Comparative Survey of German Hate Speech Datasets: Background, Characteristics and Biases

Markus Bertram[1], Johannes Schäfer[1] and Thomas Mandl[1]

[1]*University of Hildesheim, Germany*

Abstract

The large fraction of hate speech and other offensive and objectionable content online poses a vast challenge to societies. Offensive language such as insulting, hurtful, derogatory, or obscene content directed from one person to another and open to others undermines objective discussions. Hate speech detection quality depends on the datasets available for training. Potential bias needs to be identified in order to increase the generalization performance of the trained classifiers. This article gives an overview on nine German hate speech datasets. We apply a framework from the literature to gain insights into potential bias. Using different methods, our analysis shows that there are various topics in the different datasets. The results are shown and compared for LSI, Topic models, Mutual Information and Shapley values.

Keywords

Hate speech, datasets, reliability, PMI, LSI, Shapley values

1. Introduction

Hate speech and its detection has gotten significant attention in recent years [1, 2]. Increasingly, governments are trying to police social media platforms and require the implementation of automatic detection methods of illegal content such as hate speech and disinformation. A notable example of this is the *Digital Services Act* adopted by the EU parliament in July 2022, with the aim of "protection of users' rights online" [3]. This shows that hate speech and its detection raises important ethical questions with regards to free speech, protecting users and groups, as well as promoting social good. Hate speech intends to do harm to individuals and groups which motivates private and government actors to discourage it.

Because of the vast amount of online communication, which makes a manual review of all communication infeasible, there exists a need to automatically filter or detect potential hate speech [4]. Hate speech detection methods try to automatically predict how likely a given form of online communication contains hate speech, and thereby can assist humans in filtering. To properly train these systems, reliable datasets need to be available. So far, there has been little work regarding the analysis of the quality of datasets and the comparison of datasets in natural language processing (NLP) in general.

With this paper we address this issue and present a survey of nine German hate speech datasets. In the following sections we review related work (see Section 2) and outline the

LWDA, Lernen, Wissen, Daten, Analysen. October 9-11, 2023, Marburg, Germany

johannes.schaefer@uni-hildesheim.de (J. Schäfer); mandl@uni-hildesheim.de (T. Mandl)

0000-0002-5226-4206 (J. Schäfer); 0000-0002-8398-9699 (T. Mandl)

comparison framework by Wich et al. [5] in Section 3 as basis for our analysis. In Section 4 we discuss the datasets included in our survey and present the results of the framework analysis in Section 5. Finally, we conclude in Section 6.

2. Research in Hate Speech Detection

The question of the quality of databases has been approached from several angles and there is concern that current datasets do not lead to classifiers with a good level of generalization [6, 7].

A recent study analyzed six different English language hate speech datasets, with different but related labels like *hate speech*, *offensive*, *aggression*, *toxicity* etc. [8]. The authors visualized how similar and compatible classes are within and across the datasets and measured how well each class affects performance of hate speech classifiers. They grouped semantically similar classes and calculated centroids using pre-trained word embedding for each class, which are then used to calculate distances between them.

Several other works explored hate speech datasets with regards to their biases and characteristics, as well as their generalizability. A study by Nejadgholi and Kiritchenko [9] explored two different types of bias in hate speech datasets and their effect on cross-dataset generalization: *topic bias* and *task formulation bias*. The former is a type of selection bias and was identified using keyword search. The authors showed that some topics are more generalizable than others. The latter bias describes the difference in the definitions of classes between the datasets. The effect of this bias was estimated by training classifiers on different tasks. *HATECHECK* for English hate speech datasets is a collection of 29 tests, each designed to offer insight into specific weaknesses of hate speech classifiers [10]. For each test, a manually annotated test set was created. The authors showed that in their setting, models tend to focus on specific terms and not take the context into account. Lastly, Yin and Zubiaga [11] summarized the cross-dataset performance of English hate speech detection models and provided reasons for why models fail to generalize. They argued that models fail to generalize because of different grammar and vocabulary used in hate speech datasets, too few labeled data, sampling bias, representation biases, i.e. the failure of models to take the language of minority groups into account, as well as models failing to detect implicit hate speech.

3. Bias and Comparison Framework

For the analysis of German hate speech collections, we applied the framework introduced by Wich et al. [5] that can be used to show the biases and characteristics of such datasets. This bias framework can visualize the difference of the probability distributions between and within hate speech datasets. It has been applied to English and Arabic hate speech datasets.

3.1. LSI-based Similarity

The first approach implemented is based on *Latent Semantic Indexing* (LSI), presented by Deerwester et al. [12], and is a way to visualize the intra-dataset similarity between classes. LSI is a method to find a transformation $\mathcal{X} \rightarrow \mathcal{Z}$ that embeds documents in a lower-dimensional latent space based on their semantic similarities. Specifically, for each dataset, all unique words are extracted and a document-term matrix is created. Then, Singular-Value Decomposition with k singular values is applied. After that, the datasets are then filtered by classes and each document is transformed into a bag-of-word vector, which is then transformed into the latent space.

Lastly, the average cosine-similarity between the documents of the same and the other classes are computed. The average cosine-similarity is a measure of how similar the classes are with itself and the other classes and are therefore a measure of intra-dataset similarity.

3.2. Word-embedding-based Similarity

This approach focuses on visualizing the intra- and inter-dataset similarity using pre-trained word embeddings. Different to Wich et al. [5], we use word embeddings produced by the pre-trained gBERT model. Because gBERT does not embed whole sentences directly, the embedding of the first [CLS] token as sentence representation of the document is used [13]. The word embedding vectors of each dataset are averaged into a centroid which acts as a representation of the entire dataset. Then, Principal Component Analysis (PCA) is performed to project each centroid into a two-dimensional vector space in order to visualize the inter-dataset similarity. In addition to the suggestion in the framework [5], we intended to visualize inter- and intra-dataset similarity on a class level. We separated the documents into classes before averaging. This way, a centroid of each class for each dataset is obtained before performing PCA.

3.3. MI-based Word Rankings

The third method is a ranking of the 10 most relevant terms of the hate speech classes for each dataset. The framework used pointwise-mutual information (PMI) to rank the most relevant terms for each class.

However, experiments showed that the datasets have a significant amount of terms which only occur in one class. The PMI would then be the same highest value for all the words regardless of how often it occurs. Therefore, instead Mutual Information is used, which is the PMI weighted by the expectation of the joint word-class distribution, i.e. the relative word-class frequency. PMI is more useful, because a word that occurs more often is more relevant to a class than a word that does not.

3.4. Cross-Dataset Topic Model

This approach visualizes the most relevant topics of the hate speech datasets using *CluWords* [14]. First, a sample from all hate speech documents of all datasets is taken. Then, a vocabulary $\mathcal{V}$ of all unique terms in the sample is constructed. For each term t in the vocabulary, a word embedding vector is computed. Since this transformation is context-less in contrast to gBERT, German Fasttext word embeddings are used [15].

Lastly, the topics and CluWords are projected into a two-dimensional vector space using *t-SNE*, introduced by van der Maaten and Hinton [16], in order to visualize them.

Dataset name	Reference	Source	# of labeled samples	# of unlabeled samples	abusive % of labeled data	Inter-rater agreement
Covid2021	[19]	Twitter	4,960	0	22.28%	$\alpha = 91.50\%$
De-reddit-corpus	Unpub.	Reddit	0	2,992,835	-	-
Germeval2018	[20]	Twitter	8,541	0	33.84%	$\alpha = 78\%$
Germeval2019	[21]	Twitter	9,862	0	51.74%	$\kappa = 59\%$
Hasoc2019	[22]	Facebook, Twitter	4,669	0	11.63%	$\kappa = 88\%$
Hasoc2020	[23]	Twitter	3,400	0	28.62%	$\kappa = 83.3\%$
iHS	[24]	Twitter	1,249	275,022	40.19%	$\kappa = 44\%, 55\%$
IWG Hate. pub.	[25]	Twitter	469	0	23.45%	$\alpha = 38.29\%$
Telegram	[26]	Telegram	1,149	5,421,845	15.75%	$\alpha = 73.87\%$

Table 1
Overview of German hate speech datasets.

3.5. Inter-rater Reliability

The next approach focuses on the inter-rater reliability of the dataset annotators. Since not all datasets provide labeling information, this can only be calculated for those that do. The inter-rater reliability is calculated using Krippendorff's alpha [17].

3.6. SHAP Feature Importance

The last approach is based on SHAP (SHapley Additive exPlanations) [18]. It is a way to explain the importance of features for different hate speech classifiers. For each hate speech classifier f, an *explanation model* g is approximated. g uses simplified binary valued inputs x' that map to the original inputs through a mapping function h, i.e. $x \approx h(x')$. g then tries to approximate $g(x') \approx f(h(x'))$.

SHAP learns the values of the factors ϕ_i for each explanation model. Since g approximates our hate speech classifier f, the higher the value ϕ_i for a feature x'_i, the more important this feature is to the classifier. Instead of displaying the feature importance plot for a single example, we instead calculate the global feature importance for each classifier using SHAP barplots.

4. Hate Speech Data Collections

This section presents the datasets included in our analysis. Each dataset consists of recent real world examples of hate speech in online communication that were written in German. Our goal was to select largest and most recent datasets that can be found for this purpose. An overview is given in Table 1. The datasets are explained in detail in the following sections.

4.1. Covid2021

The first dataset contains German tweets collected from Twitter with COVID-19 as topic, published in 2021 [19]. The tweets were sampled from an *annotation pool* that is comprised in equal parts from three other pools: a *replies pool*, a *community pool* and a *topic pool*.

The replies pool was sampled from replies to posts published between 01.01.2020 and 20.02.2021 by three Twitter seed accounts that were identified as being influential and spreading COVID-19 misinformation. Only the tweets that contained one of 65 COVID-19 related keywords were considered for this purpose. The community pool then was fed from tweets sampled from the timeline of the accounts that replied to the seed accounts. The topic pool was sampled from tweets related to COVID-19 and hate speech. Lastly, tweets were sampled from the annotation pool and labeled by three annotators using a binary labeling scheme. A tweet was labeled ABUSIVE if it contains attacks or threats, insults, harassment, hate as well as degradation. Tweets were labeled NEUTRAL if otherwise. In total, 4,960 tweets were labeled, of which 1,105 were classified as abusive, and 3,855 as neutral. The Krippendorff's alpha is 91.5%.

4.2. Germeval2018

GermEval Shared Task on the Identification of Offensive Language, in short *Germeval2018* [20] consists of German tweets collected from Twitter. Specifically, tweets were sampled from the timeline of around 100 different users, each of which was selected because they posted both offensive, as well as non-offensive tweets. In total, 8,541 tweets were sampled and then manually annotated by one of three annotators using two different labeling schema, *coarse-grained* and *fine-grained*.

Coarse-grained is a binary classification scheme that labels a tweet as *OFFENSE* if it includes abusive language, insults or profanity, and *NEUTRAL* if not. Because the tweets were sampled around the time of the so called *refugee crisis* in Germany, the dataset creators noticed that certain non-offensive words had a high-frequency in the documents labeled as the hate speech class, but that did not appear in the non-hate speech class. Therefore, in order to debias the datasets, they further added non-hate speech tweets containing these words. Lastly, they split the dataset into a training and test set. The tweets sampled from each user only appear in one of the sets. In total, 2,890 tweets were labeled as abusive and 5,651 as neutral. It has a Krippendorff's alpha of 78%.

4.3. De-Reddit-corpus

De-Reddit-corpus was built by the authors of this paper containing posts from the German */r/de* subreddit from the reddit.com website. In total the corpus comprises 2,992,835 comments from 272,661 submissions created in 2019 or earlier. The comments were pseudo-labeled using a CNN model with word embeddings described by Schäfer and Burtenshaw [27]. The model was trained on Germeval2018 where it achieved an F1-score of 73.35%. Each comment was assigned a binary pseudo-label as well as the predicted label probability. While this dataset provides a significant number of examples of the phenomenon and can be useful for analyses, we do not recommend using it as a training dataset due to a lack of manual annotation supervision.

4.4. Germeval2019

GermEval Shared Task 2 on the Identification of Offensive Language from 2019, in short *Germeval2019* [21] also consists of a training and a test set. The training set of Germeval2019 consists primarily of tweets from the training and test set of Germeval2018, as well as some newly sampled tweets. The test set consists of entirely newly sampled tweets using the same method illustrated in Germeval2018. In addition, this time a specific effort was made to include tweets from users across the whole political spectrum.

The tweets were then manually annotated using the same labeling scheme as in Germeval2018. In total, Germeval2019 consists of 9,862 labeled tweets, 5,103 of which are labeled as abusive and the other 4,759 as neutral. The Cohen's kappa inter-rater reliability is $\kappa = 59\%$.

4.5. Hasoc2019

The fifth dataset is *Hate Speech and Offensive Content Identification in IndoEuropean Languages*, in short *Hasoc2019* [22]. The German subset of this dataset is used. The posts included in Hasoc2019 were sampled from Facebook and Twitter. The posts were manually labeled using a binary as well as a fine-grained labeling scheme. The binary labeling scheme annotates a post as either hate speech/offensive (HOF) or as non-hate speech (NOT). Here, hate speech is defined as posts containing Hate, offensive words, aggression, or profanity. In total, 4,699 posts were annotated, 543 of which were labeled as abusive, 4,126 as neutral, with a Cohen's kappa inter-rater agreement of $\kappa = 88\%$.

4.6. Hasoc2020

The sixth dataset is the German *Hate Speech and Offensive Content Identification in IndoEuropean Languages* dataset from 2020, *Hasoc2020* [28]. It consists of tweets sampled from a collection of tweets created in May 2019. First, non-German tweets were filtered using the language attribute metadata provided by Twitter. Then, the tweets were sampled using a Support Vector Machine (SVM) hate speech classifier trained on Germeval2018 and Hasoc2019. The classifier was trained in such a way that it achieves an F1-score of around 0.5. All tweets that were labeled hateful by the classifier were included in the sample. In addition, 5% of tweets that the classifier did not label as hateful were also included.

For the binary hate speech classification task, tweets were labeled as HOF when they contained hate, offensive or profane content, and NOT when otherwise. All tweets in the dataset were manually labeled twice by two different annotators. In cases when the two annotators disagreed on a label, a third annotator who did not yet see the tweet assigned the label. 3,400 tweets were labeled in total, with 973 assigned to abusive and 2,427 to neutral. Cohen's kappa inter-rater agreement is $\kappa = 83.3\%$.

4.7. iHS

iHS is an unpublished dataset of potentially illegal hate speech [24]. The creation of this dataset consisted of two steps. First, court cases were collected in which German social media posts were identified to be violating certain laws associated with hate speech. Then, using these posts as examples, 102 tweets from Germeval2019 were manually extracted that were deemed to potentially violate German law in order to create manual annotation guidelines [29].

Lastly, text posts from Twitter were annotated using these guidelines in several annotation rounds. Tweets were assigned to one of six categories: *Public incitement to commit offences*, *Incitement of masses*, *malicious gossip* and *defamation*, *insults*, *offensive language* and *other*. The offensive language category contains 214 tweets that are not illegal but still deemed hateful. The remaining 747 tweets are labeled as *other*. The Fleiss kappa inter-rater agreement ranged between $\kappa = 44\%$ and $\kappa = 55\%$.

For the purpose of binary hate speech classification, the first five categories are considered as abusive, the *other* category as neutral. In addition to the 1,249 labeled tweets, 275,022 unlabeled tweets from were also available.

4.8. IWG Hatespeech public

IWG Hatespeech public [25] contains German hate speech in the context of the refugee crisis in Europe. This dataset consists of tweets from Twitter written in 2016. They were sampled using keyword search with 10 different hashtags that were considered to likely contain a disproportionate amount of hate speech. After filtering these, the tweets were manually annotated by splitting the dataset into six parts, each of which was annotated by two of six annotators. They used a binary labeling scheme of labeling a tweet as hate speech if it violates the Twitter definition on *hateful conduct*, and neutral if not. In total, the dataset contains 469 tweets, 110 abusive and 359 neutral. The Krippendorff's alpha is 38.29%.

4.9. Telegram

The last dataset that is analyzed is referred to as *Telegram* [26]. It contains messages from German Telegram channels that were posted between 01.01.2019 and 15.03.2021.

The authors used a snowball sampling strategy. Specifically, they first collected messages from 51 public seed channels known to spread hate. Using these as starting points, they then collected all messages from channels that were mentioned by them or forwarded to them from other channels. This procedure was repeated once again for the newly acquired messages. To filter out languages other than German, the message texts were fed into a classifier using multilingual word vectors from fastText [15]. The messages were manually labeled as abusive and neutral by five annotators using the same labeling scheme as the covid2021 dataset [19]. In total, 1,149 messages were labeled, of which 181 were abusive and 968 neutral. The Krippendorff's alpha is 73.87%. In addition, the unlabeled dataset was also provided to me. It consists of 5,421,845 Telegram messages that were pseudo-labeled by a hate speech classifier [26].

5. Experiment Results

We apply the framework by Wich et al. [5] as described in Section 3 on the datasets discussed in the previous section. In this section, we present and discuss the results of this analysis. We provide the code used for this research on GitHub[1].

5.1. LSI-based Similarity

We now examine the results of the LSI-based intra-dataset similarity experiment. Table 2 shows the similarity values of the binary hate speech classes within each of the nine hate speech

[1] https://github.com/MarkusBertram/Cross-Dataset-Generalization-of-German-Hate-Speech-Datasets

Dataset	abusive → abusive	abusive → neutral, neutral → abusive	neutral → neutral
Covid2021	0.70	0.71	0.72
De-reddit-corpus	0.29	0.26	0.24
Germeval2018	0.39	0.41	0.44
Germeval2019	0.41	0.40	0.36
Hasoc2019	0.53	0.57	0.61
Hasoc2020	0.48	0.50	0.56
iHS	0.47	0.49	0.51
IWG Hatespeech public	0.28	0.17	0.21
Telegram	0.34	0.37	0.44

Table 2
LSI-based intra-dataset class similarity with 16 dimensions.

datasets using 16 LSI-dimensions. The left column are the respective datasets, the top row indicates the direction of the LSI similarity between the classes. The experiment was repeated for different LSI-dimensions with no significant changes.

In general, the differences between the classes of each dataset seem rather small. In Germeval2018, Hasoc2019, Hasoc2020, Covid2021 and Telegram, the neutral class is most similar with itself than the others. In Germeval2019, De-reddit-corpus, iHS, and IWG Hatespeech public, the hate speech class is most similar with itself.

The low absolute and relative difference between the classes can be interpreted as indicating high intra-dataset similarity, i.e. a small difference in the marginal distributions of the covariates in each dataset for each class. A classifier therefore is less likely to simply memorize class-specific phrases or words.

5.2. Word-embedding-based Similarity

Figure 1 shows the two-dimensional PCA projection of the word embedding centroids of each dataset. For this, all classes of each dataset were combined. Interestingly, Germeval2019 is closer to De-reddit-corpus than to Germeval2018. This is surprising because Germeval2019 is comprised in large parts of tweets from Germeval2018.

Figure 2 depicts the two-dimensional projection of each individual class. At least in this projection, there is no clear *abusive* or *neutral* cluster. Therefore, this experiment is an indication that the cluster assumption may not hold true. However, we have to keep in mind that we are projecting a 768-dimensional vector space into two dimensions which limits the interpretability of the result.

5.3. MI-based Word Rankings

The Mutual Information-based word rankings for both the abusive and the neutral class in each dataset show which terms can be considered most relevant for each class. They are displayed in Table 3 in descending order. Unsurprisingly, the majority of datasets rank several different terms that indicate an insult or profanity highly, i.e. *idiot*, *dumm*, *scheiß*, *abschaum*, *schwein*,

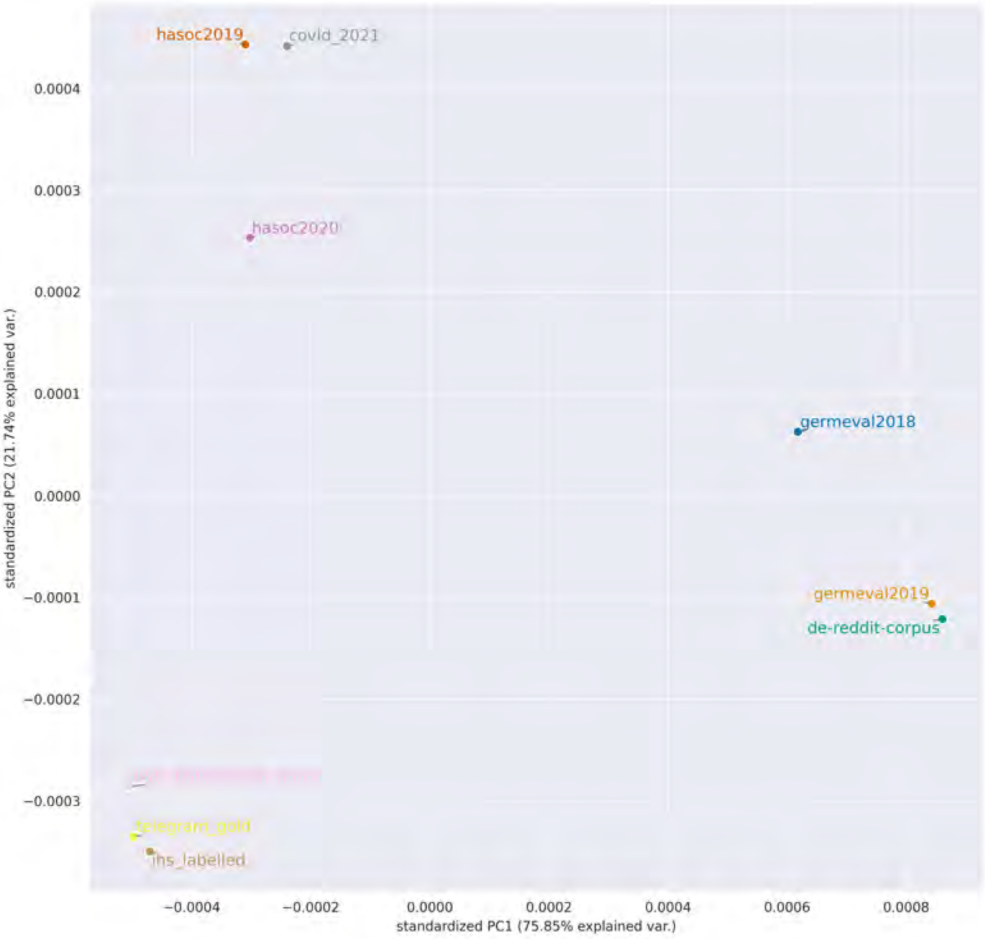

Figure 1: Word-embedding-based dataset similarity.

ferkel, *hure*, *hurensohn* and *nutte*. The latter three terms, together with the term *frau*, clearly show misogyny also being a focus in these hate speech datasets. Terms often used in a racist manner can also be found, like *flüchtling* or *islam*.

We can also conclude that there is a clear temporal shift that one should consider when generalizing across datasets. Terms like *Merkel* will clearly be more popular in some time-periods than others.

5.4. Cross-Dataset Topic Model

20 topic clusters were calculated using CluWords. The topics and two-dimensional projection of each document can be seen in Figure 3.

Most topics do not appear to be relevant to hate speech, with the exception of three topics in the lower right half of the plot. Topic T4 (*terroristen*, *faschisten*, *moslems*, etc.), topic T6 (*feministen*, *terrorgruppen*) as well as topic T15 (*inhaftierung*, *abschieberaten*, etc.) can be

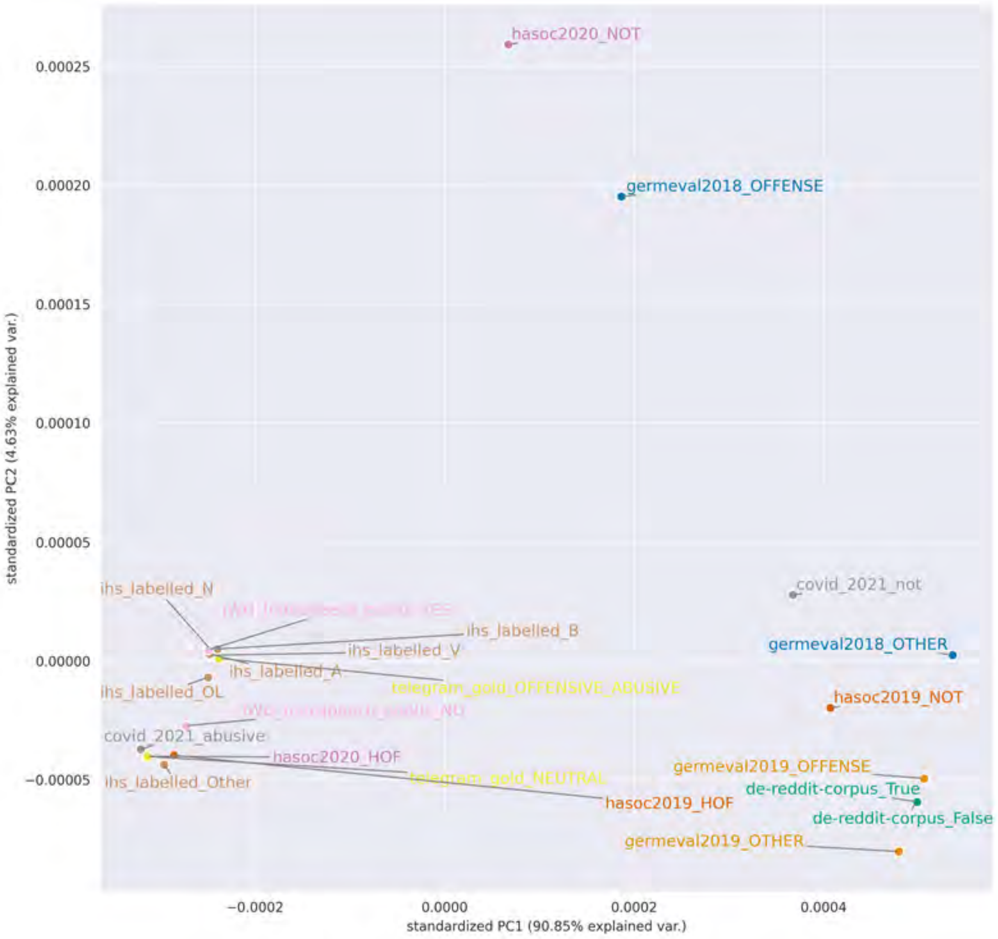

Figure 2: Word-embedding-based class similarity.

attributed to hate speech. In addition, there is no clear clustering of datasets to specific topics. This indicates that the combined hate speech datasets have several different topics and no obvious bias.

5.5. Feature Importance using Shapley Values

This analysis investigates the most important features of classifiers trained on each dataset, in descending order. We show the results for Germeval2019 as an example in 4. The y-axis contains the most important sub-tokens while the x-axis displays the global importance of each feature. For all classifiers, the feature importance was measured on the same dataset which was sampled from all hate speech datasets, combined.

The results show that several classifiers give high weights to the same features. For example, the word *Sklaven* is important for the classifiers trained on Covid2021, Germeval2018, Hasoc2019 and Hasoc2020. Another example is the word *Schweine*, which is important in Hasoc2020, iHS

Dataset	MI-based word rankings for the hate speech class
Covid2021	corona, dumm, merkel, mensch, virus, geben, glauben, anderer, idiot, einfach
De-reddit-corpus	einfach, geben, halt, anderer, sehen, leute, sagen, mensch, finden, eigentlich
Germeval2018	merkel, frau, deutsch, deutschland, dumm, geben, grüne, sehen, deutsche, land
Germeval2019	merkel, frau, deutschland, deutsch, dumm, sehen, land, geben, spd, deutsche
Hasoc2019	alias, loch, deutschland, papa, merkel, capitol, land, frau, sagen, sehen
Hasoc2020	arsch, hurensohn, scheiß, porno, dumm, deutsch, gratis, frau, ficken, halt
iHS	fuck, arsch, scheiße, ficken, nutte, dumm, idiot, abschaum, hure, einfach
IWG Hatespeech public	flüchtling, kind, frau, absagen, vergewaltigen, finden, schwimmbad, menschenwürde, verstoß, sexuell
Telegram	kind, geben, volk, mensch, deutsch, deutschland, anderer, bringen, krank, sehen

Table 3
Mutual Information-based word rankings for hate speech classes.

and Telegram. However, the majority of features that the classifier consider are ranked and weighted differently. This shows that, given the same sub-token, each classifier may come to a different conclusion regarding the predicted hate speech label.

6. Conclusion

In this work, we present a survey of German hate speech datasets and apply the framework suggested by Wich et al. [5] to compare their contents. This analysis sheds some light on the datasets and how they are similar in some regards, but diverse in topics. Our analysis shows that the contained topics are quite heterogeneous and thus a cross-dataset classification would be rather difficult. Although the analysis helps to better understand the datasets, it cannot alone determine how good the datasets are and how they can lead to better generalizability. Further experiments on the the performance of classifiers across German datasets are necessary [30].

Further new directions in hate speech detection include the creation of datasets for low resource languages (e.g. [31]), the analysis of context [32], the generation of counter speech [33] and the design of interfaces for diverse user groups of such AI systems [34].

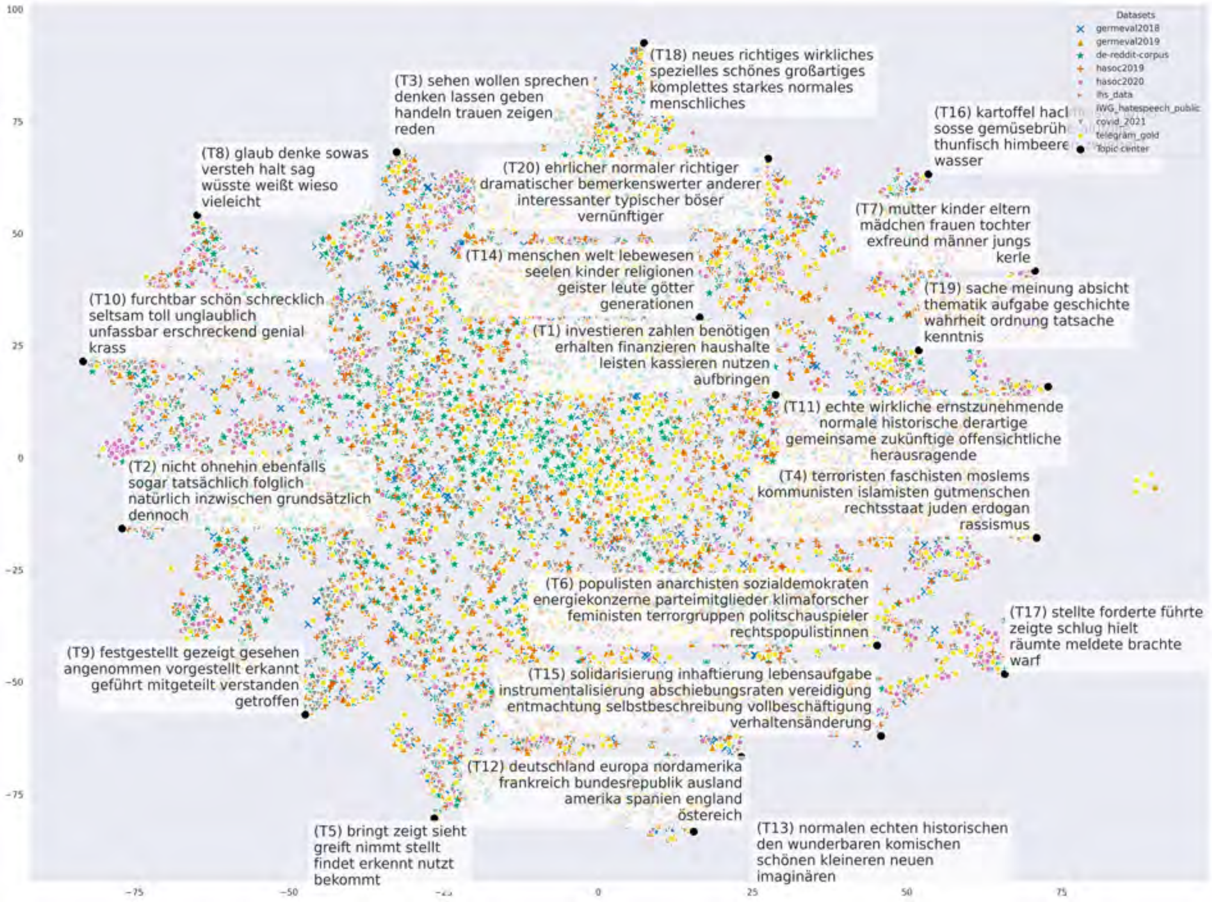

Figure 3: Topic model for the hate speech classes.

References

[1] S. Jaki, S. Steiger, Digitale Hate Speech: Interdisziplinäre Perspektiven auf Erkennung, Beschreibung und Regulation, Springer Nature, 2023. doi:10.1007/978-3-662-65964-9.

[2] B. Di Fátima, Hate Speech on Social Media: A Global Approach„ LabCom - Comunicação e Artes - Universidade da Beira Interior, 2023. URL: https://labcomca.ubi.pt/hate-speech-on-social-media-a-global-approach/.

[3] European Commission, Laws on digital services and markets: European Commission welcomes yes from the European Parliament, 2022. URL: https://ec.europa.eu/commission/presscorner/detail/en/ip_22_4313.

[4] S. Modha, T. Mandl, P. Majumder, D. Patel, Tracking hate in social media: Evaluation, challenges and approaches, SN Computer Science 1 (2020) 1–16. doi:10.1007/s42979-020-0082-0.

[5] M. Wich, T. Eder, H. Kuwatly, G. Groh, Bias and comparison framework for abusive

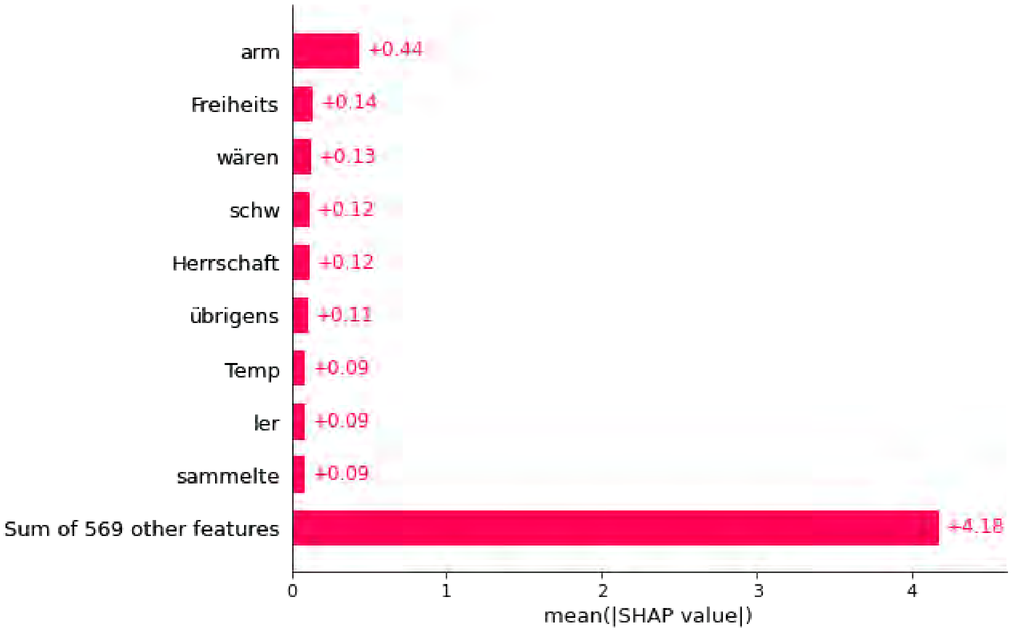

Figure 4: Global feature importance of the Germeval2019 dataset.

language datasets, AI and Ethics 2 (2022) 1–23. doi:10.1007/s43681-021-00081-0.

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

[7] T. Mandl, KI-Verfahren für die Hate Speech Erkennung: Die Gestaltung von Ressourcen für das maschinelle Lernen und ihre Zuverlässigkeit, in: Digitale Hate Speech, Springer Berlin Heidelberg, 2023, pp. 111–130. doi:10.1007/978-3-662-65964-9_6.

[8] P. Fortuna, J. Soler, L. Wanner, Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets, in: Twelfth Language Resources and Evaluation Conference, ELRA, Marseille, France, 2020, pp. 6786–6794. URL: https://aclanthology.org/2020.lrec-1.838.

[9] I. Nejadgholi, S. Kiritchenko, On cross-dataset generalization in automatic detection of online abuse, 2020. URL: https://arxiv.org/abs/2010.07414. doi:10.48550/ARXIV.2010.07414.

[10] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, J. Pierrehumbert, HateCheck: Functional Tests for Hate Speech Detection Models, in: 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL, Online, 2021, pp. 41–58. doi:10.18653/v1/2021.acl-long.4.

[11] W. Yin, A. Zubiaga, Towards generalisable hate speech detection: a review on obstacles and solutions, 2021. URL: https://arxiv.org/abs/2102.08886. doi:10.48550/ARXIV.2102.08886.

[12] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science 41 (1990) 391–407. doi:https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[14] F. Viegas, S. Canuto, C. Gomes, W. Luiz, T. Rosa, S. Ribas, L. Rocha, M. A. Gonçalves, Cluwords: Exploiting semantic word clustering representation for enhanced topic modeling, in: Twelfth ACM International Conference on Web Search and Data Mining, WSDM '19, ACM, New York, NY, USA, 2019, p. 753–761. doi:10.1145/3289600.3291032.

[15] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, arXiv preprint arXiv:1607.04606 (2016).

[16] L. van der Maaten, G. Hinton, Visualizing Data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[17] K. Krippendorff, Computing Krippendorff's Alpha-Reliability, 2011. URL: https://repository.upenn.edu/asc_papers/43.

[18] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems 30 (2017). URL: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[19] M. Wich, S. Räther, G. Groh, German abusive language dataset with focus on COVID-19, in: Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021), KONVENS 2021 Organizers, Düsseldorf, Germany, 2021, pp. 247–252. URL: https://aclanthology.org/2021.konvens-1.26.

[20] M. Wiegand, M. Siegel, J. Ruppenhofer, Overview of the GermEval 2018 shared task on the identification of offensive language, in: Proceedings of the GermEval 2018 Workshop, 14th Conference on Natural Language Processing KONVENS 2018, 2018, pp. 1–10.

[21] J. M. Struß, M. Siegel, J. Ruppenhofer, M. Wiegand, M. Klenner, Overview of GermEval Task 2, 2019 shared task on the identification of offensive language, in: 15th Conference on Natural Language Processing (KONVENS), Oct. 9 – 11, 2019, Erlangen-Nürnberg, German Society for Computational Linguistics & Language Technology, München [u.a.], 2019, pp. 352–363. URL: https://nbn-resolving.org/urn:nbn:de:bsz:mh39-93197.

[22] T. Mandl, S. Modha, P. Majumder, D. Patel, Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages), in: Working Notes of the Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE, CEUR-WS, 2019. URL: http://ceur-ws.org/Vol-2517/T3-1.pdf.

[23] T. Mandl, S. Modha, A. Kumar M, B. R. Chakravarthi, 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, ACM, New York, NY, USA, 2020, p. 29–32. doi:10.1145/3441501.3441517.

[24] J. Schäfer, K. Boguslu, Towards annotating illegal hate speech: A computational linguistic

approach, Technical Report, Detect Then Act (DTCT) Technical Report 3, 2021. URL: https://dtct.eu/wp-content/uploads/2021/10/DTCT-TR3-CL.pdf, iSSN 2736-6391.

[25] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, M. Wojatzki, Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis, in: M. Beißwenger, M. Wojatzki, T. Zesch (Eds.), NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, volume 17, Bochum, 2016, pp. 6–9.

[26] M. Wich, A. Gorniak, T. Eder, D. Bartmann, B. E. Çakici, G. Groh, Introducing an Abusive Language Classification Framework for Telegram to Investigate the German Hater Community, 2021. URL: https://arxiv.org/abs/2109.07346. doi:10.48550/ARXIV.2109.07346.

[27] J. Schäfer, B. Burtenshaw, Offence in dialogues: A corpus-based study, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), INCOMA Ltd., Varna, Bulgaria, 2019, pp. 1085–1093. URL: https://aclanthology.org/R19-1125. doi:10.26615/978-954-452-056-4_125.

[28] T. Mandl, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, J. Schäfer, Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification in Indo-European Languages, in: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, volume 2826, CEUR-WS.org, 2020, pp. 87–111. URL: http://ceur-ws.org/Vol-2826/T2-1.pdf.

[29] K. Boguslu, J. Schäfer, Annotationsrichtlinien für illegale Hassrede, 2021. URL: https://dtct.eu/wp-content/uploads/2021/09/Annotationsrichtlinien_iHS.pdf.

[30] N. Seemann, Y. S. Lee, J. Höllig, M. Geierhos, Generalizability of Abusive Language Detection Models on Homogeneous German Datasets, Datenbank-Spektrum 23 (2023) 15–25. doi:10.1007/s13222-023-00438-1.

[31] B. R. Chakravarthi, P. K. Kumaresan, R. Sakuntharaj, A. K. Madasamy, S. Thavareesan, B. Premjith, S. K, S. C. Navaneethakrishnan, J. P. McCrae, T. Mandl, Overview of the HASOC-DravidianCodeMix Shared Task on Offensive Language Detection in Tamil and Malayalam, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 13-17., volume 3159 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 589–602. URL: https://ceur-ws.org/Vol-3159/T3-1.pdf.

[32] H. Madhu, S. Satapara, S. Modha, T. Mandl, P. Majumder, Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments, Expert Systems with Applications 215 (2023) 119342.

[33] S. S. Tekiroglu, Y. Chung, M. Guerini, Generating Counter Narratives against Online Hate Speech: Data and Strategies, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL , Online, July 5-10, Association for Computational Linguistics, 2020, pp. 1177–1190. doi:10.18653/v1/2020.acl-main.110.

[34] L. Sontheimer, J. Schäfer, T. Mandl, Enabling Informational Autonomy through Explanation of Content Moderation: UI Design for Hate Speech Detection, in: Mensch und Computer 2022-Workshopband, Gesellschaft für Informatik e.V., 2022. doi:10.18420/MUC2022-MCI-WS12-260.