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.

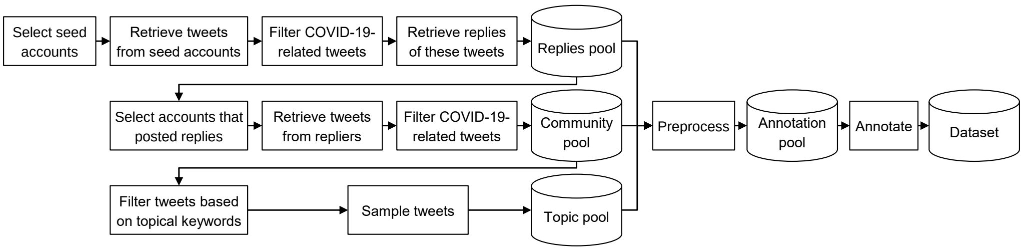
Select seed accounts
Retrieve tweets from seed accounts
Filter COVID-19-related tweets
Retrieve replies of these tweets
Replies pool
Select accounts that posted replies
Retrieve tweets from repliers
Filter COVID-19-related tweets
Community pool
Filter tweets based on topical keywords
Sample tweets
Topic pool
Preprocess
Annotation pool
Annotate
Dataset

Figure 1: Dataset creation process adapted from Räther (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; Struß 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; Struß 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. Consequently, 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[1] 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.

[1] 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 preprocessing 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 5×10^{-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, durchseuchung, einreisesperre, einreiseverbot, epidemi, existenzangst, fallzahl, gesichtsvisier, gesundheitsamt, grundrechte, hygienedemo, hygienemaßnahme, immun, impf, infekt, influenza, inkubationszeit, intensivbett, inzidenz, kontaktbeschränkung, kontaktverbot, lockdown, lockerungen, mundschutz, 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 Räther (2021, p. 84))

Topic (sCAN, 2020)	**Keywords**
"Anti-Asian racsim"	asiat, chines, ccp, wuhan, chinavirus
"Misinformation and geopolitical strategy"	amerika, militär, biowaffe
"Resurgence of old antisemitic stereotypes"	jude, jüdisch, 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

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

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 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`.

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*. Institue 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`.

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 multiverse 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 Breitinger, 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*.