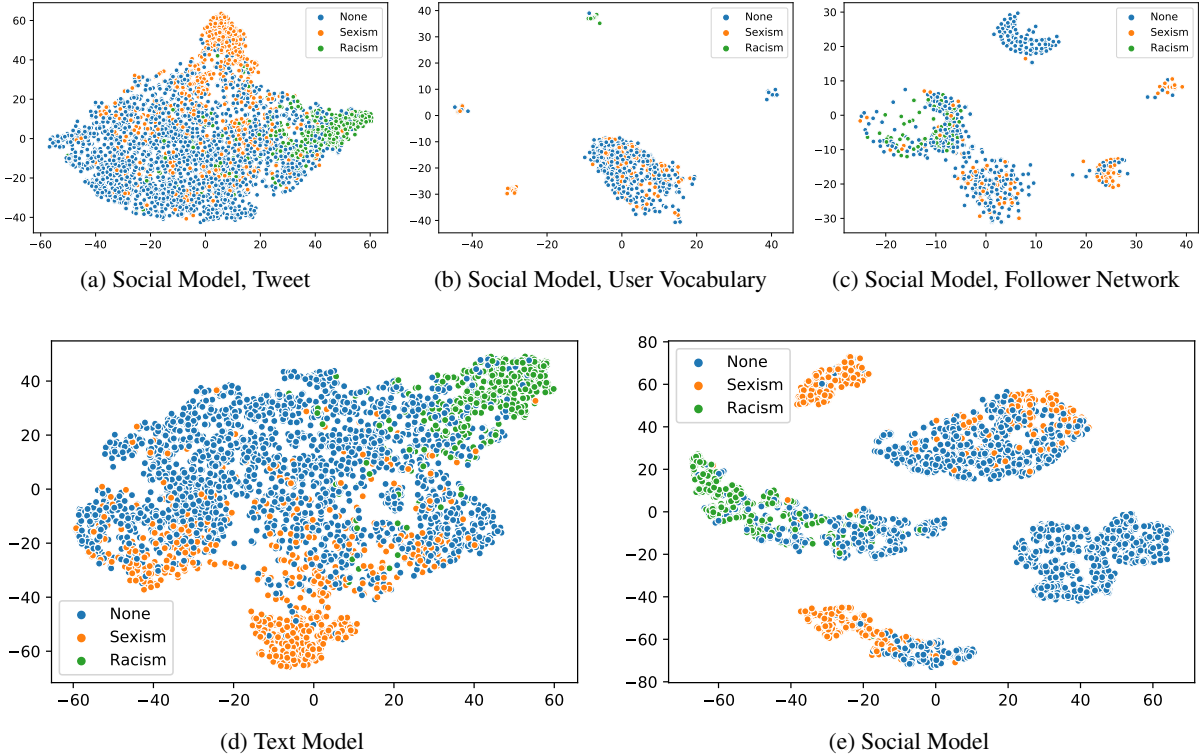


Figure 4: WASEEM tweets, colored by label, in the features space learned by our text model (d) and social model (a,b,c for the independent branches, e combined).

Intuitively, being able to divide users into different clusters based on their behavior should be helpful for classification at later layers. This is confirmed by the combined feature space plot (figure 4e). Indeed, tweets are now structured in multiple clusters instead of a single one as for our text model. Also in this case, we observe several pure or almost-pure groups.

The corresponding visualizations and results for DAVIDSON can be found in appendix A.

#### 4.4 Targeted Behavioral Analysis: Explaining a Novel Tweet

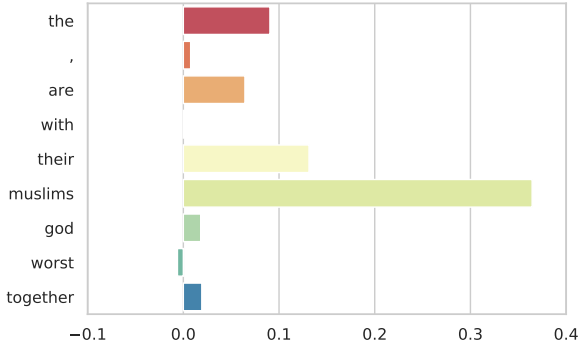
We have seen how different explainability techniques convey different types of information on the examined model. Computing Shapley values and visualizing the learned feature space can also be used in combination as they complement each other. If used together, they can both quantify the relevance of each feature as well as show how certain types of features are leveraged by the model to better distinguish between classes.

So far, our explanations are relative to the datasets used for model training and testing. However, to better understand a classifier it should also be tested beyond its test set. This can be sim-

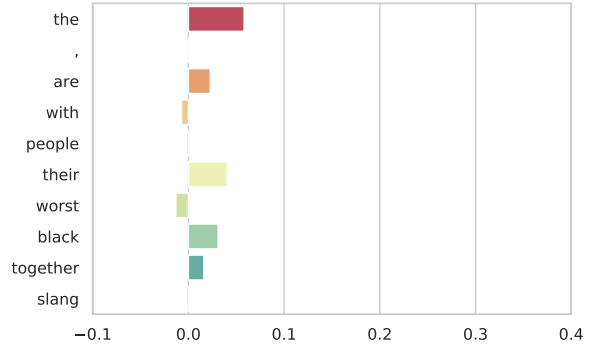
ply done by feeding the model with a novel tweet. Via artificially crafting tweets, we can check the model’s behavior in specific cases. For instance, we can inspect how it reacts to specific sub-types of hate.

Let us consider the anti-Islamic tweet “*muslims are the worst, together with their god*”. If fed to our model, it is classified as racist with a 75% confidence following our expectations. Figures 5a and 5c show explanations for the tweet. We can see that the word “*muslim*” plays a big role by looking at its corresponding Shapley value. At the same time, the projection of the novel tweet onto the feature space shows how the sample is collocated together with the other racist tweets by the text model.

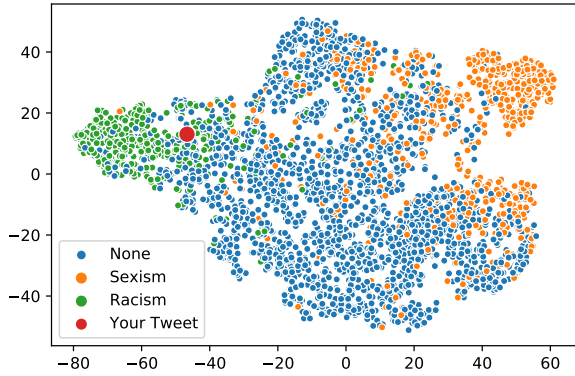
If we now change our hypothetical tweet to be anti-black—“*black people are the worst, together with their slang*”—we observe a different model behavior (figures 5b and 5d). In fact, now the tweet is not classified as racist. No word has a substantial impact on the prediction. We can also notice a slight shift of the sample in the features space, away from the racism cluster. If changing the target of the hate changes the prediction, then the model/dataset probably contains bias against that target. Model interpretability further reveals how



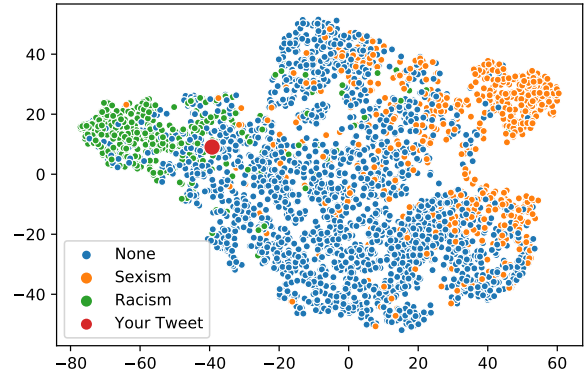
(a) Anti-Islam, Shapley Values



(b) Anti-Black, Shapley Values



(c) Anti-Islam, Embedding in Latent Space



(d) Anti-Black, Embedding in Latent Space

Figure 5: Features contribution (Shapley values w.r.t. the racism class) and embedding in the text model’s latent space of an islamophobic and a anti-black racist tweets. The two sentences had, according to our text model, the 75% and 24% probability of being racist respectively.

its behavior reacts to different targets.

We run the same experiment with our social model. This time, it correctly classifies the anti-black tweet as racist (55% confidence). This suggests that text bias could be mitigated by using models that do not only rely on the text input. However, the social model is much more sensitive to changes in the user-derived features. To test this, we feed the model the same tweet and only change the author that generated it. For a fair comparison, we pick one random user with other racist tweets, one random user with other sexist tweets, and one random user with no hateful tweets in the dataset. We refer to these users as racist, sexist, and regular users respectively.

Our crafted tweet is classified as racist when coming from a racist user (64%). However, it is instead judged non-hateful in both the other cases (12% and 19% for a sexist and user with no hate background respectively). Evidently, racist tweets also need some contribution from the social features to be judged as racist.

A very informative explanation comes again from both the Shapley values and the feature space exploration (figure 6). On the left side, we can see the Shapley value for the racist and regular users. Results relative to the sexist user are analogous to the regular user and reported in the supplementary material (A.3). All the words have a similar contribution to the racism class in all cases. However, the difference in the authors plays a substantial role in the decision. Only the racist user positively contributes to the racism class. On the right side of 6, we can see the embedding in the latent space for each case. Different input authors cause the tweet to be embedded in different clusters. Only in the first one the model actually considers the possibility of the tweet being racist.

Hence, while adding user-derived features might mitigate the effects of bias in the text, it generates a new form of bias that could discriminate users based on their previous behavior and hinder the model from classifying correctly hateful content.

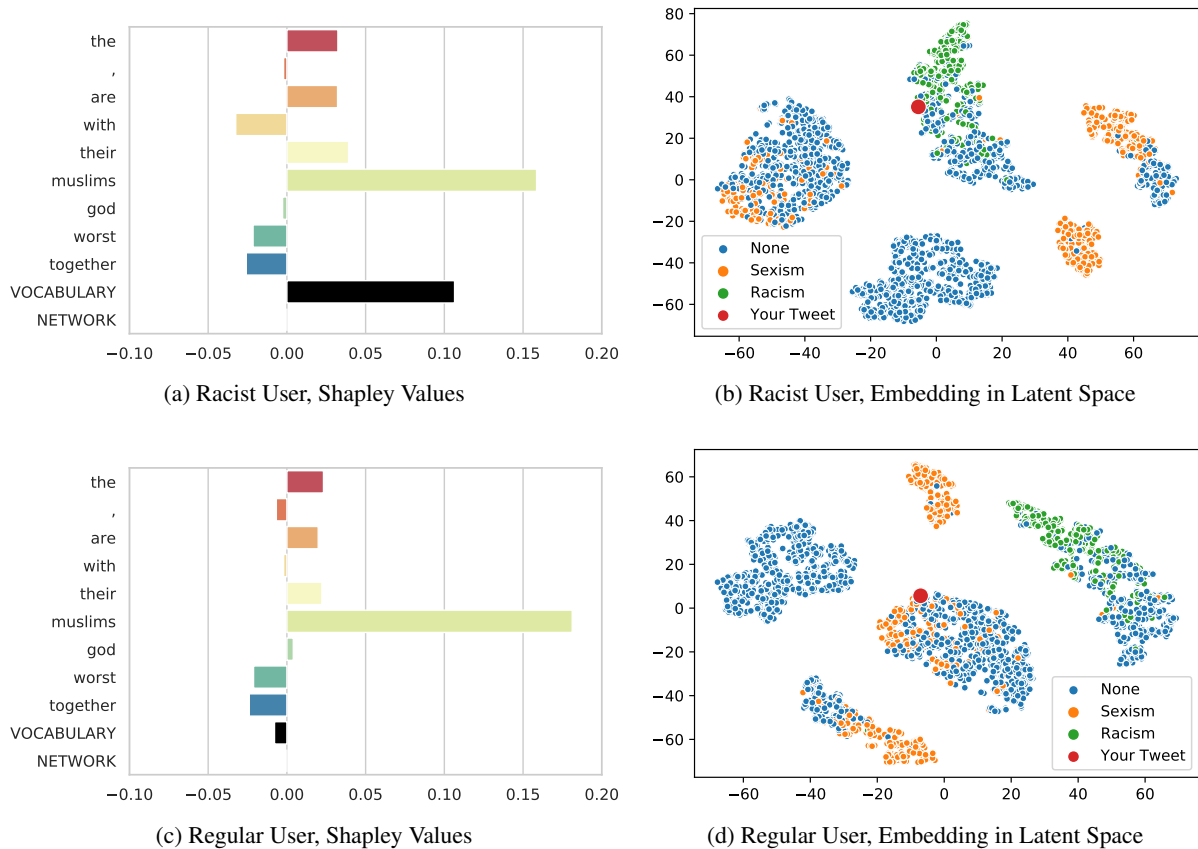


Figure 6: Features contribution (w.r.t. racism class) and embeddings of the islamophobic tweet in the social model’s latent space. The two pairs of plots are w.r.t. two predictions done with different users as input: a racist one (a,b, 64%), and a regular one (c,d, 19%).

## 5 Conclusion and Future Work

In our work, we investigated the effects of user features in hate speech detection. In previous studies, this was done by comparing models based on performance metric. We have shown that post-hoc explainability techniques provide a much deeper understanding of the models’ behavior. In our case, when applied to two models that differ specifically on the usage of context features, the in-depth comparison reveals the impact that such additional features can have.

The two utilized techniques—*Shapley values estimation* and *learned feature space exploration*—convey different kinds of information. The first one quantifies how each feature plays a role but does not tell us what is happening in the background. The second one illustrates the model’s perception of the tweets but does not provide any quantitative information for the prediction. Furthermore, we have seen that artificially crafting and modifying a tweet can be useful to examine the models’ behavior in particular scenarios. In concrete exam-

ples, the two approaches worked as bias detectors present in the text as well as in the user features.

We believe that analyzing detection models is vital for understanding how certain features shape the way data is processed. Accuracy alone is by no means a sufficient metric to decide which model to prefer. Our work shows that even models that perform significantly better can potentially lead to new types of bias. We urge researchers in the field to compare recognition approaches beyond accuracy to avoid potential harm to affected users.

Data scarcity is still a main issue faced by current researchers, especially when it comes to context features. We believe that larger and more complete datasets will improve our understanding of how certain features interact and will help future research in advancing both in accuracy and bias mitigation.

## Acknowledgments

This paper is based on a joined work in the context of Edoardo Mosca’s master’s thesis (Mosca, 2020).

## References

- Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bénézet, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.
- Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS one*, 10(7).
- Christos Baziotis, Nikos Pelekis, and Christos Doukouridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754.
- Noé Cecillon, Vincent Labatut, Richard Dufour, and Georges Linares. 2019. Abusive language detection in online conversations by combining content- and graph-based features. In *ICWSM International Workshop on Modeling and Mining Social-Media-Driven Complex Networks*, volume 2, page 8. Frontiers.
- Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Maeve Duggan. 2017. *Online harassment 2017*. Pew Research Center.
- Elise Fehn Unsvåg and Björn Gambäck. 2018. The Effects of User Features on Twitter Hate Speech Detection. In *Proc. 2nd Workshop on Abusive Language Online*, pages 75–85.
- Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proc. 11th ICWSM*, pages 491–500.
- Patxi Galán-García, José Gáviria de la Puerta, Carlos Laorden Gómez, Igor Santos, and Pablo García Bringas. 2016. Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Logic Journal of the IGPL*, 24(1):42–53.
- Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Zachary C Lipton. 2018. The mythos of model interpretability. *Queue*, 16(3):31–57.
- Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774.
- Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).
- Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PLoS one*, 14(8).
- Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182.
- Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author profiling for abuse detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1088–1098.
- Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2019a. Abusive language detection with graph convolutional networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 2145–2150.
- Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019b. Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024*.
- Edoardo Mosca. 2020. Explainability of hate speech detection models. Master’s thesis, Technical University of Munich. Advised and supervised by Maximilian Wich and Georg Groh.
- Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. *Studies in Computational Intelligence*, 881 SCI:928–940.
- Emily R Munro. 2011. The protection of children online: a brief scoping review to identify vulnerable groups. *Childhood Wellbeing Research Centre*.

- Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Twelfth international AAAI conference on web and social media*.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining*, pages 1135–1144.
- Julian Risch, Robin Ruff, and Ralf Krestel. 2020. Offensive language detection explained. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 137–143.
- Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proc. 5th Intl. Workshop on Natural Language Processing for Social Media*, pages 1–10.
- Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.
- Ellen Spertus. 1997. Smokey: Automatic recognition of hostile messages. In *Proceedings of Innovative Applications of Artificial Intelligence (IAAI)*, pages 1058–1065.
- Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665.
- Prashanth Vijayaraghavan, Hugo Larochelle, and Deb Roy. 2019. [Interpretable Multi-Modal Hate Speech Detection](#). In *Intl. Conf. Machine Learning AI for Social Good Workshop*.
- Cindy Wang. 2018. Interpreting neural network hate speech classifiers. In *Proc. 2nd Workshop on Abusive Language Online*, pages 86–92.
- William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.
- Zeeraq Waseem. 2016. [Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter](#). In *Proc. First Workshop on NLP and Computational Social Science*, pages 138–142.
- Zeeraq Waseem, Thomas Davidson, Dana Warmusley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.
- Zeeraq Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.
- Maximilian Wich, Jan Bauer, and Georg Groh. 2020. Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 54–64.
- Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2020. Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, 60(1):93–117.

## A Results on the Davidson Dataset

### A.1 Feature Space learned by the Text Model

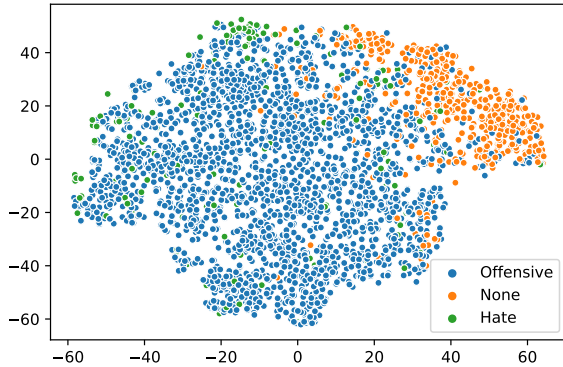


Figure 7: DAVIDSON tweets, colored by label, in the feature space learned by the text model.

Figure 7 shows the feature space learned by our text model on DAVIDSON. Overall, the distribution looks similar as the one of WASEEM visualized in figure 4d. We can notice that hate tweets are extremely sparse and mixed with the offensive ones. This is reflected by the poor model performance on the hate class, possibly caused by the conceptual overlap that these two classes have. On the other hand, non-harmful tweets are mostly concentrated in one area of the plot, confirming the satisfactory F1 score achieved.

### A.2 Feature Space learned by the Social Model

Figure 8 shows the feature space learned by our social model on DAVIDSON. As done for WASEEM, we report the plots both for the single branches as well as for their combination. The tweet branch (figure 8a) has a similar structure to figure 7. However, hateful tweets are also concentrated in a small portion of the space. This reflects the improved performance that the social model had on the hate class. This suggests that the information coming from the other input sources reinforces the signal backpropagated to the tweet branch, resulting in a less chaotic mixture of hateful and offensive tweets. The user vocabulary (figure 8b) and the follower network branch (figure 8c) do not present the same characteristics as seen on WASEEM. In this case, we do not have the data points separated into multiple clusters. The same goes for the overall learned feature space (figure 8d), where all the tweets are contained in one single cloud. This is consistent with what we observed in terms of F1 Scores. In

contrast to what occurred on WASEEM, user features did not cause a substantial impact on the feature space on DAVIDSON and thus did not produce a large leap in performance.

### A.3 Complement to Figure 6

Figure 6 compares the model’s behavior on the same tweet but with different authors, one racist and one regular. For completeness, figure 9 shows the corresponding plots—Shapley values and embedding onto the features space—for the same tweet when generated by a sexist user. The result is analogous to the one obtained with the regular user. Also in this case the tweet is not classified as racist (12% confidence). The estimated Shapley values show a substantial impact of the user vocabulary against the racism class. The embedding onto the latent space shows once more that changing the author caused the tweet to embed in a different cluster, hence excluding the possibility of the content being classified correctly.

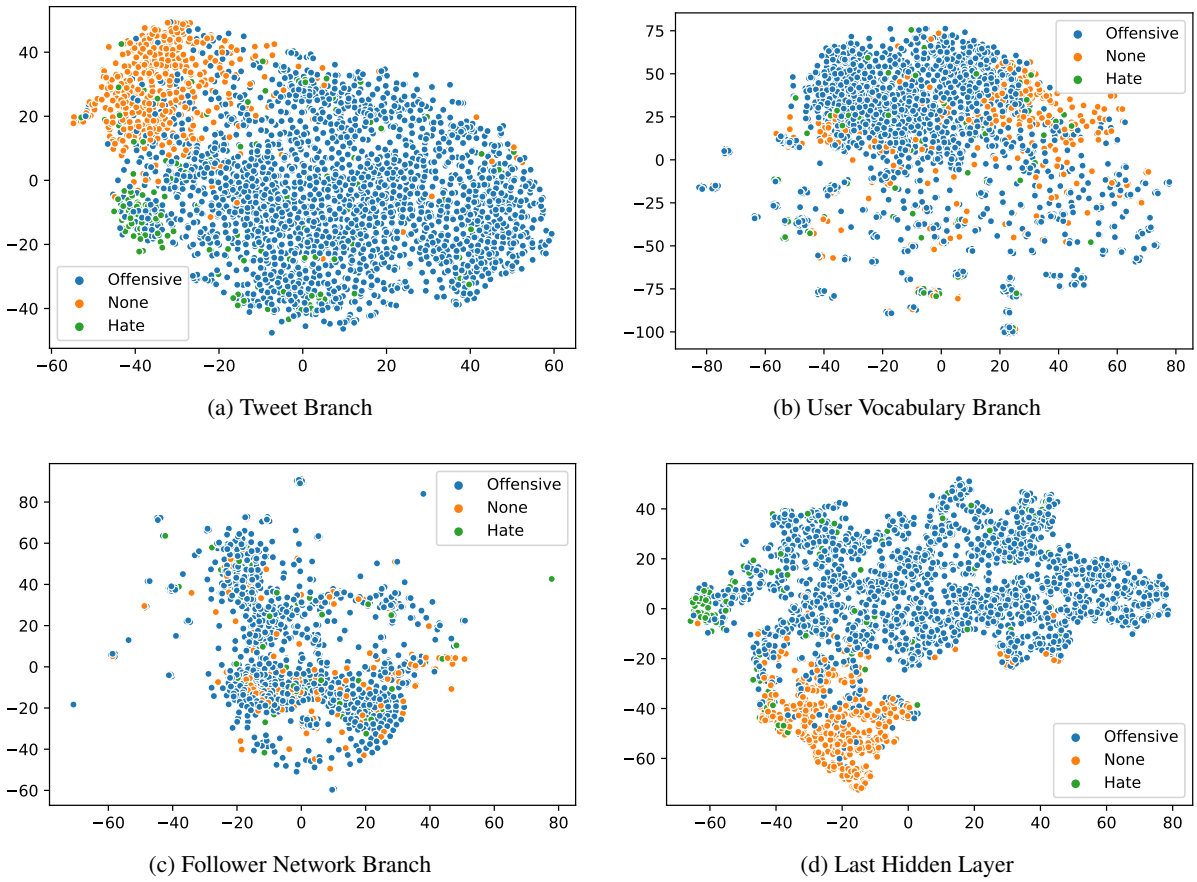


Figure 8: Latent space visualization of our social model on DAVIDSON, colored by label. The features are extracted from the single branches before the concatenation: tweet (a), user’s vocabulary (b), follower network (c). The last plot (d) shows instead the final learned features space, after all branches are combined and processed together.

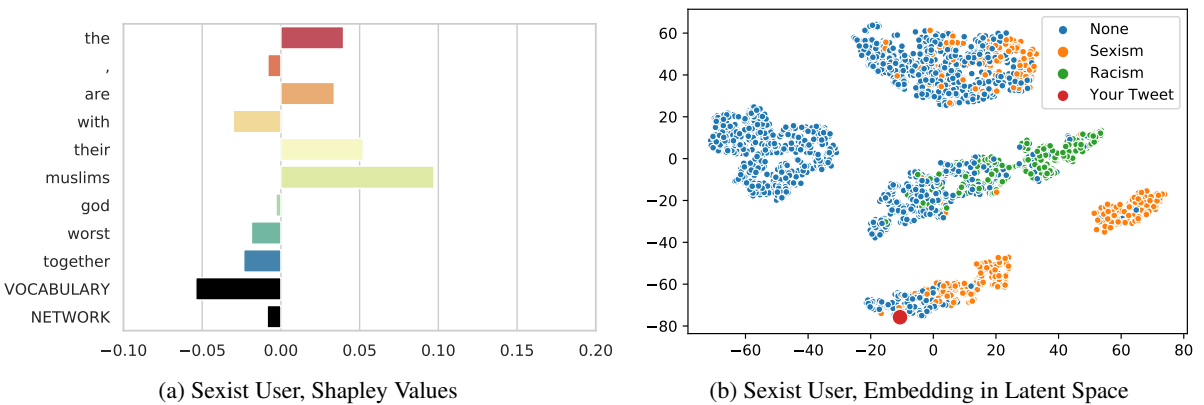


Figure 9: Features contribution (w.r.t. racism class) and embeddings of the islamophobic tweet in the social model’s latent space. The pair of plots are w.r.t. the prediction done with sexist author.

B.6 STUDY XII

©2022 Association for the Advancement of Artificial Intelligence

Reprinted with permission from:

Maximilian Wich, Adrian Gorniak, Tobias Eder, Daniel Bartmann, Burak Enes Çakici, and Georg Groh (May 2022). "Introducing an Abusive Language Classification Framework for Telegram to Investigate the German Hater Community." In: *Proceedings of the International AAAI Conference on Web and Social Media* 16.1, pp. 1133–1144. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/19364>

This thesis includes the accepted version of publication and not the final published version.



### *Publication Summary*

"Because traditional social media platforms continue to ban actors spreading hate speech or other forms of abusive languages (a process known as deplatforming), these actors migrate to alternative platforms that do not moderate user content to the same degree. One popular platform relevant for the German community is Telegram for which limited research efforts have been made so far. This study aimed to develop a broad framework comprising (i) an abusive language classification model for German Telegram messages and (ii) a classification model for the hatefulness of Telegram channels. For the first part, we use existing abusive language datasets containing posts from other platforms to develop our classification models. For the channel classification model, we develop a method that combines channel-specific content information collected from a topic model with a social graph to predict the hatefulness of channels. Furthermore, we complement these two approaches for hate speech detection with insightful results on the evolution of German speaking communities focused on hateful content on the Telegram platform. We also propose methods for conducting scalable network analyses for social media platforms to the hate speech research community. As an additional output of this study, we provide an annotated abusive language dataset containing 1,149 annotated Telegram messages." (Wich, Gorniak, et al., 2022, p. 1)

### *Author Contributions*

Maximilian Wich headed the research project. He developed the initial idea, the concept, and the methodology of the study. He collected the dataset, developed substantial parts of the code, and participated in the annotation process. Besides, he wrote a significant portion of the manuscript. Adrian Gorniak contributed to research questions 3 and 4 by implementing parts of the code and conducting analysis. He participated in the annotation process. In addition, he wrote smaller parts of the paper and provided feedback on the content. Tobias Eder reviewed the research methodology, provided feedback on the study's content, and proofread the draft. Daniel Bartmann and Burak Enes Çakici implemented parts of the code to classify Telegram messages as part of their NLP Lab Course project supervised by Maximilian Wich. Furthermore, Daniel Bartmann substantially contributed to the annotation process of selected Telegram messages. Georg Groh regularly discussed the ideas and concepts with the team and provided feedback on the manuscript.

# Introducing an Abusive Language Classification Framework for Telegram to Investigate the German Hater Community

Maximilian Wich<sup>1</sup>, Adrian Gorniak<sup>1</sup>, Tobias Eder<sup>1</sup>,  
Daniel Bartmann<sup>1</sup>, Burak Enes Çakici<sup>1</sup>, Georg Groh<sup>1</sup>

<sup>1</sup> Technical University of Munich, Munich, Germany  
maximilian.wich@tum.de, adrian.gorniak@tum.de, tobias.eder@in.tum.de,  
daniel.bartmann@in.tum.de, burak-enes.cakici@tum.de, grohg@in.tum.de

## Abstract

Because traditional social media platforms continue to ban actors spreading hate speech or other forms of abusive languages (a process known as deplatforming), these actors migrate to alternative platforms that do not moderate user content to the same degree. One popular platform relevant for the German community is Telegram for which limited research efforts have been made so far.

This study aimed to develop a broad framework comprising (i) an abusive language classification model for German Telegram messages and (ii) a classification model for the hatefulness of Telegram channels. For the first part, we use existing abusive language datasets containing posts from other platforms to develop our classification models. For the channel classification model, we develop a method that combines channel-specific content information collected from a topic model with a social graph to predict the hatefulness of channels. Furthermore, we complement these two approaches for hate speech detection with insightful results on the evolution of German speaking communities focused on hateful content on the Telegram platform. We also propose methods for conducting scalable network analyses for social media platforms to the hate speech research community. As an additional output of this study, we provide an annotated abusive language dataset containing 1,149 annotated Telegram messages.

## Introduction

Hate speech and other forms of abusive language are a severe challenge that social media platforms, such as Facebook, Twitter, and YouTube, are facing nowadays (Duggan 2017). Moreover, this problem is not only limited to the online world; studies have shown that online hate correlates with physical crimes in the real world (Müller and Schwarz 2021; Williams et al. 2020), making the phenomenon a societal challenge for everybody.

To enforce a fast reaction to harmful content on social media platforms, Germany has passed a set of laws (Network Enforcement Act or NetzDG) to force social media companies to take action against hate speech on their platforms (Rafael 2019; Echikson and Knodt 2018). These actions range from deleting single posts that contain hateful content to banning actors from the platform, which is called deplatforming (Fielitz and Schwarz 2020). While deplatforming helps limit the reach of these hate actors (Fielitz and Schwarz 2020), it often makes them migrate to less or

un-regulated platforms and continue their hateful communication (Rogers 2020; Fielitz and Schwarz 2020; Urman and Katz 2020); one such alternative social media platform is Telegram (Rogers 2020; Fielitz and Schwarz 2020; Urman and Katz 2020). In Germany, Telegram has become the focal point for right-wing extremists, conspiracy theorists, and COVID-19 deniers (Fielitz and Schwarz 2020; Urman and Katz 2020; Eckert, Leipertz, and Schmidt 2021). Along with this rapid increase in popularity and usage by various user types, two important challenges regarding abusive language detection arise: first, the automatic detection of abusive content in such texts and, second, an aggregated view on the account level to identify hateful accounts. For both challenges, we propose a machine learning-based approach.

Previously, most research efforts on detecting hate speech focused on posts and comments from Twitter and Facebook (Ross et al. 2016; Bretschneider and Peters 2017; Struß et al. 2019; Wiegand, Siegel, and Ruppenhofer 2018; Mandl et al. 2019, 2020; Wich, Räther, and Groh 2021; Wich et al. 2021a) with very little focus on Telegram. This is particularly the case for content in German. At the same time, Telegram channels and chat groups are known for being a prime driving factor of online hate within the German language community. We want to bridge this gap and build abusive language classification models for Telegram messages. Because there is no abusive language dataset available that contains labeled Telegram messages in German, our approach is to use existing abusive language datasets in German, collected from other platforms and construct a classification model for Telegram. This leads to the first research question for this study:

**RQ1** Can existing abusive language datasets from other platforms be used to develop an abusive language classification model for Telegram messages?

Because the development of an abusive language classification model requires significant amounts of data, we collected such data from the platform (Telegram) over a longer period of time. By collecting the data, we are also able to formulate additional questions about the type of content and its spread on the platform. Because there is little research on these types of communication channels and their content, we were also interested in how this content has changed over the observed time period, during which deplatforming was

occurring on other social media. Thus, we formulate an additional research question in terms of message contents:

**RQ2** How did the prevalence of abusive content evolve in the last years on Telegram?

Moving away from the message-level approach and towards an user-based approach for abusive language detection, so far no methodology has been introduced to address this problem for Telegram. As a solution, we propose developing a graph model leveraging topical information for channels in the German hater community on Telegram to find suitable representations, leading to the third research question:

**RQ3** Can a classification model be used to predict whether a Telegram channel is hateful or not?

Lastly, maintaining the channel perspective, we were interested to investigate whether our approach would allow for the derivation of channel clusters and communities, which is another important aspect regarding online hate. For this, we analyzed the topical distribution and the graph embeddings for each channel, resulting in research question four:

**RQ4** Can we leverage the topical distribution and graph embeddings to derive meaningful clusters from channels?

As an additional contribution, we release an abusive language dataset containing 1,149 Telegram messages labeled as *abusive* or *neutral*.<sup>1</sup>

## Related Work

Studies on Telegram are limited, but the number began to grow in the past years. Baumgartner et al. (2020) released an unlabeled dataset containing 317,224,715 Telegram messages from 27,801 channels, which were posted between 2015 and 2019. They used a snowball sampling strategy to discover channels and collect messages, starting with approximately 250 seed channels (mainly right-wing channels or channels about cryptocurrency). Rogers (2020) conducted an empirical study on actors who were deplatformed on traditional social media and migrated to Telegram. As part of their study, they used a classification model based on hatebase.org to detect messages with hateful language (Rogers 2020). Previous studies on the platform Twitter have shown that identifying networks and user context for social media have significant beneficial impact on classification tasks, such as hate speech detection (Mosca, Wich, and Groh 2021; Wich et al. 2021b) and motivate further in-depth studies on these communities on other platforms. Urman and Katz (2020) conducted an in-depth network analysis of a far-right community on Telegram. They used a snowball sampling strategy to uncover this community, starting with a German-speaking far-right actor. Fielitz and Schwarz (2020) analyzed German hate actors across various social media platforms and investigated the impact of deplatforming activities on these actors. According to them, "Telegram has become the most important online platform for hate actors in

Germany" (Fielitz and Schwarz 2020, p. 5). With a focus on COVID-19, Hohlfeld et al. (2021) and Holzer (2021) investigated public German-speaking channels on Telegram. The only labeled abusive language dataset with Telegram messages that we found is provided by Solopova, Scheffler, and Popa-Wyatt (2021). They released a dataset containing 26,431 messages in English from a channel supporting Donald Trump. To the best of our knowledge, no study has developed an abusive language classification model for German Telegram messages or channels.

Because there is no annotated German Telegram dataset available, we decided to train our classification model on existing German abusive language datasets. In total, we found eight of such datasets (Ross et al. 2016; Bretschneider and Peters 2017; Wiegand, Siegel, and Ruppenhofer 2018; Struß et al. 2019; Mandl et al. 2019, 2020; Wich et al. 2021a; Wich, Räther, and Groh 2021). We decided to use five of them—which constitute the most recent ones, excluding Wich et al. (2021a). These five datasets have comparable label schemata, and a large portion of the data is from the same period as our collected Telegram data. Wich et al. (2021a) was excluded because their data were only pseudo-labeled. More details on the selected datasets can be found in the following section.

As part of our methodology we worked with the Perspective API<sup>2</sup> to classify subsets of messages from Telegram for our semi-supervised baseline comparison. Recent studies that also dealt with the Perspective API have shown systemic bias in their classification framework, which could lead to minority groups being overly flagged by such hate speech systems (Sap et al. 2019, 2021). Sen et al. (2021) similarly performed a study on the Perspective API to discuss potential scientific pitfalls with the usage of automated classification for the social sciences. In our case this problem is dampened firstly by the clear focus on German language text, in which minority German speaker's vernacular is not as pronounced or flagged as offensive speech, but moreover secondly by our sampling strategy, which aims to capture right-wing hate groups and their networks. Through this we are interested in determining the potential toxicity of a very specific subgroup, which in the past was deplatformed for reasons of toxicity and hate speech already. Still we are aware of the limitations of a semi-supervised approach and further studies of the matter should verify results by including domain experts, such as anti-hate speech groups and activists.

## Methodology

In the first part, we describe how we collected data from Telegram. After that, we provide details on how we developed the abusive classification model for Telegram messages based on datasets from other platforms. In the third part, we describe how we developed a classification model to predict whether a channel belong to the hate category based on the results from the message classifier and the social graph.

<sup>1</sup>Code and data available on GitHub:  
<https://github.com/mawic/telegram-abusive-language-classification>

<sup>2</sup><https://www.perspectiveapi.com/>

## Collecting Data

We used a snowball sampling strategy to collect data from Telegram. We only collected messages from public channels that were accessible via the website t.me. A channel is comparable to a news feed: the channel operator can broadcast messages to subscribers of the channel, but subscribers cannot directly post messages on the channel. Groups and private chats were excluded from the data collection process. As seeds for the snowball sampling strategy, we used a list of German hate actors proposed by Fielitz and Schwarz (2020). At the time of data collection, 51 channels from Fielitz and Schwarz (2020)’s list were still accessible. The list comprises, among others, far-right actors, supporters of Qanon, and alternative media.

In the first round of snowball sampling, we collected messages from all seed channels. In the next round, we collected all channels that were mentioned in messages collected from the first round or whose messages were forwarded by the channels of the first round. We repeated this procedure in the third round, but we excluded some of the newly discovered channels due to the large number of channels. We defined a threshold: a channel must be mentioned or forwarded by at least five channels to collect its messages. From all channels, we collected messages that were posted between 01/01/2019 and 03/15/2021.

After data collection, we conducted language detection on the messages because the crawling process also collects other language channels such as Russian and English and we wanted to keep the focus on German. We used multilingual word vectors from *fastText* to classify languages (Grave et al. 2018). The language detection here is based on the message text and a link preview if it exists. In a second step, the language labels of messages are aggregated on a channel level. The language of a channel is German if it is the most or second most common language in the channel. The reason for the latter is that some German channels primarily share content from foreign-language sources. In the following sections, all results refer to the German-speaking channels of the dataset.

## Building Classification Models for Telegram Messages

**Models** To classify Telegram messages, we trained several binary classification models on different German datasets. The goal is to combine multiple classifiers to improve classification performance because each dataset covers different aspects and topics of abusive languages. The reason for focusing on binary classification was that it makes combining classifiers easier.

All classification models are based on pretrained BERT base models (Devlin et al. 2019). We used `deepset/gbert-base` (Chan, Schweter, and Möller 2020) and `dbmdz/bert-base-german-cased`<sup>3</sup> depending on the model’s performance on the individual dataset. Our hyperparameters for training comprise a maximum number of eight epochs, a learning rate of  $5 \times 10^{-5}$ , and a batch size of eight. In addition, we implemented an

<sup>3</sup><https://huggingface.co/dbmdz/bert-base-german-cased>

early stopping callback that stops the training after four consecutive epochs without any improvement. We selected the model with the highest macro F1 score on the validation set.

Before training the models, texts are preprocessed. The preprocessing steps comprise, among others, masking URLs and user names and replacing emojis.

**Data** We used the following German abusive language datasets collected from different platforms (mainly Twitter) to train our models:

- *GermEval 2018*: Wiegand, Siegel, and Ruppenhofer (2018) released an offensive language dataset as part of the shared task GermEval Task 2018. It contains 8,541 tweets with a binary label (*offense, other*) and a fine-grained label (*profanity, insult, abuse, other*). We used the train/test split proposed by the authors and used a 90/10 split for the training/validation set.
- *GermEval 2019*: Struß et al. (2019) published an offensive language dataset that is part of the GermEval Task 2019. It comprises 7,025 tweets that are labeled with the same labeling schema, as the previous dataset, but a further dimension was added (*implicit, explicit*). The data were split in the same way as GermEval 2018.
- *HASOC 2019*: Mandl et al. (2019) released a multilingual hate speech and offensive language dataset, called "Hate Speech and Offensive Content Identification in Indo-European Languages" (Mandl et al. 2019, p. 1), as part of a shared task. It comprises posts from Facebook and Twitter in German, English, and Hindi. The German part comprises 4,669 records with a binary label (*non hate-offensive, hate and offensive*) and a fine-grained label (*hate, offensive, profanity*). We used the train/test split proposed by the authors and used a 90/10 split for the training/validation set.
- *HASOC 2020*: Mandl et al. (2020) published another dataset, which is comparable to the previous one. It consists of posts from YouTube and Twitter in German, English, and Hindi. The German part has a size of 3,425 records using the same labeling schema as the previous dataset. We used the proposed train/validation/test-split of 70%/15%/15%.
- *COVID-19*: Wich, Räther, and Groh (2021) released an abusive language dataset containing 4,960 German tweets that primarily focus on COVID-19. These tweets have a binary label (*neutral, abusive*). We used a train/validation/test split of 70%/15%/15%.

We trained individual classification models for all datasets, except for HASOC 2019 because we could not train a model that provides an acceptable classification performance. Furthermore, we combined the GermEval and HASOC datasets and trained two additional classifiers on the two combined datasets. Combining these datasets was possible because the respective datasets use the same labeling schema.

**Classifying Telegram Messages** Because a Telegram message can have up to 40,986 characters, the tokenized

message may exceed the maximum sequence length of the BERT model, which is 512. To tackle this problem, we split all messages that had more than 412 words into parts with a maximum length of 412 words. When splitting a message, we made sure not to split sentences. For this purpose, we used the sentence detection method of the library spaCy (Honnibal et al. 2020). There were two reasons for setting the threshold to 412 words. First, using words instead of tokens was easier during preprocessing. Second, a word can be tokenized into multiple tokens. Therefore, we set the threshold to 412 instead of 512. Every part of the split message was individually classified. The final label of the complete message results from the highest probability for the abusive class. The reason for this approach was because an abusive text can contain nonabusive parts but not the other way around. In addition to the six classification models, we used Google’s Perspective API to classify the same messages. The API returns a toxicity score between 0 and 1, representing the likelihood that a message should be considered as toxic. Additionally the API offers several models for other factors such as identity attack, insult, profanity, threat etc. In our study we chose general toxicity as the most universal label. While this includes examples like profanity, which are not strictly hate speech related, we chose the broader perspective to represent the extent of flagged content in the network. We used these general toxicity classification results as a semi-supervised baseline to benchmark our models.

**Evaluating Classification Models** To evaluate the classification performance of our trained models on Telegram messages, five annotators manually annotated 1,150 of the classified Telegram messages. More information about the annotators follows below. The 1,150 messages originated from two different sampling strategies. The first strategy uses the classification results of the six trained models and the Perspective API. For each classifier, we sampled 50 messages classified as abusive and 50 classified as neutral, resulting in a total of 700. The second strategy used a topic model trained on Telegram messages (more details on the topic model can be found in the subsection Topic Model). We randomly sampled 30 messages from the 15 most prominent topics. Finally, we ensured that the annotation candidates do not contain any duplicates. As a result, we assured that the dataset has a certain degree of abusive content and that it represents the most relevant topics.

We use the labeling schema of the *COVID-19* dataset proposed by Räther (2021) and Wich, Räther, and Groh (2021) because it is compatible with the binary schema of the *HASOC* and *GermEval* datasets:

- **ABUSIVE:** The tweet comprised "any form of insult, harassment, hate, degradation, identity attack, or the threat of violence targeting an individual or a group." (Räther 2021, p. 36)
- **NEUTRAL:** The tweet did "not fall into the *ABUSIVE* class." (Räther 2021, p. 36)

Data were annotated by four nonexperts and one expert, who are males and in their twenties or early thirties. The annotation process consisted of three phases. In phase 1, the

expert presented and explained the annotation guidelines to the four nonexperts. Subsequently, all five annotators annotated the same 50 messages. In 18 cases, the annotators did not agree on the final label. These cases were discussed in a meeting to align the five annotators. In phase 2, the annotators annotated the remainder of the 1,150 messages. Each message was annotated by two different annotators. The annotators were allowed to skip a message if they could not decide on a label. In phase 3, messages without a consensus were annotated by three additional annotators so that a majority vote was possible. We used Krippendorff’s alpha (Krippendorff 2004) to measure inter-rater reliability. To assist in annotations, we used the text annotation tool of Kili Technology (Kili Technology 2021).

**Combining Classification Models** Because the datasets and consequently the classification models cover different aspects of abusive languages, we combined the six classifiers to improve classification performance (Perspective API was not part of the combination). The labels produced by this combination were used for subsequent experiments.

**Analyzing Evolution of Abusive Content** We performed two analyses to evaluate the evolution of abusive content in the German hater community on Telegram to answer RQ2. First, we compared the number of abusive messages with all messages from the collected German channels between 01/01/2019 and 02/28/2021 on a monthly level. We excluded the messages posted in March 2021 because we did not have data for the entire month. Then, we examined the relative share (prevalence) of abusive content in the messages from all German channels for the same period and granularity. In addition, we reported the prevalence of abusive content from the seed channels and the 1st-degree network of the seed channels.

### **Building a Classification Model for Hatred of Channels**

**Channel Labels** We chose to frame the task as a classification problem deciding on a binary choice of *hater* and *neutral* channels. This formulation was preferable over a formulation as a regression problem, which predicts the relative *hatred* of channels, due to the fact that even channels with the highest amount of hate content, still contain a vast amount of non-hate messages. The average portion of hate messages in the channels of the selected network is 2.7% with a standard deviation of 0.045. Similarly, we were more interested in mapping out the overall extent of the network sharing similar content, than to focus on hotspots based entirely on the intensity of the hate, as opposed to their centrality within the network. To set up the task we had to determine a label for each channel based on whether or not the channel contained any abusive messages. We at first defined a *hater* as a channel that posted or forwarded at least one abusive message. This minimum threshold is chosen based on the fact that we want to generate a comprehensive overview of the potential extent of the spread of hate content on the platform. While it is possible to set the

bar for the hate label higher, we were primarily interested in all channels spreading this type of information and not just in the most prolific spreaders. At the same time, setting the threshold to one proved problematic due to the possibility of misclassification, meaning that false positives would cause neutral channels to be classified as haters. Instead, for each message, we calculated a threshold based on the conditional probability that a message is neutral under the condition of it being labeled as abusive. This conditional probability is retrieved from a confusion matrix (Figure 1h). As a result, we had to adjust the weighting of the confusion matrix’s rows. Because we intentionally oversampled the abusive class in the evaluation set, the ratio of abusive texts was no longer representative of the entire dataset. We assume that the relative share of abusive content is 3.1% for 2020, based on the results from the analysis of the abusive content’s evolution. The resulting conditional probability is 82.9%. Assuming an error rate of smaller than 5.0% , we need at least 17 messages that are classified as *abusive* to be certain that the abuse posted is likely to be genuine. Second, we created a directed graph representing the network of channels. Each channel is a node; a directed edge from nodes A to B exists if A either mentions B or forwards a message from B.

**Topic Model** We assigned a topic distribution vector as a feature to each node of the graph, representing the topical distribution within the messages of the channel. The topical distribution was calculated on the basis of the topic model generated with Top2Vec (Angelov 2020). We relied on the hyperparameter selection of the author, used the `distiluse-base-multilingual-cased`<sup>4</sup> pre-trained sentence transformer as embedding model, and sampled 250,000 messages (500 messages from the 500 channels containing the largest amount of messages in our dataset) as training samples. From the 100 most relevant topics, we manually chose nine topics to serve as proxies for hateful content. These topics were predominant in a larger number of channels, while simultaneously being indicative of hatefulness, predominantly by being focused on a specific kind of discriminative or otherwise *abusive* language. They are listed in Table 1: the topic name in the first column was derived on the basis of the most descriptive terms of the respective topic vectors from which we provide the first three terms in the second column (in German) and a translation of the terms in the third column. Because we are working with many channels that can be associated with German hater communities, we relied only on these topics to cluster different topical emphases with respect to potentially harmful content. We aggregated the counts of all documents in our dataset with cosine similarity to any of the selected topics greater than 0.5 and normalized these counts to create a topic distribution for each node.

**Graph Model** We used GraphSAGE to generate embeddings for the graph (Hamilton, Ying, and Leskovec 2017). The graph was the one described in the paragraph *Channel*

<sup>4</sup><https://huggingface.co/distilbert-base-multilingual-cased>

*Labels* and combined with the topic distribution vectors as node attributes from the previous paragraph. We used the Directed GraphSAGE method from the StellarGraph library (CSIRO’s Data61 2018). As we were learning unsupervised embeddings, i.e., we did not provide the learning model with labels of the channels, we used the *Corrupted Generator* of StellarGraph for sampling additional training data. During training, the model learned to differentiate between true graph instances and corrupted ones. The model was trained for 500 epochs with two layers of size 32 each, an Adam optimizer, and an early stop after 20 epochs of patience.

**Channel Classification** We developed a neural network (NN) classification model using the graph embeddings to predict the classes. The model consists of two densely connected NN layers. The input for the first layer is a 32-dimensional graph embedding. The second layer (output) has two units due to the binary task. The first layer uses a rectified linear unit activation function, whereas soft-max was applied to the output layer. To train the model, cross-entropy was used as a loss function with accuracy as the metric using an Adam optimizer. We trained the model for a maximum of 150 epochs with a batch size of eight with an early stopping strategy that had the patience of 100 epochs and a minimum delta of 0.05 for accuracy on the validation set. The dataset was split into training/validation/test sets (70%/15%/15%).

The dataset for RQ3 only used messages from 2020, as the social network on Telegram is rapidly evolving and changing, with channels and users not staying constant over longer periods of time. That means that by including older edges the overall network structure would generally be less meaningful and introduce noise into the analysis. Another aspect of this decision is that the emergence of COVID-19 strongly influenced and accelerated the evolution of the network, which did not exist pre-COVID-19 pandemic.

## Results

### Collecting Data

In total, we collected 13,822,605 messages from 4,962 channels that were posted between 01/01/2019 and 03/15/2021. 28.4% of all messages (3,931,136) are forwarded messages, showing the popularity and relevance of this feature for Telegram. In addition to the 4,962 channels, we collected the metadata of 43,142 additional channels that were either the source of forwarded messages or were mentioned in a message.

39.2% of all collected messages (5,421,845) are in German, which is the most frequent language, followed by English and Russian. 2,748 of the 4,962 crawled channels (55.4%) are classified as German-speaking according to our approach and are therefore included in the full analysis.

### Building Classification Models for Telegram Messages

**Models** Table 2 presents the classification metrics of the six trained classification models. It comprises the precision, recall, and F1 score of the abusive class as well as the macro F1 score and the used model that performed best on the

Topic	Descriptive terms	Translation
Vaccinations	impfen, geimpft, durchgeimpft	vaccinate, vaccinated, fully vaccinated
Police	Polizeigewalt, Bundespolizei, Polizeiführung	police violence, federal police, police leadership
COVID-19	Coronakrise, Corona, Coronaleugner	corona crisis, corona, corona denier
Migration	Migrantengewalt, Migranten, Refugees	migrant violence, migrants, refugees
Extremism	rechtsextremer, rechtsextremen, rechtsextreme	far-right
Racism	Rassismus, rassistischer, rassistisch	racism, racist
Islamophobia	Moslemterror, Islamisten, Islamisierung	Muslim terror, Islamists, Islamization
Violence	sterben, Massenmörder, Massenmord	die, mass murder
Antisemitism	Antisemitismus, Antisemiten, antisemitische	antisemitism, antisemites, antisemitic

Table 1: Topics selected for topic distribution along with three descriptive terms of the topic model.

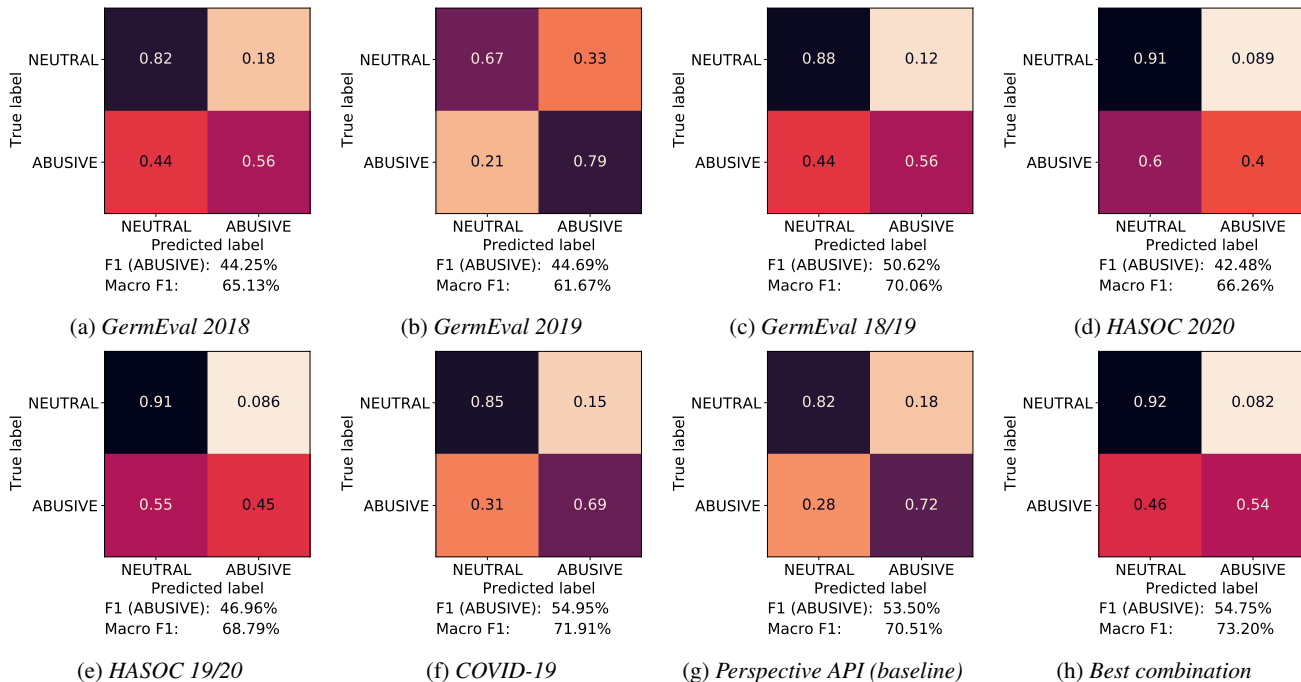


Figure 1: Classification performance of the various models on annotated Telegram evaluation set.

Dataset/Model	Prec	Rec	F1	Macro F1	Basis
GermEval 18	71.1	61.0	65.7	75.0	dbmdz
GermEval 19	72.2	85.1	78.1	77.1	dbmdz
GermEval 18/19	87.6	77.6	82.3	83.8	dbmdz
HASOC 20	69.0	73.7	71.3	80.6	deepset
HASOC 19/20	71.0	69.9	70.4	80.3	dbmdz
COVID-19	73.9	69.9	71.8	82.3	deepset

Table 2: Classification performance of the classifiers

dataset. The last column contains the name of the pretrained model that was used as basis for fine-tuning.

**Evaluating Classification Models** To test the trained classification models, we annotated 1,150 Telegram messages. One message was removed during the annotation process because it did not contain any text, resulting in 1,149 annotated messages. 968 (84.2%) were labeled as *neutral* and 126 (15.8%) as *abusive*. The Krippendorff’s alpha was 73.87%,

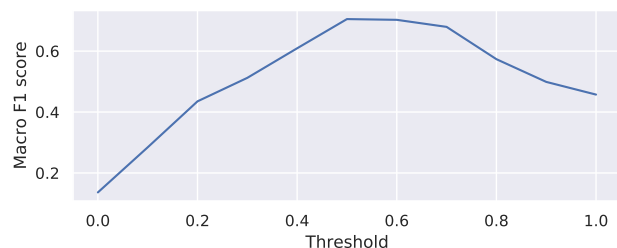
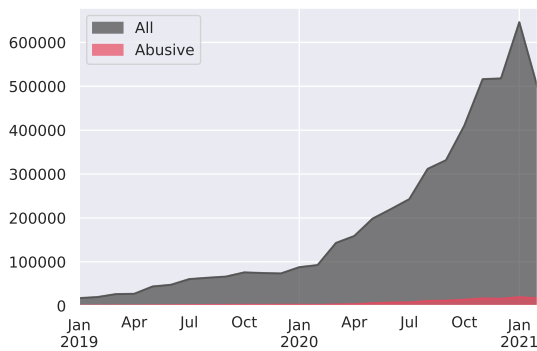


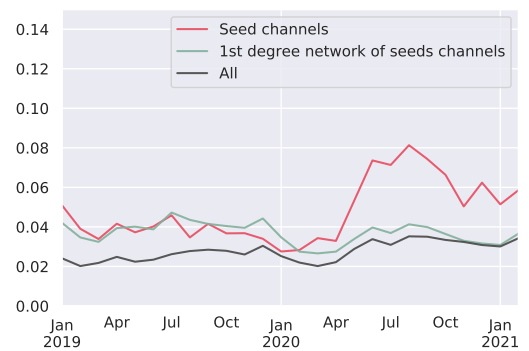
Figure 2: Macro F1 score dependent on various threshold for Perspective API on test set.

which is a good inter-rater reliability score in the context of hate speech and abusive language (Kurrek, Saleem, and Ruths 2020).

Figure 1 visualizes the classification performance of the various classifiers on the evaluation set. It presents the con-



(a) Absolute number of all and abusive messages from German channels.



(b) Relative share of abusive messages for German channels.

Figure 3: Evolution of abusive messages in absolute and relative terms.

fusion matrix, the F1 score of the abusive class, and the macro F1 score of the six trained classification models (a–f), the Perspective API (g), and the best combination of the six classifiers (h). Let us first compare the six classification models that we trained on the different datasets. The best-performing model is COVID-19; it outperformed the other models in terms of F1 score (54.95%) and macro F1 score (71.91%). In comparison to the COVID-19 test set, however, the performance drastically decreased. This should not be surprising because Telegram messages differ from tweets in terms of structure and content.

To benchmark the performance of our classification model, we used Google’s Perspective API to classify messages. The API returns a toxicity score between 0 and 1 which represents the probability of the message being toxic. We translated this value by setting a threshold. If the value is above or equal to the threshold, the label is *abusive*; otherwise, the label is *neutral*. We initially set the threshold for abusive messages to 0.5. Results after validation of other thresholds are collected in Figure 2; the highest macro F1 score on the test set is also achieved by setting a threshold of 0.5. Comparing the performance of the Perspective API with our best-performing model, our model achieves a slightly higher F1 score (54.95% vs. 53.50%) and macro F1 score (71.91% vs 70.51%) in the case of the chosen threshold. The model also achieves comparable results with slightly higher thresholds, with increasing decay in performance for higher toxicity scores, as more messages fall into the false positive category.

Because the datasets cover different aspects of abusive language, we also examined whether a combination of all six classifiers can improve performance. Performance indicates that a majority vote (at least four classifiers vote for *abusive*) of all six models is the best-performing combination in terms of the macro F1 score, as shown in Figure 1h. It outperforms the Perspective API and the classifier trained on the COVID-19 dataset in terms of macro F1 score. To validate the result, we applied the McNemar’s test (Dietterich 1998) to show that the best combination performs significantly differently ( $p < 0.05$ ) from the Perspective API ( $p = 2.69 \times 10^{-5}$ ) and

COVID-19 ( $p = 1.02 \times 10^{-3}$ ). Therefore, the best combination is the majority vote with at least four classifiers voting for *abusive*, which we used for the following two case studies.

**Analyzing the Evolution of Abusive Content** Figure 3a shows how the number of messages in the German Telegram channels has increased between the beginning of 2019 and 2021. We can trace the growth of these channels back to the phenomenon of deplatforming. Deplatforming means that actors are permanently banned on traditional social media platforms (e.g., Facebook, Twitter, and YouTube), resulting in them moving to less moderated or unregulated platforms (e.g., Telegram and Gab) (Rogers 2020; Fielitz and Schwarz 2020; Urman and Katz 2020). Notably, the increase in messages accelerated substantially with the rise of the COVID-19 pandemic (February 2020). The reasons for this are likely similar. Traditional social media platforms (e.g., Twitter and YouTube) blocked accounts of hate actors spreading conspiracy theories regarding COVID-19, causing migration to Telegram and alternative platforms (Fielitz and Schwarz 2020; Holzer 2021). Simultaneously with the growing number of messages every month (black curve), abusive content also increased (red curve).

To answer the question of whether the abusive content has grown only proportionally, we plotted the relative share of abusive content in Figure 3b. The black line represents the relative share for all messages. We observe that the share of abusive content increased from 2.4% to 3.4% during the 26 months. The red line shows the portion of abusive messages in the seed channels. It is not surprising that the share is significantly higher because these channels were labeled as hater channels by Fielitz and Schwarz (2020). The line follows the trend: the abusive content of the selected channels is growing. The green line visualizing the percentage of abusive messages in the channels being in the first-degree network of the seed channels<sup>5</sup> does not mirror the same trend.

<sup>5</sup>A channel is in the first-degree network if a seed channel mentions the channel or forwards a message from this channel and vice versa.



A potential explanation is that the number of channels in the first-degree network has increased over time, causing the alignment of the relative share with the overall average. In total, the prevalence of abusive content for the entire period is 3.1% for all channels, 5.3% for the seed channels, and 3.5% for the 1st degree network of the seed channels.

In summary, we observe the trend that messages classified as *abusive* by our combined model increase in absolute and relative terms in the German hater community on Telegram and are particularly pervasive in the central seed channels.

### Building a Classification Model for the Hatfulness of Channels

In this section, we report the results of our classification model for identifying hateful users, along with additional findings in the process of setting up our model.

**Channel Labels** The dataset for developing a channel classification model contains 2,420 German channels that were active in 2020 and posted 3,232,721 messages. 809 of 2,420 channels (33.4%) are labeled as *hater*, the rest as *neutral*. Each channel is represented by a node in the directed graph. In total, we identified 146,865 edges between channels, which represent messages from one channel which are shared in another or mentioning another channel in a message (unidirectional). This leads to a density of 0.0251 and an average in- and out-degree of 60.73.

**Topical Distribution** As the first result, we examined clusters based on the topical distribution of the seed channels. To do this, the similarity between the topical distribution of each pair of users has been computed using the Jensen—Shannon divergence. For the resulting similarity matrix, a hierarchical clustering approach has been used to group similar users into clusters, as described in Figure 4. While we only disclose an anonymized version of our results, we report that the upper left cluster consists only of sources for alternative news and the large cluster in the center mainly contains actors who belong to the far-right network.

**Graph Embeddings** Before using the graph embeddings from the directed GraphSAGE model for the classification model, we investigated the expressiveness of the embeddings for community detection. For this, we applied the dimensional reduction method UMAP to our embeddings to find more dense representations. In the second step, we used DB-SCAN to cluster these reduced embeddings. In Figure 5, we report the results of the community detection, along with a visualization indicating the label of each node (channel). Seed channels are marked with a large square instead of a dot. The clustering algorithm recognizes four distinct communities along with one outlier class. The channels/nodes of the outlier class are dark green and spread over the figure (cf. Figure 5a). The large community in the center does not only contain most of the seed channels in our dataset but also the largest proportion of channels labeled as *hater* (38%). In the other communities, we find a significantly lower proportion

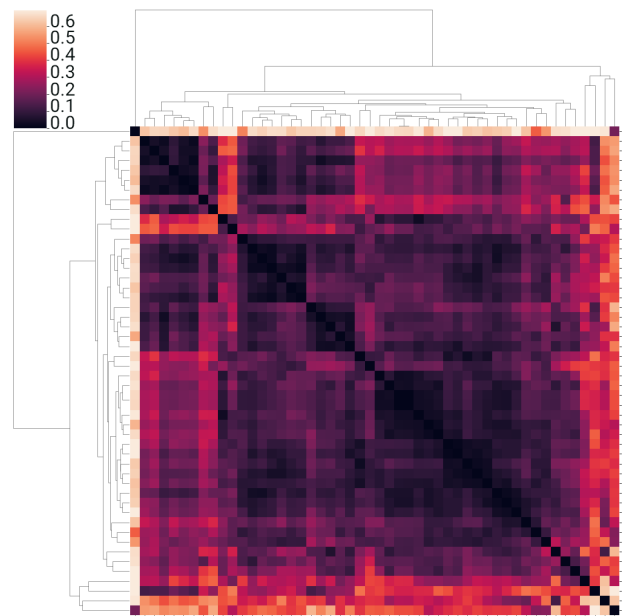


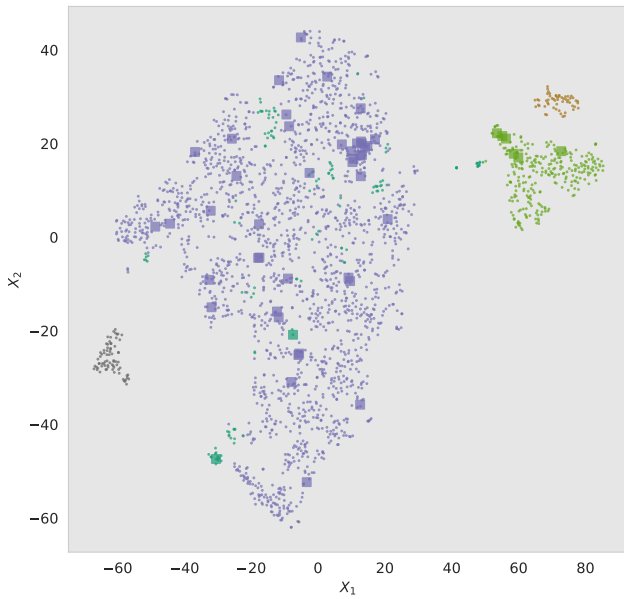
Figure 4: Similarity matrix for the seed channels of the Telegram dataset. A detailed list of channels can be found in the resources on Github, see above.

of hatefully classified users (5%-24%). In the outlier class, 33% are hater. From that, we conclude that hateful users appear more often in communities with other hateful users.

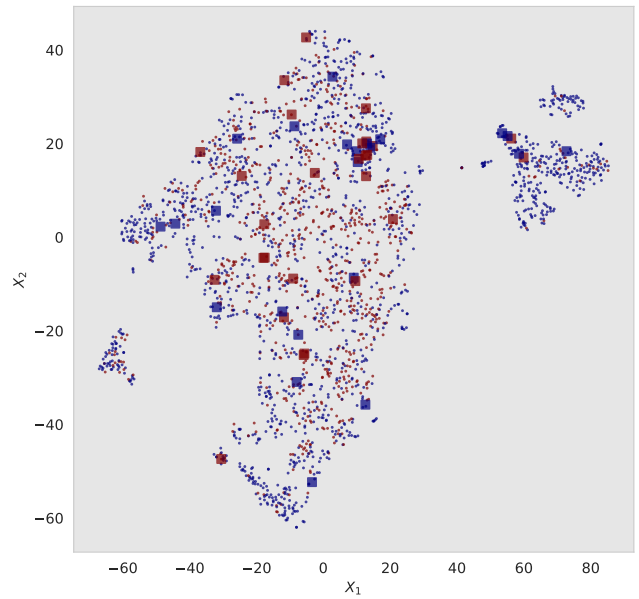
**Channel Classification** The classification model trained to distinguish between *hater* and *neutral* channels achieves a macro F1 score of 69.5% (*neutral*: 74.2%; *hater*: 64.9%). It is important to stress that this performance is reached solely on the unsupervised graph embeddings as input and does not use any additional semantic or text data. Figure 6 visualizes the confusion matrix of the classification model for the test set. We observe that the model performs well in predicting the labels of the German Telegram channels.

### Discussion

In RQ1 we asked whether existing abusive language datasets can be used to train language classification models for the Telegram platform. The short answer to this is yes. However, we have to accept a decline in classification performance. Comparing the macro F1 scores of the classifiers on the original test and evaluation sets, we observe an average decline of approximately 12.5pp. To better assess this value, it is helpful to look into the study on the generalizability of abusive language datasets from Swamy, Jamatia, and Gambäck (2019). They trained models on different abusive language datasets and evaluated them on each other. The average performance decline is 18.1pp if a classifier is evaluated on another test set. Considering this aspect, we can claim that our models perform decently, especially the combination of all six classification models with a threshold of four. This claim



(a) Graph embeddings with community labels



(b) Graph embeddings with hate class labels (red=haters)

Figure 5: Comparison of graph embeddings with community and hate class labels.

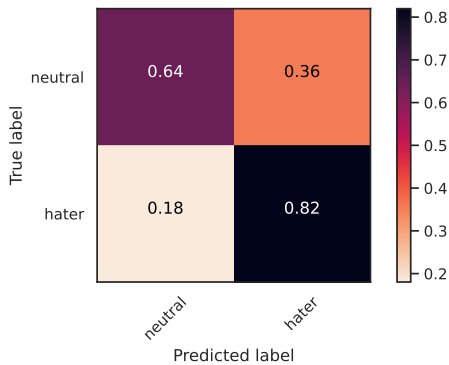


Figure 6: Confusion matrix of model to classify channels.

is supported by the fact that the combined models outperform the Perspective API in terms of F1 score. We integrated this external model provided by Google as a benchmark because it is developed to handle different types of texts (e.g., comments, posts, and emails), and it is in production (Jigsaw 2021). Further, utilizing Twitter abusive language data has proven as particularly helpful in this case, as it offers the largest amount of labeled datasets in German currently available. The availability of the data is not easily compensated by a smaller labeled dataset entirely focused on Telegram, or other social media platforms, such as Reddit. In the end, while it is to be expected that platform specific data would be beneficial for better performance on the task, the core idea of the research was also about tracing the effects of deplatforming and the shift from one social media platform,

such as Twitter, to another. It is to be expected that despite the changing particularities on Telegram, deplatformed actors would still choose to communicate in a similar manner as on the previous platform and talk about similar topics. Further the experiments are also fruitful in case additional platforms, such as Telegram would in the future choose to deplatform certain actors, in which case data would need to be collected from the ground up again. Consequently, we can state that our approach is successful, but it still provides room for improvement.

In RQ2 we wanted to observe the changes in abusive content on Telegram over the deplatforming period on other social media sites. We observe an increasing prevalence of abusive messages in the collected Telegram subnetwork, especially in the group of the seed channels. Notably, the rise of COVID-19 coincided with a significant increase of abusive messages. One may argue that the absolute share of abusive content is unreliable because our combined classification model is imperfect. However, the observed change in the relative share of abusive messages provides a reliable indication of an increasing amount of overall abusive content since it was classified using the same classification model. We trace this trend back to the deplatforming activities of large social media platforms and Telegram’s lack of content moderation. However we also have to point out that the prevalence of abusive content is unrepresentative of the entire German Telegram network. Due to our snowball sampling approach, we have an obvious selection bias because we started with channels that were classified as hate actors by Fielitz and Schwarz (2020). Nevertheless, we assume that the prevalence of abusive content is larger on Telegram than on traditional social media platforms, such as Twitter, Face-

book, and YouTube, that have implemented rigorous reporting and monitoring processes and take an active stance in content moderation. In the case of Telegram, such processes are missing, or entirely in the hands of the respective channel owners.

In RQ3 we asked whether classifying hateful content on a channel level was possible using only aggregate information and the overall network structure. We thus developed a classification model to predict whether a channel is a hate actor. It uses the network structure and the topic distribution of messages in each channel for prediction. Our model achieves a macro F1 score of 69.5%. To the best of our knowledge, we are the first to develop such a classification model for Telegram channels. Therefore, we do not have a baseline to compare our results with. However, Ribeiro et al. (2018) and Li et al. (2021) developed comparable models classifying Twitter accounts as hateful or normal. For the same dataset, Ribeiro et al. (2018) and Li et al. (2021) achieved F1 scores of 67.0% and 79.9%, respectively. Our F1 score of 64.9% is not directly comparable with these results, but it is in a similar order of magnitude, supporting our approach.

In RQ4 we wanted to find out whether we can leverage topical distributions combined with graph embeddings, to derive meaningful clusters from channels. We presented two approaches that allow clustering: The first approach leverages the topical distribution of channels to group actors based on the topical similarity of the content they spread. Applying this to the seed channels for the collection of the dataset indicates promising results for future research attempts in clustering actors on social media based on the content of their postings in a time-saving manner. The second method we propose in this context leverages embeddings learned from the social graph that we generated from the dataset in an unsupervised manner. The advantage of our approach over traditional community detection methods, such as the Louvian method (Blondel et al. 2008), is that it can handle node attributes, meaning additional data can be incorporated in the community detection. This enabled us to combine network data (i.e. relations between the channels) with data about the topics that are discussed in the channels. Our results indicate different communities that vary by the number of hateful users present. Large communities appear to be spanned by seed users which was to be expected based on our data collection approach; however, we also detected smaller communities that do not contain any seed users, indicating that our sampling approach was able to find communities beyond the direct sphere of influence of the initial seed set. For a more precise evaluation of these results, more general information about the German hater community and its relative extend would have been helpful. However, no such studies are currently available.

## Conclusion and Future Work

To the best of our knowledge, we are the first to develop abusive language classification models for German messages on Telegram. Our results look promising. The text model outperforms Google’s Perspective API in terms of

F1 score (macro F1: 73.2%). Similarly, the channel classification model provides good performance in detecting *hater* channels (macro F1: 69.5%). In addition, we have outlined methods for facilitating and scaling abusive language analysis on a message level as well as on the channel level. In the latter case, we fully relied on unsupervised learning methods, which makes these approaches particularly appealing. Furthermore, we publish the first abusive language dataset consisting of German Telegram messages.

There are multiple possible directions for future work in this research field. Firstly, the research community would benefit from larger annotated corpora, which should also include media files shared in those channels (e.g., photos with messages, memes, and videos). Because such media files (e.g., memes) can be used to transport hate (Kiela et al. 2021), they are relevant for the problem of detecting abusive content but were not part of this study.

Regarding the classification model for *hater* channels, integrating additional data (e.g., metadata of the channels) and enhancing the NN architecture could improve classification performance. An explorative network analysis of the sub-network could help identify additional features and give a better overview of the relative extent of hateful communities on Telegram. In addition, a larger overall sample size of Telegram should be collected to mitigate the selection bias introduced by our selection of hateful seed users.

We also encourage researchers from various core disciplines, such as machine learning and social sciences, to synergize in their research efforts and validate the performances achieved by sophisticated learning frameworks applied to large amounts of data with perspectives from social and political science on these phenomena. Due to the unstoppable increase in content produced on social platforms such as Telegram, automatic methods for generating insights will become indispensable. Finally, the hate speech detection community should look into applying approaches such as the ones presented here to other alternative social media platforms as hate actors will congregate there as deplatforming efforts continue.

## References

- Angelov, D. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.
- Baumgartner, J.; Zannettou, S.; Squire, M.; and Blackburn, J. 2020. The Pushshift Telegram Dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 840–847.
- Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10): P10008. doi:10.1088/1742-5468/2008/10/p10008. URL <https://doi.org/10.1088/1742-5468/2008/10/p10008>.
- Bretschneider, U.; and Peters, R. 2017. Detecting offensive statements towards foreigners in social media. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

- Chan, B.; Schweter, S.; and Möller, T. 2020. German's Next Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6788–6796. Barcelona, Spain (Online): International Committee on Computational Linguistics.
- CSIRO's Data61. 2018. StellarGraph Machine Learning Library. <https://github.com/stellargraph/stellargraph>. Accessed: 2022-03-31.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.
- Dietterich, T. G. 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation* 10(7): 1895–1923. ISSN 0899-7667. doi:10.1162/089976698300017197. URL <https://doi.org/10.1162/089976698300017197>.
- Duggan, M. 2017. *Online harassment 2017*. Pew Research Center.
- Echikson, W.; and Knodt, O. 2018. Germany's NetzDG: A key test for combatting online hate. *CEPS Policy Insight*.
- Eckert, S.; Leipertz, S.; and Schmidt, C. 2021. Querdenker: Wie die Corona-Krise zu Radikalisierung führte. *Norddeutscher Rundfunk* URL <https://story.ndr.de/querdenker/>. Visited on 11/20/2021.
- Fielitz, M.; and Schwarz, K. 2020. *Hate not Found?! Deplatforming the Far-Right and its Consequences*. Institut für Demokratie und Zivilgesellschaft: Jena.
- Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; and Mikolov, T. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.
- Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 1025–1035.
- Hohlfeld, R.; Bauerfeind, F.; Braglia, I.; Butt, A.; Dietz, A.-L.; Drexel, D.; Fedlmeier, J.; Fischer, L.; Gandl, V.; Glaser, F.; Habertzettel, E.; Helling, T.; Käsbauer, I.; Kast, M.; Krieger, A.; Lächner, A.; Malkanova, A.; Raab, M.-K.; Rech, A.; and Weymar, P. 2021. Communicating COVID-19 against the backdrop of conspiracy ideologies: How public figures discuss the matter on Facebook and Telegram.
- Holzer, B. 2021. Zwischen Protest und Parodie : Strukturen der »Querdenken«-Kommunikation auf Telegram (und anderswo). In Reichardt, S., ed., *Die Misstrauensgemeinschaft der »Querdenker« : Die Corona-Protteste aus kultur- und sozialwissenschaftlicher Perspektive*, 125–157. Frankfurt: Campus Verlag.
- Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A. 2020. spaCy: Industrial-strength Natural Language Processing in Python. URL <https://doi.org/10.5281/zenodo.1212303>.
- Jigsaw. 2021. Perspective API - Case Studies URL <https://www.perspectiveapi.com/case-studies/>. Visited on 11/20/2021.
- Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Fitzpatrick, C. A.; Bull, P.; Lipstein, G.; Nelli, T.; Zhu, R.; et al. 2021. The Hateful Memes Challenge: Competition Report. In *NeurIPS 2020 Competition and Demonstration Track*, 344–360. PMLR.
- Kili Technology. 2021. Text annotation tool. URL <https://kili-technology.com>. Visited on 11/20/2021.
- Krippendorff, K. 2004. *Content Analysis: An Introduction to Its Methodology*. Content Analysis: An Introduction to Its Methodology. Sage.
- Kurrek, J.; Saleem, H. M.; and Ruths, D. 2020. Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for Online Slur Usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 138–149. Online: Association for Computational Linguistics.
- Li, S.; Zaidi, N. A.; Liu, Q.; and Li, G. 2021. Neighbours and Kinsmen: Hateful Users Detection with Graph Neural Network. In Karlapalem, K.; Cheng, H.; Ramakrishnan, N.; Agrawal, R. K.; Reddy, P. K.; Srivastava, J.; and Chakraborty, T., eds., *Advances in Knowledge Discovery and Data Mining*, 434–446. Cham: Springer International Publishing.
- Mandl, T.; Modha, S.; Kumar M, A.; and Chakravarthi, B. R. 2020. Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In *Forum for Information Retrieval Evaluation, FIRE 2020*, 29–32. New York, NY, USA: Association for Computing Machinery.
- Mandl, T.; Modha, S.; Majumder, P.; Patel, D.; Dave, M.; Mandlia, C.; and Patel, A. 2019. Overview of the HASOC Track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation, FIRE '19*, 14–17. New York, NY, USA: Association for Computing Machinery. ISBN 9781450377508.
- Mosca, E.; Wich, M.; and Groh, G. 2021. Understanding and interpreting the impact of user context in hate speech detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, 91–102.
- Müller, K.; and Schwarz, C. 2021. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association* 19(4): 2131–2167.
- Rafael, Simone; Ritzmann, A. 2019. *Hate Speech and Radicalisation Online - The OCCI Research Report*, chapter Background: the ABC of hate speech, extremism and the NetzDG. ISD Global.

- Räther, S. 2021. *Investigating Techniques for Learning with Limited Labeled Data for Hate Speech Classification*. Master's thesis, Technical University of Munich. Advised and supervised by Maximilian Wich and Georg Groh.
- Ribeiro, M.; Calais, P.; Santos, Y.; Almeida, V.; and Meira Jr, W. 2018. Characterizing and Detecting Hateful Users on Twitter. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)*.
- Rogers, R. 2020. Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication* 35(3): 213–229.
- Ross, B.; Rist, M.; Carbonell, G.; Cabrera, B.; Kurowsky, N.; and Wojatzki, M. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In Beißwenger, M.; Wojatzki, M.; and Zesch, T., eds., *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, volume 17 of *Bochumer Linguistische Arbeitsberichte*, 6–9. Bochum.
- Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, A. N. 2019. The risk of racial bias in hate speech detection. In *ACL*.
- Sap, M.; Swayamdipta, S.; Vianna, L.; Zhou, X.; Choi, Y.; and Smith, N. A. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.
- Sen, I.; Flöck, F.; Weller, K.; Weiß, B.; and Wagner, C. 2021. Applying a total error framework for digital traces to social media research. In *Handbook of Computational Social Science, Volume 2*, 127–139. Routledge.
- Solopova, V.; Scheffler, T.; and Popa-Wyatt, M. 2021. A Telegram corpus for hate speech, offensive language, and online harm. *Journal of Open Humanities Data* 7.
- Struß, J. M.; Siegel, M.; Ruppenhofer, J.; Wiegand, M.; and Klenner, M. 2019. Overview of GermEval Task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, 354–365.
- Swamy, S. D.; Jamatia, A.; and Gambäck, B. 2019. Studying Generalisability across Abusive Language Detection Datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 940–950. Hong Kong, China: Association for Computational Linguistics.
- Urman, A.; and Katz, S. 2020. What they do in the shadows: examining the far-right networks on Telegram. *Information, Communication & Society* 0(0): 1–20.
- Wich, M.; Breiting, M.; Strathern, W.; Naimarevic, M.; Groh, G.; and Pfeffer, J. 2021a. Are your Friends also Haters? Identification of Hater Networks on Social Media: Data Paper. In *Companion Proceedings of the Web Conference 2021 (WWW'21 Companion)*.
- Wich, M.; Mosca, E.; Gorniak, A.; Hingerl, J.; and Groh, G. 2021b. Explainable Abusive Language Classification Leveraging User and Network Data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 481–496. Springer.
- Wich, M.; Räther, S.; and Groh, G. 2021. German Abusive Language Dataset with Focus on COVID-19. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*.
- Wiegand, M.; Siegel, M.; and Ruppenhofer, J. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)*.
- Williams, M. L.; Burnap, P.; Javed, A.; Liu, H.; and Ozalp, S. 2020. Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. *The British Journal of Criminology* 60(1): 93–117.



## BIBLIOGRAPHY

---

- Adadi, Amina and Mohammed Berrada (2018). "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." In: *IEEE Access* 6, pp. 52138–52160. DOI: [10.1109/ACCESS.2018.2870052](https://doi.org/10.1109/ACCESS.2018.2870052).
- Aken, Betty van, Julian Risch, Ralf Krestel, and Alexander Löser (Oct. 2018). "Challenges for Toxic Comment Classification: An In-Depth Error Analysis." In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, pp. 33–42. DOI: [10.18653/v1/W18-5105](https://doi.org/10.18653/v1/W18-5105). URL: <https://aclanthology.org/W18-5105>.
- Akhtar, Sohail, Valerio Basile, and Viviana Patti (Oct. 2020). "Modeling Annotator Perspective and Polarized Opinions to Improve Hate Speech Detection." In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 8. 1, pp. 151–154. URL: <https://ojs.aaai.org/index.php/HCOMP/article/view/7473>.
- Aksenov, Dmitrii, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno-Schneider, and Georg Rehm (Aug. 2021). "Fine-grained Classification of Political Bias in German News: A Data Set and Initial Experiments." In: *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. Online: Association for Computational Linguistics, pp. 121–131. DOI: [10.18653/v1/2021.woah-1.13](https://doi.org/10.18653/v1/2021.woah-1.13). URL: <https://aclanthology.org/2021.woah-1.13>.
- Al Kuwatly, Hala, Maximilian Wich, and Georg Groh (Nov. 2020). "Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics." In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, pp. 184–190. DOI: [10.18653/v1/2020.alw-1.21](https://doi.org/10.18653/v1/2020.alw-1.21). URL: <https://www.aclweb.org/anthology/2020.alw-1.21>.
- Alperin, Juan Pablo, Erik Warren Hanson, Kenneth Shores, and Stefanie Haustein (2017). "Twitter Bot Surveys: A Discrete Choice Experiment to Increase Response Rates." In: *Proceedings of the 8th International Conference on Social Media & Society. #SMSociety17*. Toronto, ON, Canada: Association for Computing Machinery. ISBN: 9781450348478. DOI: [10.1145/3097286.3097313](https://doi.org/10.1145/3097286.3097313).
- Alsafari, Safa, Samira Sadaoui, and Malek Mouhoub (2020). "Hate and offensive speech detection on Arabic social media." In: *Online Social Networks and Media* 19. ISSN: 24686964. DOI: [10.1016/j.osnem.2020.100096](https://doi.org/10.1016/j.osnem.2020.100096).
- Aluru, Sai Saketh, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee (2021). "A Deep Dive into Multilingual Hate Speech Classification." In: *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*. Ed. by Yuxiao Dong, Georgiana Ifrim, Dunja Mladenić, Craig Saunders, and Sofie Van Hoecke. Cham: Springer International

- Publishing, pp. 423–439. ISBN: 978-3-030-67670-4. DOI: [10.1007/978-3-030-67670-4\\_26](https://doi.org/10.1007/978-3-030-67670-4_26).
- Arango, Aymé, Jorge Pérez, and Barbara Poblete (2019). “Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation.” In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR’19. Paris, France: Association for Computing Machinery, pp. 45–54. ISBN: 9781450361729. DOI: [10.1145/3331184.3331262](https://doi.org/10.1145/3331184.3331262).
- Arras, Leila, Franziska Horn, Grégoire Montavon, Klaus Robert Müller, and Wojciech Samek (2017). ““What is relevant in a text document?”: An interpretable machine learning approach.” In: *PLOS ONE* 12.8, pp. 1–24. arXiv: [1612.07843](https://arxiv.org/abs/1612.07843).
- Arthur, Rob (July 10, 2019). “We Analyzed More Than 1 Million Comments on 4chan. Hate Speech There Has Spiked by 40% Since 2015.” In: *Vice*. URL: <https://www.vice.com/en/article/d3nbzy/we-analyzed-more-than-1-million-comments-on-4chan-hate-speech-there-has-spiked-by-40-since-2015> (visited on 10/25/2021).
- Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek (July 2015). “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation.” In: *PLOS ONE* 10.7, pp. 1–46. DOI: [10.1371/journal.pone.0130140](https://doi.org/10.1371/journal.pone.0130140).
- Badjatiya, Pinkesh, Shashank Gupta, Manish Gupta, and Vasudeva Varma (2017). “Deep Learning for Hate Speech Detection in Tweets.” In: *Proceedings of the 26th International Conference on World Wide Web Companion*. WWW ’17 Companion. Perth, Australia: International World Wide Web Conferences Steering Committee, pp. 759–760. ISBN: 9781450349147. DOI: [10.1145/3041021.3054223](https://doi.org/10.1145/3041021.3054223).
- Baldauf, Johannes, Julia Ebner, and Jakob Guhl (2019). *Hate Speech and Radicalisation Online: The OCCI Research Report*. Institute for Strategic Dialogue. URL: <https://www.isdglobal.org/isd-publications/hate-speech-and-radicalisation-online-the-occi-research-report/>.
- Banks, James (Nov. 2010). “Regulating hate speech online.” In: *Intl. Review of Law, Computers & Technology* 24.3, pp. 233–239. URL: <https://www.tandfonline.com/doi/full/10.1080/13600869.2010.522323>.
- Barredo Arrieta, Alejandro et al. (2020). “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.” In: *Information Fusion* 58, pp. 82–115. ISSN: 1566-2535. DOI: [10.1016/j.inffus.2019.12.012](https://doi.org/10.1016/j.inffus.2019.12.012). URL: <https://www.sciencedirect.com/science/article/pii/S1566253519308103>.
- Bashar, Md Abul, Richi Nayak, Khanh Luong, and Thirunavukarasu Balasubramaniam (July 2021). “Progressive domain adaptation for detecting hate speech on social media with small training set and its application to COVID-19 concerned posts.” In: *Social Network Analysis and Mining* 11.1, p. 69. ISSN: 1869-5469. DOI: [10.1007/s13278-021-00780-w](https://doi.org/10.1007/s13278-021-00780-w).



- Basile, Valerio (2020). "It's the End of the Gold Standard as we Know it. On the Impact of Pre-aggregation on the Evaluation of Highly Subjective Tasks." In: *2020 AIXIA Discussion Papers Workshop, AIXIA 2020 DP*. Vol. 2776. CEUR-WS, pp. 31–40.
- Bender, Emily M. and Batya Friedman (2018). "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." In: *Transactions of the Association for Computational Linguistics 6*, pp. 587–604. DOI: [10.1162/tacl\\_a\\_00041](https://doi.org/10.1162/tacl_a_00041). URL: <https://www.aclweb.org/anthology/Q18-1041>.
- Binns, Reuben, Michael Veale, Max Van Kleek, and Nigel Shadbolt (2017). "Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation." In: *Social Informatics*. Ed. by Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri. Cham: Springer International Publishing, pp. 405–415. ISBN: 978-3-319-67256-4. DOI: [10.1007/978-3-319-67256-4\\_32](https://doi.org/10.1007/978-3-319-67256-4_32).
- Bretschneider, Uwe and Ralf Peters (2017). "Detecting offensive statements towards foreigners in social media." In: *Proceedings of the 50th Hawaii International Conference on System Sciences*, pp. 2213–2222. URL: [https://aisel.aisnet.org/hicss-50/dsm/social\\_media\\_culture/3/](https://aisel.aisnet.org/hicss-50/dsm/social_media_culture/3/).
- Chatzakou, Despoina, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali (2017). "Mean Birds: Detecting Aggression and Bullying on Twitter." In: *Proceedings of the 2017 ACM on Web Science Conference*. WebSci '17. Troy, New York, USA: Association for Computing Machinery, pp. 13–22. ISBN: 9781450348966. DOI: [10.1145/3091478.3091487](https://doi.org/10.1145/3091478.3091487).
- Chaudhry, Prateek and Matthew Lease (2020). *You Are What You Tweet: Profiling Users by Past Tweets to Improve Hate Speech Detection*. arXiv: [2012.09090](https://arxiv.org/abs/2012.09090).
- Chetty, Naganna and Sreejith Alathur (2018). "Hate speech review in the context of online social networks." In: *Aggression and Violent Behavior 40*, pp. 108–118. ISSN: 1359-1789. DOI: <https://doi.org/10.1016/j.avb.2018.05.003>. URL: <https://www.sciencedirect.com/science/article/pii/S1359178917301064>.
- Chun, Soon Ae, Richard Holowczak, Kannan Neten Dharan, Ruoyu Wang, Soumaydeep Basu, and James Geller (2019). "Detecting political bias trolls in Twitter data." In: *15th International Conference on Web Information Systems and Technologies, WEBIST 2019*. SciTePress, pp. 334–342. DOI: [10.5220/0008350303340342](https://doi.org/10.5220/0008350303340342).
- Cieslak, Matthew C, Ann M Castelfranco, Vittoria Roncalli, Petra H Lenz, and Daniel K Hartline (2020). "t-Distributed Stochastic Neighbor Embedding (t-SNE): A tool for eco-physiological transcriptomic analysis." In: *Marine genomics 51*, p. 100723. DOI: [10.1016/j.margen.2019.100723](https://doi.org/10.1016/j.margen.2019.100723).
- Commission, European (2020). "White Paper on Artificial Intelligence: A European approach to excellence and trust." In: *Com (2020) 65 Final*. URL: [https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust\\_en](https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en).
- Cotik, Viviana, Natalia Debandi, Franco M Luque, Paula Miguel, Agustín Moro, Juan Manuel Pérez, Pablo Serrati, Joaquin Zajac, and Demián Zayat

- (2020). *A study of Hate Speech in Social Media during the COVID-19 outbreak*. presented at ACL 2020 NLP COVID-19 Workshop. Seattle, WA, USA. URL: <https://openreview.net/pdf?id=01e0ESDhbSW>.
- Dadvar, Maral, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong (2013). "Improving Cyberbullying Detection with User Context." In: *Advances in Information Retrieval*. Ed. by Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, Stefan Ruger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 693–696. ISBN: 978-3-642-36973-5. DOI: [10.1007/978-3-642-36973-5\\_62](https://doi.org/10.1007/978-3-642-36973-5_62).
- Danilevsky, Marina, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen (Dec. 2020). "A Survey of the State of Explainable AI for Natural Language Processing." In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, pp. 447–459. URL: <https://aclanthology.org/2020.aacl-main.46>.
- Das, Mithun, Punyajoy Saha, Ritam Dutt, Pawan Goyal, Animesh Mukherjee, and Binny Mathew (2021). "You Too Brutus! Trapping Hateful Users in Social Media: Challenges, Solutions & Insights." In: *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*. HT '21. Virtual Event, USA: Association for Computing Machinery, pp. 79–89. ISBN: 9781450385510. DOI: [10.1145/3465336.3475106](https://doi.org/10.1145/3465336.3475106).
- Davani, Aida Mostafazadeh, Mark Dıaz, and Vinodkumar Prabhakaran (2021). *Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations*. arXiv: [2110.05719 \[cs.CL\]](https://arxiv.org/abs/2110.05719).
- Davidson, Thomas, Debasmita Bhattacharya, and Ingmar Weber (Aug. 2019). "Racial Bias in Hate Speech and Abusive Language Detection Datasets." In: pp. 25–35. DOI: [10.18653/v1/W19-3504](https://doi.org/10.18653/v1/W19-3504). URL: <https://aclanthology.org/W19-3504>.
- Davidson, Thomas, Dana Warmesley, Michael Macy, and Ingmar Weber (May 2017). "Automated Hate Speech Detection and the Problem of Offensive Language." In: 11.1, pp. 512–515. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/14955>.
- Devakumar, Delan, Geordan Shannon, Sunil S Bhopal, and Ibrahim Abubakar (2020). "Racism and discrimination in COVID-19 responses." In: *The Lancet* 395.10231, p. 1194. DOI: [10.1016/S0140-6736\(20\)30792-3](https://doi.org/10.1016/S0140-6736(20)30792-3).
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423). URL: <https://aclanthology.org/N19-1423>.
- Dixon, Lucas, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman (2018). "Measuring and Mitigating Unintended Bias in Text Classification."

- In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '18. New Orleans, LA, USA: Association for Computing Machinery, pp. 67–73. ISBN: 9781450360128. DOI: [10.1145/3278721.3278729](https://doi.org/10.1145/3278721.3278729).
- Djuric, Nemanja, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati (2015). "Hate Speech Detection with Comment Embeddings." In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15 Companion. Florence, Italy: Association for Computing Machinery, pp. 29–30. ISBN: 9781450334730. DOI: [10.1145/2740908.2742760](https://doi.org/10.1145/2740908.2742760).
- Doshi-Velez, Finale and Been Kim (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. arXiv: [1702.08608](https://arxiv.org/abs/1702.08608) [stat.ML].
- Duggan, Maeve (2017). *Online harassment 2017*. Pew Research Center.
- Ebrahimi, Javid, Anyi Rao, Daniel Lowd, and Dejing Dou (July 2018). "Hot-Flip: White-Box Adversarial Examples for Text Classification." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 31–36. DOI: [10.18653/v1/P18-2006](https://doi.org/10.18653/v1/P18-2006). URL: <https://aclanthology.org/P18-2006>.
- Erjavec, Karmen and Melita Poler (Nov. 2012). "'You Don't Understand, This is a New War!' Analysis of Hate Speech in News Web Sites' Comments." In: *Mass Communication and Society* 15, pp. 899–920. DOI: [10.1080/15205436.2011.619679](https://doi.org/10.1080/15205436.2011.619679).
- Evkoski, Bojan, Andraz Pelicon, Igor Mozetic, Nikola Ljubescic, and Petra Kralj Novak (2021). "Retweet communities reveal the main sources of hate speech." In: arXiv: [2105.14898](https://arxiv.org/abs/2105.14898).
- Fan, Lizhou, Huizi Yu, and Zhanyuan Yin (2020). "Stigmatization in social media: Documenting and analyzing hate speech for COVID-19 on Twitter." In: *Proceedings of the Association for Information Science and Technology*. Association for Information Science and Technology 57.1, e313–e313. ISSN: 2373-9231. DOI: [10.1002/pra2.313](https://doi.org/10.1002/pra2.313). URL: <https://pubmed.ncbi.nlm.nih.gov/33173820>.
- Fehn Unsvåg, Elise and Björn Gambäck (Oct. 2018). "The Effects of User Features on Twitter Hate Speech Detection." In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, pp. 75–85. DOI: [10.18653/v1/W18-5110](https://doi.org/10.18653/v1/W18-5110). URL: <https://aclanthology.org/W18-5110>.
- Fielitz, Maik and Karolin Schwarz (2020). *Hate not Found?! Deplatforming the Far-Right and its Consequences*. Institut für Demokratie und Zivilgesellschaft: Jena. URL: <https://www.idz-jena.de/forschung/hate-not-found-das-deplatforming-der-extremen-rechten/>.
- Fortuna, Paula and Sérgio Nunes (July 2018). "A Survey on Automatic Detection of Hate Speech in Text." In: 51.4. ISSN: 0360-0300. DOI: [10.1145/3232676](https://doi.org/10.1145/3232676).
- Founta, Antigoni, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis (June 2018). "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior." In: *Proceedings of the*

- International AAI Conference on Web and Social Media* 12.1. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/14991>.
- Founta, Antigoni Maria, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis (2019). "A Unified Deep Learning Architecture for Abuse Detection." In: *Proceedings of the 10th ACM Conference on Web Science*. New York, NY, USA: Association for Computing Machinery, pp. 105–114. ISBN: 9781450362023. DOI: [10.1145/3292522.3326028](https://doi.org/10.1145/3292522.3326028).
- Gambäck, Björn and Utpal Kumar Sikdar (Aug. 2017). "Using Convolutional Neural Networks to Classify Hate-Speech." In: *Proceedings of the First Workshop on Abusive Language Online*. Vancouver, BC, Canada: Association for Computational Linguistics, pp. 85–90. DOI: [10.18653/v1/W17-3013](https://doi.org/10.18653/v1/W17-3013). URL: <https://aclanthology.org/W17-3013>.
- Gao, Lei, Alexis Kuppersmith, and Ruihong Huang (Nov. 2017). "Recognizing Explicit and Implicit Hate Speech Using a Weakly Supervised Two-path Bootstrapping Approach." In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 774–782. URL: <https://aclanthology.org/I17-1078>.
- Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford (2020). *Datasheets for Datasets*. arXiv: [1803.09010](https://arxiv.org/abs/1803.09010) [cs.DB].
- Geva, Mor, Yoav Goldberg, and Jonathan Berant (Nov. 2019). "Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 1161–1166. DOI: [10.18653/v1/D19-1107](https://doi.org/10.18653/v1/D19-1107). URL: <https://aclanthology.org/D19-1107>.
- Glaser, April (Nov. 11, 2019). "Where 8channers Went After 8chan." In: *SLATE*. URL: <https://slate.com/technology/2019/11/8chan-8kun-white-supremacists-telegram-discord-facebook.html> (visited on 10/25/2021).
- Google (2021). *Perspective API*. URL: <https://www.perspectiveapi.com/> (visited on 10/25/2021).
- Gordon, Joshua, Marzieh Babaeianjelodar, and Jeanna Matthews (2020). "Studying Political Bias via Word Embeddings." In: *Companion Proceedings of the Web Conference 2020*. WWW '20. Taipei, Taiwan: Association for Computing Machinery, pp. 760–764. ISBN: 9781450370240. DOI: [10.1145/3366424.3383560](https://doi.org/10.1145/3366424.3383560).
- Guest, Ella, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts (Apr. 2021). "An Expert Annotated Dataset for the Detection of Online Misogyny." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 1336–1350.

- DOI: [10.18653/v1/2021.eacl-main.114](https://doi.org/10.18653/v1/2021.eacl-main.114). URL: <https://www.aclweb.org/anthology/2021.eacl-main.114>.
- Guhl, Jakob and Lea Gerster (2020). *Krise und Kontrollverlust: Digitaler Extremismus im Kontext der Corona-Pandemie*. Institute for Strategic Dialogue. URL: <https://www.isdglobal.org/isd-publications/krise-und-kontrollverlust-digitaler-extremismus-im-kontext-der-corona-pandemie/>.
- Hajare, Prasad, Sadia Kamal, Siddharth Krishnan, and Arunkumar Bagavathi (2021). *A Machine Learning Pipeline to Examine Political Bias with Congressional Speeches*. arXiv: [2109.09014](https://arxiv.org/abs/2109.09014) [cs.CY].
- Hamilton, William L., Rex Ying, and Jure Leskovec (2017). "Inductive Representation Learning on Large Graphs." In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., pp. 1025–1035. ISBN: 9781510860964.
- Hanisch, Astrid and Margarete Jäger (2011). "Das Stigma "Gutmensch"." In: *Duisburger Institut für Sprach- und Sozialforschung* 22.
- Hevner, Alan and Samir Chatterjee (2010). "Design Science Research in Information Systems." In: *Design Research in Information Systems*. Vol. 22. Springer, pp. 9–22. DOI: [10.1007/978-1-4419-5653-8\\_2](https://doi.org/10.1007/978-1-4419-5653-8_2).
- Hevner, Alan R (2007). "A Three Cycle View of Design Science Research." In: *Scandinavian Journal of Information Systems* 19.2, p. 4. URL: <https://aisel.aisnet.org/sjis/vol19/iss2/4>.
- Hevner, Alan R., Salvatore T. March, Jinsoo Park, and Sudha Ram (2004). "Design Science in Information Systems Research." In: *MIS Quarterly* 28.1, pp. 75–105. ISSN: 02767783. URL: <http://www.jstor.org/stable/25148625>.
- Hildebrandt, Mireille (2019). "Privacy as protection of the incomputable self: From agnostic to agonistic machine learning." In: *Theoretical Inquiries in Law* 20.1, pp. 83–121. DOI: [10.1515/til-2019-0004](https://doi.org/10.1515/til-2019-0004).
- Hine, Gabriel, Jeremiah Onalapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn (May 2017). "Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web." In: vol. 11. 1, pp. 92–101. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/14893>.
- Holzer, Boris (2021). "Zwischen Protest und Parodie: Strukturen der »Querdenken«-Kommunikation auf Telegram (und anderswo)." In: *Die Misstrauensgemeinschaft der Querdenker: Die Corona-Proteste aus kultur- und sozialwissenschaftlicher Perspektive*. Ed. by Sven Reichardt. Frankfurt/New York: Campus Verlag, pp. 125–157.
- Horta Ribeiro, Manoel, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West (Oct. 2021). "Do Platform Migrations Compromise Content Moderation? Evidence from r/The\_Donald and r/Incels." In: vol. 5. CSCW2. New York, NY, USA: Association for Computing Machinery. DOI: [10.1145/3476057](https://doi.org/10.1145/3476057).

- Hosterman, Alec R, Naomi R Johnson, Ryan Stouffer, and Steven Herring (2018). "Twitter, social support messages, and the #metoo movement." In: *The Journal of Social Media in Society* 7.2, pp. 69–91.
- Howard, Philip N, Aiden Duffy, Deen Freelon, Muzammil M Hussain, Will Mari, and Marwa Maziad (2011). "Opening closed regimes: what was the role of social media during the Arab Spring?" In: *Project on Information Technology and Political Islam Data Memo 2011.1*. DOI: [10.2139/ssrn.2595096](https://doi.org/10.2139/ssrn.2595096).
- Jiang, Shan, Ronald E. Robertson, and Christo Wilson (Apr. 2020). "Reasoning about Political Bias in Content Moderation." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 09, pp. 13669–13672. DOI: [10.1609/aaai.v34i09.7117](https://doi.org/10.1609/aaai.v34i09.7117). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/7117>.
- Johnson, N. F., R. Leahy, N. Johnson Restrepo, N. Velasquez, M. Zheng, P. Manrique, P. Devkota, and S. Wuchty (2019). "Hidden resilience and adaptive dynamics of the global online hate ecology." In: *Nature* 573.7773, pp. 261–265. DOI: [10.1038/s41586-019-1494-7](https://doi.org/10.1038/s41586-019-1494-7).
- Kanclerz, Kamil, Alicja Figas, Marcin Gruza, Tomasz Kajdanowicz, Jan Kocon, Daria Puchalska, and Przemyslaw Kazienko (Aug. 2021). "Controversy and Conformity: from Generalized to Personalized Aggressiveness Detection." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 5915–5926. DOI: [10.18653/v1/2021.acl-long.460](https://doi.org/10.18653/v1/2021.acl-long.460). URL: <https://aclanthology.org/2021.acl-long.460>.
- Kiritchenko, Svetlana, Isar Nejadgholi, and Kathleen C. Fraser (Sept. 2021). "Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective." In: *Journal of Artificial Intelligence Research* 71, pp. 431–478. ISSN: 1076-9757. DOI: [10.1613/jair.1.12590](https://doi.org/10.1613/jair.1.12590).
- Knuttila, Lee (Sept. 2011). "User unknown: 4chan, anonymity and contingency." In: *First Monday* 16.10. DOI: [10.5210/fm.v16i10.3665](https://doi.org/10.5210/fm.v16i10.3665). URL: <https://firstmonday.org/ojs/index.php/fm/article/view/3665>.
- Kocoń, Jan, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko (2021). "Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach." In: *Information Processing & Management* 58.5, p. 102643. ISSN: 0306-4573. DOI: [10.1016/j.ipm.2021.102643](https://doi.org/10.1016/j.ipm.2021.102643). URL: <https://www.sciencedirect.com/science/article/pii/S0306457321001333>.
- Konle, Leonard and Fotis Jannidis (2020). "Domain and Task Adaptive Pre-training for Language Models." In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020), Amsterdam, The Netherlands, November 18-20, 2020*. Ed. by Folgert Karsdorp, Barbara McGillivray, Adina Nerghe, and Melvin Wevers. Vol. 2723. CEUR Workshop Proceedings. CEUR-WS.org, pp. 248–256. URL: <http://ceur-ws.org/Vol-2723/short33.pdf>.
- Kreißel, Philip, Julia Ebner, Alexander Urban, and Jakob Guhl (2018). *Hass auf Knopfdruck. Rechtsextreme Trollfabriken und das Ökosystem koordinierter*

- Hasskampagnen im Netz*. Institute for Strategic Dialogue. URL: [https://www.isdglobal.org/wp-content/uploads/2018/07/ISD\\_Ich\\_Bin\\_Hier\\_2.pdf](https://www.isdglobal.org/wp-content/uploads/2018/07/ISD_Ich_Bin_Hier_2.pdf).
- Kümpel, Anna Sophie and Diana Rieger (2019). *Wandel der Sprach- und Debattenkultur in sozialen Online-Medien: Ein Literaturüberblick zu Ursachen und Wirkungen von inziiviler Kommunikation*. Konrad-Adenauer-Stiftung e. V. DOI: [10.5282/ubm/epub.68880](https://doi.org/10.5282/ubm/epub.68880).
- Kurrek, Jana, Haji Mohammad Saleem, and Derek Ruths (Nov. 2020). "Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for Online Slur Usage." In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, pp. 138–149. DOI: [10.18653/v1/2020.alw-1.17](https://doi.org/10.18653/v1/2020.alw-1.17). URL: <https://www.aclweb.org/anthology/2020.alw-1.17>.
- Larimore, Savannah, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler (June 2021). "Reconsidering Annotator Disagreement about Racist Language: Noise or Signal?" In: *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. Online: Association for Computational Linguistics, pp. 81–90. DOI: [10.18653/v1/2021.socialnlp-1.7](https://doi.org/10.18653/v1/2021.socialnlp-1.7). URL: <https://aclanthology.org/2021.socialnlp-1.7>.
- Le, Quoc and Tomas Mikolov (June 2014). "Distributed Representations of Sentences and Documents." In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Beijing, China: PMLR, pp. 1188–1196. URL: <https://proceedings.mlr.press/v32/le14.html>.
- Lerch, Laurence, Maximilian Wich, Tobias Eder, and Georg Groh (2022). "Mediale Hasssprache und technologische Entscheidbarkeit: Zur ethischen Bedeutung subjektiv-perzeptiver Datenannotationen in der Hate Speech Detection." In: *Medien – Demokratie – Bildung: Normative Vermittlungsprozesse und Diversität in mediatisierten Gesellschaften*. Ed. by Gudrun Marci-Boehncke, Matthias Rath, Malte Delere, and Hanna Höfer. Wiesbaden: Springer Fachmedien Wiesbaden, pp. 295–310. ISBN: 978-3-658-36446-5. DOI: [10.1007/978-3-658-36446-5\\_17](https://doi.org/10.1007/978-3-658-36446-5_17).
- Li, Shu, Nayyar A. Zaidi, Qingyun Liu, and Gang Li (2021). "Neighbours and Kinsmen: Hateful Users Detection with Graph Neural Network." In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Kamal Karlapalem, Hong Cheng, Naren Ramakrishnan, R. K. Agrawal, P. Krishna Reddy, Jaideep Srivastava, and Tanmoy Chakraborty. Cham: Springer International Publishing, pp. 434–446. ISBN: 978-3-030-75762-5. DOI: [10.1007/978-3-030-75762-5\\_35](https://doi.org/10.1007/978-3-030-75762-5_35).
- Linden, Ilse van der, Hinda Haned, and Evangelos Kanoulas (2019). *Global Aggregations of Local Explanations for Black Box models*. arXiv: [1907.03039 \[cs.IR\]](https://arxiv.org/abs/1907.03039).
- Lipton, Zachary C. (June 2018). "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery." In: *Queue* 16.3, pp. 31–57. ISSN: 1542-7730. DOI: [10.1145/3236386.3241340](https://doi.org/10.1145/3236386.3241340).

- Liu, Hui, Qingyu Yin, and William Yang Wang (July 2019). "Towards Explainable NLP: A Generative Explanation Framework for Text Classification." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5570–5581. DOI: [10 . 18653 / v1 / P19 - 1560](https://doi.org/10.18653/v1/P19-1560). URL: <https://aclanthology.org/P19-1560>.
- Lundberg, Scott M. and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions." In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., pp. 4768–4777. ISBN: 9781510860964.
- Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing Data using t-SNE." In: *Journal of Machine Learning Research* 9.86, pp. 2579–2605. URL: <http://jmlr.org/papers/v9/vandermaaten08a.html>.
- MacAvaney, Sean, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder (Aug. 2019). "Hate speech detection: Challenges and solutions." In: *PLOS ONE* 14.8, pp. 1–16. DOI: [10 . 1371 / journal . pone . 0221152](https://doi.org/10.1371/journal.pone.0221152).
- Madukwe, Kosisochukwu, Xiaoying Gao, and Bing Xue (Nov. 2020). "In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets." In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, pp. 150–161. DOI: [10 . 18653 / v1 / 2020 . alw - 1 . 18](https://doi.org/10.18653/v1/2020.alw-1.18). URL: <https://www.aclweb.org/anthology/2020.alw-1.18>.
- Mandl, Thomas, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi (2020). "Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German." In: *Forum for Information Retrieval Evaluation*. FIRE 2020. Hyderabad, India: Association for Computing Machinery, pp. 29–32. ISBN: 9781450389785. DOI: [10 . 1145 / 3441501 . 3441517](https://doi.org/10.1145/3441501.3441517).
- Mandl, Thomas, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel (2019). "Overview of the HASOC Track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages." In: *Proceedings of the 11th Forum for Information Retrieval Evaluation*. FIRE '19. Kolkata, India: Association for Computing Machinery, pp. 14–17. ISBN: 9781450377508. DOI: [10 . 1145 / 3368567 . 3368584](https://doi.org/10.1145/3368567.3368584).
- Mathew, Binny, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee (May 2021). "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.17, pp. 14867–14875. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/17745>.
- McHugh, Mary L. (2012). "Interrater reliability: the kappa statistic." In: *Biochemia medica* 22.3, pp. 276–282. ISSN: 1330-0962. URL: <https://pubmed.ncbi.nlm.nih.gov/23092060>.
- MDZ Digital Library (2021). *dbmdz BERT models*. <https://github.com/dbmdz/berts> (accessed on 1.2.2021).



- Mill, John Stuart (2011). *On Liberty*. Cambridge Library Collection - Philosophy. Cambridge University Press.
- Mishra, Pushkar, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova (Aug. 2018). "Author Profiling for Abuse Detection." In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1088–1098. URL: <https://www.aclweb.org/anthology/C18-1093>.
- (June 2019). "Abusive Language Detection with Graph Convolutional Networks." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 2145–2150. DOI: [10.18653/v1/N19-1221](https://doi.org/10.18653/v1/N19-1221). URL: <https://aclanthology.org/N19-1221>.
- Mishra, Pushkar, Helen Yannakoudakis, and Ekaterina Shutova (2020). *Tackling Online Abuse: A Survey of Automated Abuse Detection Methods*. arXiv: [1908.06024 \[cs.CL\]](https://arxiv.org/abs/1908.06024).
- (Nov. 2021). "Modeling Users and Online Communities for Abuse Detection: A Position on Ethics and Explainability." In: pp. 3374–3385. URL: <https://aclanthology.org/2021.findings-emnlp.287>.
- Molnar, Christoph (2019). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. URL: <https://christophm.github.io/interpretable-ml-book/>.
- Mosca, Edoardo, Maximilian Wich, and Georg Groh (June 2021). "Understanding and Interpreting the Impact of User Context in Hate Speech Detection." In: *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. Online: Association for Computational Linguistics, pp. 91–102. DOI: [10.18653/v1/2021.socialnlp-1.8](https://doi.org/10.18653/v1/2021.socialnlp-1.8). URL: <https://aclanthology.org/2021.socialnlp-1.8>.
- Mubarak, Hamdy, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa (May 2020). "Overview of OSACT4 Arabic Offensive Language Detection Shared Task." In: *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*. Marseille, France: European Language Resource Association, pp. 48–52. ISBN: 979-10-95546-51-1. URL: <https://aclanthology.org/2020.osact-1.7>.
- Müller, Karsten and Carlo Schwarz (Oct. 2020). "Fanning the Flames of Hate: Social Media and Hate Crime." In: *Journal of the European Economic Association* 19.4, pp. 2131–2167. ISSN: 1542-4766. DOI: [10.1093/jeea/jvaa045](https://doi.org/10.1093/jeea/jvaa045).
- Niemann, Marco (2019). "Abusiveness is Non-Binary: Five Shades of Gray in German Online News-Comments." In: *2019 IEEE 21st Conference on Business Informatics (CBI)*. Vol. 01, pp. 11–20. DOI: [10.1109/CBI.2019.00009](https://doi.org/10.1109/CBI.2019.00009).
- Niemann, Marco, Dennis M. Riehle, Jens Brunk, and Jörg Becker (2020). "What Is Abusive Language?" In: *Disinformation in Open Online Media*. Ed. by Christian Grimme, Mike Preuss, Frank W. Takes, and Annie Waldherr. Cham: Springer International Publishing, pp. 59–73. ISBN: 978-3-030-39627-5. DOI: [10.1007/978-3-030-39627-5\\_6](https://doi.org/10.1007/978-3-030-39627-5_6).

- Nobata, Chikashi, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang (2016). "Abusive Language Detection in Online User Content." In: *Proceedings of the 25th International Conference on World Wide Web*. WWW '16. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, pp. 145–153. ISBN: 9781450341431. DOI: [10.1145/2872427.2883062](https://doi.org/10.1145/2872427.2883062).
- Nozza, Debora (Aug. 2021). "Exposing the limits of Zero-shot Cross-lingual Hate Speech Detection." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, pp. 907–914. DOI: [10.18653/v1/2021.acl-short.114](https://doi.org/10.18653/v1/2021.acl-short.114). URL: <https://aclanthology.org/2021.acl-short.114>.
- Obermaier, Magdalena, Michaela Hofbauer, and Carsten Reinemann (2018). "Journalists as targets of hate speech. How German journalists perceive the consequences for themselves and how they cope with it." In: *SCM Studies in Communication and Media* 7.4, pp. 499–524. ISSN: 2192-4007. DOI: [10.5771/2192-4007-2018-4-499](https://doi.org/10.5771/2192-4007-2018-4-499).
- Owen, Tess (Oct. 7, 2019). "How Telegram Became White Nationalists' Go-To Messaging Platform." In: *Vice*. URL: <https://www.vice.com/en/article/59nk3a/how-telegram-became-white-nationalists-go-to-messaging-platform> (visited on 10/25/2021).
- Papakyriakopoulos, Orestis, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco (2020). "Bias in Word Embeddings." In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT\* '20. Barcelona, Spain: Association for Computing Machinery, pp. 446–457. ISBN: 9781450369367. DOI: [10.1145/3351095.3372843](https://doi.org/10.1145/3351095.3372843). URL: <https://doi.org/10.1145/3351095.3372843>.
- Papegnies, Etienne, Vincent Labatut, Richard Dufour, and Georges Linarès (2017). "Graph-Based Features for Automatic Online Abuse Detection." In: *Statistical Language and Speech Processing*. Ed. by Nathalie Camelin, Yannick Estève, and Carlos Martín-Vide. Cham: Springer International Publishing, pp. 70–81. ISBN: 978-3-319-68456-7. DOI: [10.1007/978-3-319-68456-7\\_6](https://doi.org/10.1007/978-3-319-68456-7_6).
- Pitsilis, Georgios K., Heri Ramampiaro, and Helge Langseth (Dec. 2018). "Effective hate-speech detection in Twitter data using recurrent neural networks." In: *Applied Intelligence* 48.12, pp. 4730–4742. ISSN: 1573-7497. DOI: [10.1007/s10489-018-1242-y](https://doi.org/10.1007/s10489-018-1242-y).
- Poerner, Nina, Hinrich Schütze, and Benjamin Roth (July 2018). "Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 340–350. DOI: [10.18653/v1/P18-1032](https://doi.org/10.18653/v1/P18-1032). URL: <https://aclanthology.org/P18-1032>.
- Poletto, Fabio, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti (June 2021). "Resources and benchmark corpora for hate speech detection: a systematic review." In: *Language Resources and Evaluation* 55.2, pp. 477–523. ISSN: 1574-0218. DOI: [10.1007/s10579-020-09502-8](https://doi.org/10.1007/s10579-020-09502-8).

- Prince, Matthew (Aug. 5, 2019). "Terminating Service for 8Chan." In: *The Cloudflare Blog*. URL: <https://blog.cloudflare.com/terminating-service-for-8chan/> (visited on 10/25/2021).
- Qian, Jing, Mai ElSherief, Elizabeth Belding, and William Yang Wang (June 2018). "Leveraging Intra-User and Inter-User Representation Learning for Automated Hate Speech Detection." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 118–123. DOI: [10.18653/v1/N18-2019](https://doi.org/10.18653/v1/N18-2019). URL: <https://aclanthology.org/N18-2019>.
- Raisi, Elaheh and Bert Huang (June 2016). "Cyberbullying Identification Using Participant-Vocabulary Consistency." In: *Proceedings of the 2016 ICML Workshop on #Data4Good: Machine Learning in Social Good Applications*. New York, NY, USA, pp. 46–50.
- (2017). "Cyberbullying Detection with Weakly Supervised Machine Learning." In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ASONAM '17. Sydney, Australia: Association for Computing Machinery, pp. 409–416. ISBN: 9781450349932. DOI: [10.1145/3110025.3110049](https://doi.org/10.1145/3110025.3110049).
- Ranasinghe, Tharindu and Marcos Zampieri (Nov. 2020). "Multilingual Offensive Language Identification with Cross-lingual Embeddings." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 5838–5844. DOI: [10.18653/v1/2020.emnlp-main.470](https://doi.org/10.18653/v1/2020.emnlp-main.470). URL: <https://aclanthology.org/2020.emnlp-main.470>.
- Rangel, Francisco, GLDLP Sarracén, BERTa Chulvi, Elisabetta Fersini, and Paolo Rosso (2021). "Profiling hate speech spreaders on twitter task at PAN 2021." In: *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum*. Bucharest, Romania.
- Rawls, John (1999). *A Theory of Justice*. Cambridge, Massachusetts: Harvard University Press. ISBN: 9780674000773. DOI: [10.2307/j.ctvkjb25m](https://doi.org/10.2307/j.ctvkjb25m).
- Ribeiro, Manoel, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. (June 2018). "Characterizing and Detecting Hateful Users on Twitter." In: *Proceedings of the International AAAI Conference on Web and Social Media 12.1*. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/15057>.
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, pp. 1135–1144. ISBN: 9781450342322. DOI: [10.1145/2939672.2939778](https://doi.org/10.1145/2939672.2939778).
- (Apr. 2018a). "Anchors: High-Precision Model-Agnostic Explanations." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/11491>.

- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (July 2018b). "Semantically Equivalent Adversarial Rules for Debugging NLP models." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 856–865. DOI: [10.18653/v1/P18-1079](https://doi.org/10.18653/v1/P18-1079). URL: <https://aclanthology.org/P18-1079>.
- Richter, Marie, Chine Labbé, Virginia Padovese, and Kendrick McDonald (May 20, 2020). "Twitter: Superspreeder von Corona-Falschinformationen." In: *Newsguard*. URL: <https://www.newsguardtech.com/de/twitter-superspreaders-europe/> (visited on 10/25/2021).
- Rieger, Diana, Anna Sophie Kumpel, Maximilian Wich, Toni Kiening, and Georg Groh (Oct. 2021). "Assessing the Extent and Types of Hate Speech in Fringe Communities: A Case Study of Alt-Right Communities on 8chan, 4chan, and Reddit." In: *Social Media + Society* 7.4. DOI: [10.1177/205630512111052906](https://doi.org/10.1177/205630512111052906).
- Risch, Julian, Robin Ruff, and Ralf Krestel (May 2020). "Offensive Language Detection Explained." In: *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. Marseille, France: European Language Resources Association (ELRA), pp. 137–143. ISBN: 979-10-95546-56-6. URL: <https://aclanthology.org/2020.trac-1.22>.
- Risch, Julian, Philipp Schmidt, and Ralf Krestel (Aug. 2021). "Data Integration for Toxic Comment Classification: Making More Than 40 Datasets Easily Accessible in One Unified Format." In: *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. Online: Association for Computational Linguistics, pp. 157–163. DOI: [10.18653/v1/2021.woah-1.17](https://doi.org/10.18653/v1/2021.woah-1.17). URL: <https://aclanthology.org/2021.woah-1.17>.
- Rogers, Richard (2020). "Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media." In: *European Journal of Communication* 35.3, pp. 213–229. DOI: [10.1177/0267323120922066](https://doi.org/10.1177/0267323120922066).
- Roose, Kevin (June 30, 2020). "Reddit's C.E.O. on Why He Banned 'The\_Donald' Subreddit." In: *The New York Times*. URL: <https://www.nytimes.com/2020/06/30/us/politics/reddit-bans-steve-huffman.html> (visited on 10/25/2021).
- Rosenthal, Sara, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov (Aug. 2021). "SOLID: A Large-Scale Semi-Supervised Dataset for Offensive Language Identification." In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 915–928. DOI: [10.18653/v1/2021.findings-acl.80](https://doi.org/10.18653/v1/2021.findings-acl.80). URL: <https://aclanthology.org/2021.findings-acl.80>.
- Ross, Björn, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki (Sept. 2016). "Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis." In: *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*. Vol. 17, pp. 6–9.
- Röttger, Paul, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert (Aug. 2021). "HateCheck: Functional Tests for

- Hate Speech Detection Models." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 41–58. DOI: [10.18653/v1/2021.acl-long.4](https://doi.org/10.18653/v1/2021.acl-long.4). URL: <https://aclanthology.org/2021.acl-long.4>.
- Rudin, Cynthia and Berk Ustun (2018). "Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice." In: *INFORMS Journal on Applied Analytics* 48.5, pp. 449–466. DOI: [10.1287/inte.2018.0957](https://doi.org/10.1287/inte.2018.0957).
- Russell, Stuart and Peter Norvig (2009). *Artificial Intelligence: A Modern Approach*. 3rd. USA: Prentice Hall Press. ISBN: 0136042597.
- Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf (2020). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv: [1910.01108](https://arxiv.org/abs/1910.01108) [cs.CL].
- Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith (July 2019). "The Risk of Racial Bias in Hate Speech Detection." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 1668–1678. DOI: [10.18653/v1/P19-1163](https://doi.org/10.18653/v1/P19-1163). URL: <https://www.aclweb.org/anthology/P19-1163>.
- sSCAN (2020). *Hate speech trends during the Covid-19 pandemic in a digital and globalised age*. sSCAN project – Platforms, Experts, Tools: Specialised Cyber-Activists Network. URL: <https://scan-project.eu/wp-content/uploads/sSCAN-Analytical-Paper-Hate-speech-trends-during-the-Covid-19-pandemic-in-a-digital-and-globalised-age.pdf> (visited on 10/25/2021).
- Schaetz, Nadja, Laura Leibner, Pablo Porten-Cheé, Martin Emmer, and Christian Strippel (2020). *Politische Partizipation in Deutschland 2019*. Vol. 1. Weizenbaum Report. Berlin: Weizenbaum Institute for the Networked Society - The German Internet Institute, p. 13. DOI: [10.34669/wi.wr/1](https://doi.org/10.34669/wi.wr/1).
- Schmidt, Anna and Michael Wiegand (Apr. 2017). "A Survey on Hate Speech Detection using Natural Language Processing." In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Valencia, Spain: Association for Computational Linguistics, pp. 1–10. DOI: [10.18653/v1/W17-1101](https://doi.org/10.18653/v1/W17-1101). URL: <https://www.aclweb.org/anthology/W17-1101>.
- Shahrezaye, Morteza, Miriam Meckel, and Simon Hegelich (2020). "Estimating the Political Orientation of Twitter Users Using Network Embedding Algorithms." In: *Journal of Applied Business & Economics* 22.14. DOI: [10.33423/jabe.v22i14.3965](https://doi.org/10.33423/jabe.v22i14.3965).
- Shahrezaye, Morteza, Orestis Papakyriakopoulos, Juan Carlos Medina Serano, and Simon Hegelich (2019). "Estimating the Political Orientation of Twitter Users in Homophilic Networks." In: *AAAI Spring Symposium: Interpretable AI for Well-being*. URL: [http://ceur-ws.org/Vol-2448/SSS19\\_Paper\\_Upload\\_216.pdf](http://ceur-ws.org/Vol-2448/SSS19_Paper_Upload_216.pdf).

- Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje (Aug. 2017). "Learning Important Features Through Propagating Activation Differences." In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 3145–3153. URL: <https://proceedings.mlr.press/v70/shrikumar17a.html>.
- Shrikumar, Avanti, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje (2017). "Not Just a Black Box: Learning Important Features Through Propagating Activation Differences." In: arXiv: [1605.01713](https://arxiv.org/abs/1605.01713) [cs.LG].
- Solopova, Veronika, Tatjana Scheffler, and Mihaela Popa-Wyatt (2021). "A Telegram corpus for hate speech, offensive language, and online harm." In: *Journal of Open Humanities Data* 7. DOI: [10.5334/johd.32](https://doi.org/10.5334/johd.32).
- Spertus, Ellen (1997). "Smokey: Automatic Recognition of Hostile Messages." In: *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*. AAAI'97/IAAI'97. Providence, Rhode Island: AAAI Press, pp. 1058–1065. ISBN: 0262510952.
- Stappen, Lukas, Fabian Brunn, and Björn Schuller (2020). *Cross-lingual Zero- and Few-shot Hate Speech Detection Utilising Frozen Transformer Language Models and AXEL*. arXiv: [2004.13850](https://arxiv.org/abs/2004.13850) [cs.CL].
- Struß, Julia Maria, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, Manfred Klenner, et al. (2019). "Overview of GermEval Task 2, 2019 shared task on the identification of offensive language." In: *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pp. 354–365. DOI: [10.5167/uzh-178687](https://doi.org/10.5167/uzh-178687).
- Sun, Shiliang, Honglei Shi, and Yuanbin Wu (2015). "A survey of multi-source domain adaptation." In: *Information Fusion* 24, pp. 84–92. ISSN: 1566-2535. DOI: [10.1016/j.inffus.2014.12.003](https://doi.org/10.1016/j.inffus.2014.12.003). URL: <https://www.sciencedirect.com/science/article/pii/S1566253514001316>.
- Švec, Andrej, Matúš Pikuliak, Marián Šimko, and Mária Bieliková (Oct. 2018). "Improving Moderation of Online Discussions via Interpretable Neural Models." In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, pp. 60–65. DOI: [10.18653/v1/W18-5108](https://doi.org/10.18653/v1/W18-5108). URL: <https://aclanthology.org/W18-5108>.
- Swamy, Steve Durairaj, Anupam Jamatia, and Björn Gambäck (Nov. 2019). "Studying Generalisability across Abusive Language Detection Datasets." In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, pp. 940–950. DOI: [10.18653/v1/K19-1088](https://doi.org/10.18653/v1/K19-1088). URL: <https://aclanthology.org/K19-1088>.
- Telegram (2021). *Telegram FAQ*. URL: <https://telegram.org/faq> (visited on 10/25/2021).
- Timberg, Craig and Drew Harwell (Feb. 5, 2021). "TheDonald's owner speaks out on why he finally pulled plug on hate-filled site." In: *Washington Post*.

- URL: <https://www.washingtonpost.com/technology/2021/02/05/why-the-donald-moderator-left/> (visited on 10/25/2021).
- Tuters, Marc and Sal Hagen (2020). “(((They))) rule: Memetic antagonism and nebulous othering on 4chan.” In: *New Media & Society* 22.12, pp. 2218–2237. DOI: [10.1177/1461444819888746](https://doi.org/10.1177/1461444819888746).
- Ullmann, Stefanie and Marcus Tomalin (Mar. 2020). “Quarantining online hate speech: technical and ethical perspectives.” In: *Ethics and Information Technology* 22.1, pp. 69–80. ISSN: 1572-8439. DOI: [10.1007/s10676-019-09516-z](https://doi.org/10.1007/s10676-019-09516-z).
- Urman, Aleksandra and Stefan Katz (2020). “What they do in the shadows: examining the far-right networks on Telegram.” In: *Information, Communication & Society* 0.0, pp. 1–20. DOI: [10.1080/1369118X.2020.1803946](https://doi.org/10.1080/1369118X.2020.1803946).
- Velásquez, N., R. Leahy, N. Johnson Restrepo, Y. Lupu, R. Sear, N. Gabriel, O. Jha, B. Goldberg, and N. F. Johnson (2020). *Hate multiverse spreads malicious COVID-19 content online beyond individual platform control*. arXiv: [2004.00673 \[physics.soc-ph\]](https://arxiv.org/abs/2004.00673).
- Vidgen, Bertie and Leon Derczynski (Dec. 2021). “Directions in abusive language training data, a systematic review: Garbage in, garbage out.” In: *PLOS ONE* 15, pp. 1–32. DOI: [10.1371/journal.pone.0243300](https://doi.org/10.1371/journal.pone.0243300).
- Vidgen, Bertie, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble (Nov. 2020). “Detecting East Asian Prejudice on Social Media.” In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, pp. 162–172. DOI: [10.18653/v1/2020.alw-1.19](https://doi.org/10.18653/v1/2020.alw-1.19). URL: <https://aclanthology.org/2020.alw-1.19>.
- Vidgen, Bertie, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts (Aug. 2019). “Challenges and frontiers in abusive content detection.” In: *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, pp. 80–93. DOI: [10.18653/v1/W19-3509](https://doi.org/10.18653/v1/W19-3509). URL: <https://www.aclweb.org/anthology/W19-3509>.
- Vidgen, Bertie, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble (June 2021). “Introducing CAD: the Contextual Abuse Dataset.” In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 2289–2303. DOI: [10.18653/v1/2021.naacl-main.182](https://doi.org/10.18653/v1/2021.naacl-main.182). URL: <https://aclanthology.org/2021.naacl-main.182>.
- Vijayaraghavan, Prashanth, Hugo Larochelle, and Deb Roy (2019). “Interpretable multi-modal hate speech detection.” In: *Proceedings of the International Conference on Machine Learning AI for Social Good Workshop*. Long Beach, United States.
- Vogels, Emily A. (2021). *The State of Online Harassment*. Pew Research Center. URL: <https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/> (visited on 11/28/2021).

- Waldron, Jeremy (2012). *The Harm in Hate Speech*. Cambridge, Massachusetts and London, England: Harvard University Press.
- Wang, Cindy (Oct. 2018). "Interpreting Neural Network Hate Speech Classifiers." In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, pp. 86–92. DOI: [10.18653/v1/W18-5111](https://doi.org/10.18653/v1/W18-5111). URL: <https://aclanthology.org/W18-5111>.
- Waseem, Zeerak (Nov. 2016). "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter." In: *Proceedings of the First Workshop on NLP and Computational Social Science*. Austin, Texas: Association for Computational Linguistics, pp. 138–142. DOI: [10.18653/v1/W16-5618](https://doi.org/10.18653/v1/W16-5618). URL: <https://aclanthology.org/W16-5618>.
- Waseem, Zeerak, Thomas Davidson, Dana Warmesley, and Ingmar Weber (Aug. 2017). "Understanding Abuse: A Typology of Abusive Language Detection Subtasks." In: *Proceedings of the First Workshop on Abusive Language Online*. Vancouver, BC, Canada: Association for Computational Linguistics, pp. 78–84. DOI: [10.18653/v1/W17-3012](https://doi.org/10.18653/v1/W17-3012). URL: <https://aclanthology.org/W17-3012>.
- Waseem, Zeerak and Dirk Hovy (June 2016). "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter." In: *Proceedings of the NAACL Student Research Workshop*. San Diego, California: Association for Computational Linguistics, pp. 88–93. DOI: [10.18653/v1/N16-2013](https://doi.org/10.18653/v1/N16-2013). URL: <https://www.aclweb.org/anthology/N16-2013>.
- Wich, Maximilian, Hala Al Kuwatly, and Georg Groh (Nov. 2020). "Investigating Annotator Bias with a Graph-Based Approach." In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, pp. 191–199. DOI: [10.18653/v1/2020.alw-1.22](https://doi.org/10.18653/v1/2020.alw-1.22). URL: <https://www.aclweb.org/anthology/2020.alw-1.22>.
- Wich, Maximilian, Jan Bauer, and Georg Groh (Nov. 2020). "Impact of Politically Biased Data on Hate Speech Classification." In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, pp. 54–64. DOI: [10.18653/v1/2020.alw-1.7](https://doi.org/10.18653/v1/2020.alw-1.7). URL: <https://www.aclweb.org/anthology/2020.alw-1.7>.
- Wich, Maximilian, Melissa Breiterger, Wienke Strathern, Marlena Naimarevic, Georg Groh, and Jürgen Pfeffer (Apr. 2021). "Are Your Friends Also Haters? Identification of Hater Networks on Social Media: Data Paper." In: *Companion Proceedings of the Web Conference 2021*. WWW '21. Ljubljana, Slovenia: Association for Computing Machinery, pp. 481–485. ISBN: 9781450383134. DOI: [10.1145/3442442.3452310](https://doi.org/10.1145/3442442.3452310).
- Wich, Maximilian, Tobias Eder, Hala Al Kuwatly, and Georg Groh (July 2021). "Bias and comparison framework for abusive language datasets." In: *AI and Ethics*. ISSN: 2730-5961. DOI: [10.1007/s43681-021-00081-0](https://doi.org/10.1007/s43681-021-00081-0).
- Wich, Maximilian, Adrian Gorniak, Tobias Eder, Daniel Bartmann, Burak Enes Çakici, and Georg Groh (May 2022). "Introducing an Abusive Language Classification Framework for Telegram to Investigate the German Hater Community." In: *Proceedings of the International AAAI Conference on Web and*



- Social Media* 16.1, pp. 1133–1144. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/19364>.
- Wich, Maximilian, Edoardo Mosca, Adrian Gorniak, Johannes Hingerl, and Georg Groh (Sept. 2021). “Explainable Abusive Language Classification Leveraging User and Network Data.” In: *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*. Ed. by Yuxiao Dong, Nicolas Kourtellis, Barbara Hammer, and Jose A. Lozano. Cham: Springer International Publishing, pp. 481–496. ISBN: 978-3-030-86517-7. DOI: [10.1007/978-3-030-86517-7\\_30](https://doi.org/10.1007/978-3-030-86517-7_30).
- Wich, Maximilian, Svenja Räther, and Georg Groh (Sept. 2021). “German Abusive Language Dataset with Focus on COVID-19.” In: *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*. Düsseldorf, Germany: KONVENS 2021 Organizers, pp. 247–252. ISBN: 978-1-954085-83-1. URL: <https://aclanthology.org/2021.konvens-1.26>.
- Wich, Maximilian, Christian Widmer, Gerhard Hagerer, and Georg Groh (Sept. 2021). “Investigating Annotator Bias in Abusive Language Datasets.” In: *Deep Learning for Natural Language Processing Methods and Applications*. Held Online: INCOMA Ltd., pp. 1515–1525. ISBN: 978-954-452-072-4. DOI: [10.26615/978-954-452-072-4\\_170](https://doi.org/10.26615/978-954-452-072-4_170). URL: <https://aclanthology.org/2021.ranlp-1.170>.
- Wiegand, Michael, Josef Ruppenhofer, and Thomas Kleinbauer (June 2019). “Detection of Abusive Language: the Problem of Biased Datasets.” In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 602–608. DOI: [10.18653/v1/N19-1060](https://doi.org/10.18653/v1/N19-1060). URL: <https://www.aclweb.org/anthology/N19-1060>.
- Wiegand, Michael, Melanie Siegel, and Josef Ruppenhofer (2018). “Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language.” In: *Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)*. Vienna, Austria.
- Williams, Matthew L, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp (2020). “Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime.” In: *The British Journal of Criminology* 60.1, pp. 93–117. DOI: [10.1093/bjc/azz064](https://doi.org/10.1093/bjc/azz064).
- Wong, Julia Carrie (Aug. 5, 2019). “8chan: the far-right website linked to the rise in hate crimes.” In: *The Guardian*. URL: <https://www.theguardian.com/technology/2019/aug/04/mass-shootings-el-paso-texas-dayton-ohio-8chan-far-right-website> (visited on 10/25/2021).
- Wulczyn, Ellery, Nithum Thain, and Lucas Dixon (2017). “Ex Machina: Personal Attacks Seen at Scale.” In: *Proceedings of the 26th International Conference on World Wide Web. WWW '17*. Perth, Australia: International World Wide Web Conferences Steering Committee, pp. 1391–1399. ISBN: 9781450349130. DOI: [10.1145/3038912.3052591](https://doi.org/10.1145/3038912.3052591).

- Yong, Caleb (2011). "Does freedom of speech include hate speech?" In: *Res Publica* 17.4, pp. 385–403. DOI: [10.1007/s11158-011-9158-y](https://doi.org/10.1007/s11158-011-9158-y).
- Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar (June 2019). "SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)." In: pp. 75–86. DOI: [10.18653/v1/S19-2010](https://doi.org/10.18653/v1/S19-2010). URL: <https://www.aclweb.org/anthology/S19-2010>.
- Ziems, Caleb, Bing He, Sandeep Soni, and Srijan Kumar (2021). *Racism is a Virus: Anti-Asian Hate and Counterspeech in Social Media during the COVID-19 Crisis*. arXiv: [2005.12423v1](https://arxiv.org/abs/2005.12423v1) [cs.SI].