Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

GMHP7k: A Corpus of German Misogynistic Hatespeech Posts

Jonas Glasebach[1], Max-Emanuel Keller[2], Alexander Döschl[2], Peter Mandl[2]

[1]KPMG AG, Munich, Germany
[2]HM Hochschule München University of Applied Sciences, Munich, Germany
[1]jglasebach@kpmg.com, [2]max-emanuel.keller@hm.edu, [2]alexander.doeschl@hm.edu, [2]peter.mandl@hm.edu

Abstract

We provide a German corpus consisting of 7,061 posts authored by users of social media platforms. A group of volunteers annotated each post according to hatespeech and misogynistic/misogynous hatespeech in a binary fashion. The interrater reliability over all annotators according to Fleiss' Kappa is 0.6409 for hatespeech and 0.8258 for misogynistic hatespeech. Furthermore, baseline measurements with machine learning based text classification with BERT are presented. Initial experiments with the corpus achieve macro average F_1-scores up to 0.79 for hatespeech and 0.75 for misogynistic hatespeech. The dataset of the corpus on German Misogynistic Hatespeech Posts (GMHP7k) is publicly available.

Introduction

User interactions are an integral part of the internet. Whether on social media, discussion platforms, or comment sections, the quantity of postings is immense. It is evident that not all of these posts are positive; some contain offensive and, at times, dehumanizing content. One prevalent form of this hatred is directed towards women (Ging and Siapera 2018). However, identifying and removing such texts requires a significant investment of human resources. Hence, there is a concerted effort to automate the hatespeech detection process as much as possible, using machine learning methods.

In order to train a system to identify hatespeech, texts annotated as either hatespeech or not hatespeech are required. There exist numerous corpora containing posts from various sources that have been annotated for aspects such as sentiment in comments (Cieliebak et al. 2017; Narr, Hülfenhaus, and Albayrak 2012) or indications of hatespeech (Waseem 2016). However, to the best of our knowledge, there is no publicly available corpus of texts in the German language annotated specifically for misogynistic hatespeech, which often includes subtly disguised hostility towards women.

We have created a new corpus called German Misogynistic Hatespeech Posts (GMHP7k) of 7,061 comments, specialized in the detection and differentiation of general hatespeech and the specific case of misogynistic hatespeech. The posts originate from social media platforms and are written in German. The annotation process was conducted by a group of six volunteers who assessed posts for both hatespeech in general and misogynistic hatespeech.

The GMHP7k dataset facilitates the training of a classifier for detecting (misogynistic) hatespeech in new texts. Our goal is to develop a framework for the detection of (misogynistic) hatespeech in user posts on online platforms. We believe that the corpus can also prove valuable for other applications in the field of natural language processing (NLP) and classification. For this reason, we have made the dataset of the corpus publicly available to the scientific community. In this paper, we make the following contributions:

- We put our research into the context of related work in Section Related Work.
- We propose a methodology to annotate short texts in Section Dataset and describe the actual annotation process in Section Details.
- We provide a first baseline classification of the texts by (misogynistic) hatespeech using a fine-tuned BERT model in Section Experiments and discuss the results in Section Discussion.
- We make the annotated corpus publicly available for research purpose in Section How to Use the Corpus.

Related Work

In the detection of hatespeech, the creation of corpora is and has been a crucial task. The corpora with hatespeech data are used as data for training and evaluation of machine learning models for the detection of hatespeech, but can also serve as an evaluation base line for the comparison with other works. Hence, there exist a number of related works in the domain of hatspeech classification that contain data from different sources (e.g. social media) and in different languages. In this section we focus on datasets that are related to either of the two main aspects of our dataset, namely hateful texts in German language and misogyny. To avoid bias in the model, the dataset should cover a diverse range of subject areas, ensuring that there are enough positive examples for the different categories. Additionally, the team's available capacities must be taken into account when creating a new dataset, particularly concerning the manual annotation of the corpus.

Vidgen and Derczynski (2020) analyzed a total of 63 hatespeech datasets and provide the website hatespeechdata.com

that contains an overview of the datasets in several languages. They found that 25 datasets contain texts in English while only four in German. On average, the datasets contained about 8,000 texts, with approximately 36.7 % abusive texts on average.

As there are only limited datasets in German, we draw the inspiration for our research work and our methodology from the One Million Post Corpus (Schabus, Skowron, and Trapp 2017). This corpus is a collection of one million comments from an Austrian online newspaper site, of which 11,773 were classified in seven categories.

Roß et al. (2016) presented a dataset consisting of 469 anti-refugee hate posts in German from the European refugee crisis on Twitter. Bretschneider and Peters (2017) contributed about 6,000 German Facebook posts with anti-foreigner statements annotated on strength and target of the statement and about 11 % of abusive posts. A dataset for the GermEval 2018 task on identifying offensive statements was provided by Wiegand, Siegel, and Ruppenhofer (2019). This dataset consists of 8,500 posts from Twitter in German, which have been annotated on the occurrence (binary) and strength of offensive statements. Approximately 34 % of the texts in the dataset are classified as abusive. Mandl et al. (2019) supplied a dataset on hatespeech and offensive content identification with about 4,500 German posts from Twitter and Facebook annotated on availability and strength of hatespeech or offensive statements with about 24 % of abusive posts. Assenmacher et al. (2021) provided a dataset with 85,000 posts in German from the German newspaper Rheinische Post, that were annotated on availability and type of offensive language, with a quote of 8.4 % abusive posts. The annotated categories also include sexism as specific type, which however, is not restricted to misogynistic posts. Demus et al. (2022) contributed a dataset consisting of 10,000 posts in German from Twitter on offensive language, that were annotated on hatespeech, sentiment and several other categories, with a quote of about 10.85 % abusive posts.

To the best of our knowledge there exists no corpus with posts related to the detection of misogyny in German langugage. However, there have been previous works on such corpora in other languages like English, Spanish and Italian (Shushkevich and Cardiff 2019). Fersini, Rosso, and Anzovino (2018) describe a task on the identification of misogyny from 8,115 tweets with 4,138 in Spanish and 3,977 in English annotated on the type (five categories) and the target (active or passive) of misogynistic behavior, with a quote of about 48.24 % abusive posts. A second similar task of Fersini, Nozza, and Rosso (2018) targets the identification of misogyny from 10,000 tweets with 5,000 each in Italian and in English annotated on the type (five categories) and the target (active or passive) of misogynistic behavior, with a quote of about 45.85 % abusive posts. Another corpus on abusive language also targeting sexism but not specifically misogyny is the one of Waseem (2016) that contains 6,909 posts from Twitter in English, that were annotated by amateurs and experts with the labels racist, sexist, both or neither.

Dataset

The corpus presented in this work consists of 7,061 user posts. The posts come from two social media platforms, namely Facebook and X (formerly Twitter), and are written in German. The focus of the corpus is on the texts that were left as comments by users as reaction on posts published by politicians and other public figures. Each comment has an ID, the text and a publication time.

The posts were annotated by a group of six volunteers who rated posts as either *hatespeech* or *not hatespeech* and, in the case of hatespeech, whether it could be categorized as *misogynistic hatespeech* or *not misogynistic hatespeech*. All of the data mentioned above are part of the corpus that we make available to the scientific community.

In the following sections we describe the procedure for the creation of the corpus. First, we give an overview of the annotation process, which is divided into several phases (Section Annotation Process). We then describe the classes and guidelines used in the annotation process (Section Classes to Annotate) before we give the details of the data collection carried out (Section Data Collection). Finally, in Section Inter-rater Reliability, we discuss the calculation of the inter-rater reliability, which measures the consistency of the volunteers in the annotation of posts.

Annotation Process

The annotation process consists of three consecutive phases. During Phase 1, the volunteers were prepared to create a common understanding of the annotation process. The volunteers reviewed several posts and familiarized themselves with the classes that had to be annotated. Subsequently, the goal of Phase 2 was to ensure that the volunteers actually annotate according to the same rules. For this purpose, the agreement of the volunteers in the annotation was determined by calculating an inter-rater reliability. Phases 1 and 2 thus represented two training phases, hence their results were discarded, but retained in the dataset for documentation purposes. Based on the jointly agreed set of guidelines, further posts were annotated in Phase 3. Each of these posts was annotated by only one person and represent the productive part of the corpus. The annotation process presented in this paper has proven reliable for the annotation of social media posts in an earlier work (Keller et al. 2019), which served as a basis for a system that assists users in creating successful posts (Keller et al. 2018; Keller, Döschl, and Mandl 2023).

Classes to Annotate

During annotation, volunteers rated two aspects of a post: the presence of hatespeech and misogynistic hatespeech. The availability of hatespeech depends on perception of the comment text by the annotators and can be rated as *hatespeech* or *not hatespeech*. The misogynistic hatespeech, on the other hand, can be either *misogynistic hatespeech* or *not misogynistic hatespeech*.

Based on various definitions (Meta 2024; Sponholz 2018; ECRI 2016), in this work, misogynistic hatespeech is regarded as a subset of hatespeech. Misogynistic hatespeech

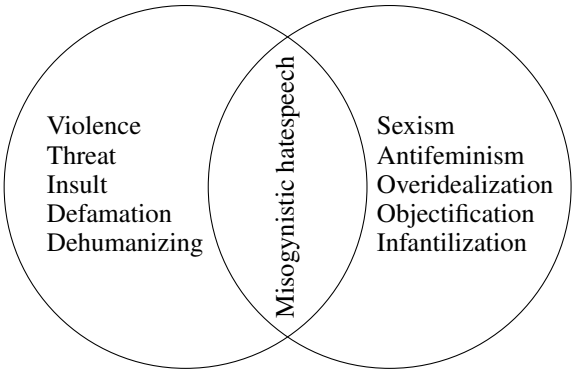


Figure 1: Categorization of (misogynistic) hatespeech.

can be characterized as written content that threatens violence or trivializes it, insults or defames, negatively stereotypes, demeans, or objectifies and dehumanizes women on the basis of their gender (see Fig. 1). Although misogynistic hatespeech is often reduced to sexism in the public perception, according to this definition it includes threats of any kind against women. This encompasses both threats of physical violence (e.g., "Wenn ich dich sehe, werde ich dich fertig machen"/"If I see you, I will hurt you") and psychological violence such as harassment through stalking or surveillance (cyberstalking). Insulting individuals with swear words and scatological terms also constitutes hatespeech. The evolution of language on the internet continually produces covert euphemisms and code words serving as insults. Sexist insults or comments about appearance objectify women (e.g., "Du siehst zwar ganz niedlich aus, aber deine politische Einstellung ist Nonsens"/"You might look cute, but your political stance is nonsense"). Employing negative stereotypes or suggesting that women deserve a particular social role (e.g., "Frauen haben in der Politik nichts zu suchen"/"Women do not belong in politics") also forms part of misogynistic hatespeech.

As previously established, hatespeech is multifaceted and often not unequivocally identifiable (Vidgen and Derczynski 2020). This became apparent in Phases 1 and 2 of the annotation process when tweets were discussed collectively. As a result, the annotators expanded and refined the definition of (misogynistic) hatespeech with the following guidelines:

- *Irony:* Ironic statements, frequently used as a rhetorical device, implies that the author of the text means the opposite of what is actually written. It is often employed to implicitly express a negative judgement or criticism. Hence, for this corpus, an ironic text could be construed as hatespeech/misogynistic hatespeech, even if the author had different intentions.
- *Vulgar terms:* Vulgar language refers to coarse and offensive language. Some words have become part of the regular vernacular and are used for swearing or in conversation among friends. Yet, when offensive content is directed at a person, it is considered an insult and is marked as hatespeech. This means that vulgar language is not automatically classified as hatespeech in the annotation, but that the conventional use of the term in the German language is taken into account. Hate comments in the form of quotes and song lyrics present a special case. It is assumed that the post author agrees with the content of the lyrics and thus shares it.
- *Context:* Judging tweets and comments can be challenging without additional context, as they often reference other tweets, hashtags, or media. Phrases like "Frauen wissen das aber nicht"/"Women are unaware of this" might not constitute hatespeech depending on the discussion context. Consider a discussion centered on the experience of being a man, where this sentence could be deemed non-malicious. Conversely, it is conceivable that the author may be implying criticism towards women or harboring misogynistic intentions. Nevertheless, research by Gomez et al. (2019) and Pavlopoulos et al. (2020) indicates that supplementary contextual information did not yield improved classification outcomes. Accordingly, we applied the same method as with irony - literal comprehension of the content was employed. However, given the sensitivity of the subject matter, such texts were predominantly classified as misogynistic hatespeech in most cases.
- *Recipient:* The recipient is an essential factor in evaluating hatespeech. Statements directed towards individuals or groups identified by innate, personal characteristics are subject to stricter scrutiny compared to critiques aimed at legal entities. The inherent vulnerability associated with personal traits necessitates a more rigorous assessment to prevent the perpetuation of discrimination and targeted harassment.
- *Gender-related content:* Hatespeech against women is not always synonymous with misogynistic hatespeech. The misogynistic condition applies only if the hatespeech is gender-related. The difference becomes clear when considering the following two examples:

  1. "Diese Frau hat den Verstand von Eierlikör."
     "This woman has the mind of eggnog."
  2. "[@TwitterUser] Danke fotze"
     "[@TwitterUser] Thanks, cunt"

  The first example is an insult directed at a woman and therefore constitutes hatespeech. The insult focuses on behavior and not on the gender of the woman, so it is not misogynistic hatespeech. In the second tweet, a female person is insulted as well. The author, however, uses a vulgar word ("Fotze"/"cunt"), which is clearly derogative towards women. Therefore, it constitutes misogynistic hatespeech.

The guidelines mentioned above do not completely encompass the complexity of hatespeech. In many cases, it is difficult for annotators to make an objective assessment. During discussion rounds, the following reasons for variations in annotations have emerged: background knowledge on a topic, different tolerance thresholds, and subjective affirmations or aversions.

User ID	Username	#Tweets
301025783	@maltegymka	1053
2285678941	@schwulemiker	2309
3037183941	@zwingelstaender	1748
1456678385560784896	@wahlbuergerde	1299
1459594844331023872	@sophialuani1510	149

Table 1: Twitter troll profiles.

Twitter trend	Timespan	#Tweets
#luisaneubauer	03/16 – 03/18/2022	529
Rbb24	01/11 – 01/13/2022	603

Table 2: Hateful Twitter trends.

Data Collection

Given a random selection of text examples, it is likely that only a small proportion of these are actually hate comments. This leads to a very high annotation effort (Founta et al. 2018). In order to obtain a great proportion of misogynistic hatespeech, while also obtaining a comprehensive dataset, a mixture of different methods were used in this study.

1. *Keyword search*: A very effective method for selection is to search for specific keywords. This was based on the word collection from Hatebase.org (Sharma, Agrawal, and Shrivastava 2018), which also lists German words that are offensive to women. In addition, English insults are also frequently used in German, which is why the list was expanded to include known English translations. An endpoint of the Twitter API was used for the search, which sets up a real-time stream (GET /2/tweets/search/stream). The following rule with words offensive towards women was used to filter the data. The (-) sign in front of a parameter indicates that it is excluded from the search.

```
(dorfmatratze OR flittchen OR Fotze
  OR Hürchen OR Hure OR Schlampe
  OR Bitch OR cunt OR hoe OR pussy
  OR slut OR whore) lang:de
-filter:links -is:retweet -has:media
```

   A minor drawback of this approach is an inherent bias in relation to the keywords used. For this reason, the number of tweets and replies used was limited to 1,000.

2. *Twitter troll profiles*: Furthermore, conspicuous user profiles that increasingly post misogynistic content were identified. These profiles were found using the results of the keyword search previously carried out. The manual selection was aided by clear statements in the profile description and the indication from Twitter that the profile may contain sensitive content. Another Endpoint of the Twitter API was used to query the user profiles (GET /2/users/:id/tweets). The selected profiles are listed in Table 1.

3. *Hateful Twitter trends*: In addition, the dataset was expanded by two randomly found Twitter trends which showed an increased amount of misogynistic comments listed in Table 2.

In addition to the primary data acquisition of new texts, we also used existing texts from other datasets, which were manually selected from the datasets to increase the amount of relevant misogynistic hatespeech. Although no existing German dataset with a focus on misogynistic hatespeech could be found (see Section Related Work for details), several datasets on hatespeech also contained texts with misogynistic hatespeech, as we define misogynistic hatespeech as a subset of hatespeech (see Section Classes to Annotate).

The dataset by Charitidis et al. (2020) is based on Twitter profiles of well-known journalists and news media, which were annotated according to hatespeech. Each tweet was only annotated once. In order to create a consistent annotation, regular quality controls were set up by a supervisor. The dataset is publicly available for research and non-commercial use only. It contains the tweet ID and the label, while the corresponding texts have to be obtained via the Twitter API. After data cleansing, a total of 325 tweets containing hatespeech were added to the dataset.

The dataset by Mandl et al. (2020) was created as part of the NLP Hate Speech and Offensive Content (HASOC) challenge and is available under the Creative Commons Attribution 4.0 International license. It is based on random tweets from the month of May 2019. Each tweet was annotated with the labels hatespeech, insult, swearing/vulgar content. Two, and in the case of disagreements three, students annotated each tweet. In the event that there was a disagreement and no third person was available, a specially defined logic was used to select the more reliable annotator. The reliability was calculated by matching previous annotations with other annotators. Based on this dataset, after data cleanup, 413 potential hatespech texts were added.

A second HASOC Challenge held in 2019 (Mandl et al. 2019) was based on a different dataset obtained from Twitter and Facebook and is available under the Creative Commons Attribution 4.0 International license as well. The authors used a keyword search to restrict the data and expanded it to include content from conspicuous user profiles. A total of 96 % of the German content was annotated by two people. The dataset contains the same label types as the HASOC 2020 dataset. In contrast to 2020, however, no tweet IDs are available. After the dataset cleanup, a total of 405 potential hatespeech texts remained.

In preparation for the annotation, the dataset first had to be cleaned of disruptive elements. These include duplicates, as they only reflect existing content. Similarly, texts consisting only of characters/hashtags/@-mention chains were deleted. It is possible that misogynistic texts were deleted as a result, as hashtags can also contain misogynistic terms. The same applies to texts that were deleted because they contain only one word. During the dataset analysis, it was noticed that vulgar content is increasingly being used for sex advertising. By excluding texts containing the German words ”feucht” and ”nass” (both Engl. wet), ”porno” (Engl. porn) and the English words ”free” and ”xxx”, it was possible to filter out the vast majority of such content. Also disturbing content

Kappa Statistic	Strength of Agreement
< 0.00	Poor
$0.00 - 0.20$	Slight
$0.21 - 0.40$	Fair
$0.41 - 0.60$	Moderate
$0.61 - 0.80$	Substantial
$0.81 - 1.00$	Almost Perfect

Table 3: Interpretation of Kappa (Landis and Koch 1977).

was removed that offers no additional information content, such as links and control characters (except spaces). Any @-mentions still present have been replaced by the neutral formulation "@TwitterUser". This helps to avoid possible prejudices of the annotators against the persons mentioned and to preserve their privacy. Hashtags were also removed, as they often address individuals. The last step resulted in the text becoming difficult to understand in some cases. For example, the following tweet loses its context when the hashtag is removed:

> "@TwitterUser Lass dich von diesen #Rechtswixer n nicht beeinflussen."
>
> "@TwitterUser Don't let these #RightwingWanker s influence you."

Inter-rater Reliability

A common method to calculate the agreement of annotators during annotation is to let the same documents be evaluated by all annotators and then compare the results. The quality of the agreement can then be calculated using an inter-rater reliability.

A well known measure for the inter-rater reliability is Cohens Kappa (Cohen 1960), shown in Eq. (1). It computes the observed match between two annotators $\bar{P}$ as well as the probability for an agreement based on chance $\bar{P}_e$ and from this it calculates the agreement κ. The height of the kappa value is thereby a measure for the quality of the annotation. An interpretation of the kappa value by Landis and Koch (1977) is given in Table 3.

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{1}$$

However, Cohens Kappa has a major disadvantage. If there are three annotators instead of two, three kappa values for the three different combinations of the annotators can be calculated, however, a simple aggregation of these values to a single value is not possible. This drawback is solved by Fleiss Kappa (Fleiss 1971), which represents a single kappa value that can be calculated with two or more annotators.

As Cohens Kappa, Fleiss kappa is calculated with Eq. (1), whereby the probabilities are defined differently as illustrated by $\bar{P}$ in Eq. (2) and Eq. (3) as well as $\bar{P}_e$ in Eq. (4) and Eq. (5). Let N represent the number of subjects, which in our case is the number of posts, while n is the number of annotations per subject, and k is the number of categories. The subjects are indexed by $i = 1, 2, \ldots N$ and the categories by $j = 1, 2, \ldots k$. Hence, n_{ij} is the number of annotators who assigned the i-th post to the j-th class. In our case, the number of categories is $k = 2$ each for hatespeech and misogynistic hatespeech.

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i \tag{2}$$

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \tag{3}$$

$$\bar{P}_e = \sum_{j=1}^{k} {p_j}^2 \tag{4}$$

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \tag{5}$$

Details

The annotation of the corpus was performed according to the procedure previously outlined in Section Dataset. In the following Annotation we first describe the procedure in greater detail. Then, in Section Corpus Statistics, we present the results of the annotation.

Annotation

A total of six volunteers, two women and four men aged between 25 and 65, all of whom speak German as their mother tongue and have an academic background, assisted with the evaluation. The annotation process was carried out over a period of two and a half months (mid-February to April 2022).

During Phase 1, the volunteers met for a discussion and reviewed 11 selected posts, which were manually selected to ensure that the posts contain different aspects of hatespeech according to our definition of hatespeech from Section Classes to Annotate. For each post the volunteers read the text and the availability of hatespeech was rated manually as *hatespeech* or *not hatespeech* and either *misogynistic hatespeech* or *not misogynistic hatespeech*. If the volunteers came to different views at first on how to assess one of these two points, this point was discussed until a consensus was reached that everyone approved. This approach was intended to create a common understanding among the volunteers on how to assess these two aspects.

The following Phase 2 should determine whether the volunteers actually evaluate with the same standards. To validate this, each of the volunteers received the same randomly selected 46 posts, which they then annotated on their own according to the rules previously established during Phase 1. Afterwards the results of the individual volunteers were used to determine the uniformity of the annotation. Therefore an inter-rater reliability with Fleiss' Kappa was calculated. The agreement for all volunteers was $\kappa = 0.5904$ for hatespeech and $\kappa = 0.4349$ for misogynistic hatespeech, which both correspond to a moderate agreement according to interpretation of Landis and Koch (1977) mentioned above in Table 3. Due to this noticeably lower result for the misogynistic hatespeech, we examined the underlying causes. In order to determine the influence of every single volunteer, we

	Phase 2a		Phase 2b		Phase 2c	
w/wo	HS	MHS	HS	MHS	HS	MHS
All	0.5904	0.4349	0.7655	0.5939	0.6409	0.8258
1	0.5776	0.3794	0.7490	0.5034	0.6242	0.7909
2	0.5919	0.3794	0.7611	0.5700	0.6435	0.7909
3	**0.6182**	0.5704	0.7587	0.6286	0.6600	**0.9045**
4	0.5656	0.3205	0.7802	0.5034	**0.6618**	0.8844
5	0.5843	0.3205	0.7384	0.5700	0.6435	0.7909
6	0.6060	**0.6044**	**0.8054**	**0.7637**	0.6103	0.7909

HS = hatespeech, MHS = misogynistic hatespeech

Table 4: Inter-rater reliability with Fleiss' Kappa.

calculated additional values for all combinations of $n-1$ volunteers. The results are given in Table 4 in column *Phase 2a*. As the interpretation shows, for hatespeech only expert 3 has a slightly higher negative influence, since without him a κ of 0.6182 could be achieved. Overall, however, the values differed only slightly. In contrast, for misogynistic hatespeech, expert 6 was identified as the expert with a significantly lower level of agreement, since without him a κ of 0.6044 is possible.

In order to improve the uniformity of annotation, the volunteers reviewed the 46 posts annotated during the first round of Phase 2 and discussed the rules of annotation again. Then a second round of Phase 2 was conducted, in which the volunteers received another 43 posts for annotation. Table 4 presents the kappa values of the second round in column *Phase 2b*. As the results show, the agreement over all volunteers regarding hatespeech increased by 0.1751 to a substantial $\kappa = 0.7655$. But also the agreement concerning the misogynistic hatespeech had increased significantly by 0.1590 to a strong moderate $\kappa = 0.5939$. To further improve the agreement, the 43 posts from the second round were reviewed, before a third round of Phase 2 was conducted, where the volunteers annotated another 46 posts. The kappa values are shown in Table 4 in column *Phase 2c*. This corresponds to the best agreement for the misogynistic hatespeech that increased by 0.2319 to an almost perfect $\kappa = 0.8258$ and the second best for hatesepeech that slightly decreased by -0.1246 to $\kappa = 0.6409$. As we can observe, volunteer 4 has the strongest negative influence on hatespeech, while volunteer 3 has it for the misogynistic hastespeech. A comparable corpus of Schabus, Skowron, and Trapp (2017) comes to similar kappa values between 0.3 and 0.6.

During the final Phase 3, the six volunteers processed a further 7,061 randomly selected posts. Due to the solid inter-rater reliability, each post was annotated by only one volunteer instead of a multiple evaluation. Table 5 summarizes the details of the annotation for the phases, including the number of annotations, number of unique posts and average number annotations of each volunteer.

Corpus Statistics

During the three phases of the annotation, a total of 7,207 posts were annotated. Of these, 146 were processed during Phases 1 and 2, which only served to train the volunteers.

		Number of annotations	
Phase	Posts	Total	Per volunteer
1	11	11	-
2a	46	276	46
2b	43	258	43
2c	46	276	46
3	7,061	7,061	∅ 1,176

Table 5: Phases of the annotation process.

The annotations of these posts were created before or for the purpose of calculating the inter-rater reliability and therefore have no guaranteed quality. Phase 3 was the first phase in which 7,061 further posts were annotated, whose quality is assured, which is why they form the core of the corpus.

It might occur that a certain subject area is especially prevalent in the dataset. This can lead to a situation where classification models trained on this data perform poorly when applied on new data. One reason for this is that certain words correlate with individual classes, even though they are not representative of the respective class (Wiegand, Ruppenhofer, and Kleinbauer 2019). Table 6 provides the quota of the 7,061 posts assigned to each class. The distribution of hatespeech reveals that 22.29 % of the post were annotated as hatespeech. The table also shows the distribution of the second criterion misogynistic hatespeech, with 6.51 % of all posts are being rated as *misogynisitc hatespeech*. Consequently, 29.22 % of hatespeech posts are also misogynistic.

Better insight can be achieved by using *Pointwise Mutual Information (PMI)*, a metric that can be used to examine a dataset for existing topics and their influence on the classes within that dataset (Wiegand, Ruppenhofer, and Kleinbauer 2019). The PMI metric from Eq. (6) calculates which words appear most frequently in combination with each other. To do this, the probability that the two words occur in the same document $P(W_1, W_2)$ is divided by the product of their individual probabilities for occurrence $P(W_1)\ P(W_2)$.

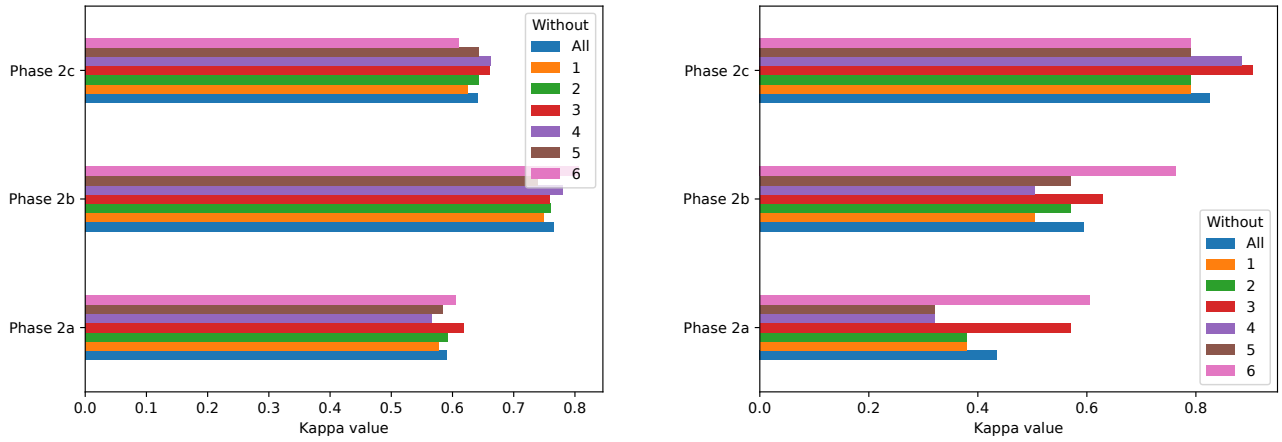
$$PMI(W_1, W_2) = log_2 \frac{P(W_1, W_2)}{P(W_1)P(W_2)} \tag{6}$$

The interpretation of the result can be illustrated using two examples:

1. W_1 = Butter (Engl. butter); W_2 = Brot (Engl. bread)
2. W_1 = Kanzler (Engl. Chancellor); W_2 = Scholz

In example 1, it is very likely that the words ("Butter"/"butter", "Brot"/"bread") often appear together. However, they are also used in many other word and sentence combinations (e.g. "Pizzabrot"/"Pizza bread", "Alles in Butter"/"Everything is fine"). The words Kanzler and Scholz, on the other hand, are also frequently used in the same sentence (e.g. "Olaf Scholz ist der aktuelle Kanzler von Deutschland"/"Olaf Scholz is the current Chancellor of Germany") but only rarely in a different context. If words only appear together, the following applies:

$$P(W_1, W_2) \approx P(W_1) \approx P(W_2) \tag{7}$$



(a) Hatespeech

(b) Misogynistic hatespeech

Figure 2: Phase 2a to 2c - Inter-rater reliability with Fleiss' Kappa per class for different annotator combinations.

Due to the calculation of the product in the denominator of PMI, it has to be expected that the PMI in example 1 is worse than the PMI in example 2. Unless we are dealing with a text corpus on German politics, the words from example 2 are rather rarely used words. However, they achieve a higher PMI value. To address the problem that rare words are overrated, the normalized version (NPMI) from Eq. (8) can be used instead (Bouma 2009), while the NPMI result have be interpreted following Eq. (9).

$$NPMI(W_1, W_2) = \frac{PMI(W_1, W_2)}{-log_2 \cdot P(W_1, W_2)} \tag{8}$$

$$NPMI = \begin{cases} 1 & \text{Terms only occur together} \\ 0 & \text{Normal distribution} \\ -1 & \text{Terms never appear together} \end{cases} \tag{9}$$

The NMPI metric can be used to reveal the co-occurrence of classes and specific words. The wordclouds shown in Fig. 3 reveal which words appear most frequently in the posts of each class. The more often a word occurs, the larger it is displayed in the wordcloud. In texts classified as neutral, meaning texts that do not contain hatespeech, very generally used, uncritical words appear most frequently (Fig. 3a). There are no specific trigger words or sensitive topics identified in these texts. In the case of hatespeech posts, the topic areas of the most frequently used words are clearly differently distributed. According to the depiction from Fig. 3b, besides attacks on personalities from politics, there are also dehumanizing events desired for specific individuals or groups (e.g., wishing for "Abschaum"/"lowlifes" to "verrecken"/"die"). In addition, crude insults such as "Hurensohn"/"son of a bitch" are frequently used. Expressions that are xenophobic or islamophobic such as "Neger"/"negroes" or "Musels"/"muzzies" are also included (Mandl et al. 2019, 2020). According to Fig. 3c, misogynistic hatespeech is usually recognized by (strongly) derogatory terms for women (e.g., "Weiber"/"broads", "Schlampen"/"sluts", "Göre"/"brat", "Fotz"/"cunt").

	Posts	%
Not hatespeech	5,487	77.71
Hatespeech	1,574	22.29
Not misogynistic hatespeech	6,601	93.49
Misogynistic hatespeech	460	6.51

Table 6: Number of posts per class in 7,061 posts.

Experiments

Based on the 7,061 annotated posts, we trained and evaluated first baseline models that classify new texts according to hate speech and misogynistic content.

This classification task is a multi-label classification characterized by the fact that each post is assigned to two out of four classes, which in our case represent *hatespeech*, *not hatespeech*, *misogynistic hatespeech* and *not misogynistic hatespeech*. A multi-label classifier would predict for a post whether it belongs to each one of the classes or not. We decided to simplify the problem by breaking it down into two binary classification problems. Therefore, we trained a separate binary classifier for hatespeech and misogynistic hatespeech, which predicts whether a post belongs to it or not.

In this study we used a pretrained German language model based on the BERT-Architecture called gbert-base (Chan, Schweter, and Möller 2020). Finetuning the complete model on the described classification task required one additional single linear layer with a Sigmoid activation function as classification head. It utilizes the [CLS] token from the BERT model's final layer, containing a representation of all tokens in a sentence. For model training, consistent hyperparameters were employed, setting the learning

(a) Neutral. (b) Hatespeech. (c) Misogynistic hatespeech.

Figure 3: Wordclouds.

rate to 1×10^{-5} and utilizing a batch size of 16. Each experiment underwent training for a maximum of 8 epochs. In instances where a model converged earlier and demonstrated superior performance at a prior checkpoint, that specific checkpoint was evaluated. Given the imbalance in the dataset concerning the number of hatespeech comments, we additionally experimented with undersampling the majority class. We reduced the number of neutral texts in a ratio of 5:1 to the number of texts containing misogynistic hatespeech. All experiments were conducted in the Google Colab Cloud on a single Nvidia T4 GPU with 16 GB RAM, enabling reproducibility on standard personal hardware. Consequently, two distinct models were trained for the classification of hatespeech (*Hatespeech* & *Hatespeech Undersampling*) and misogynistic hatespeech (*Misogyny* & *Misogyny Undersampling*).

For the evaluation of the algorithms, the corpus of 7,061 posts was split by fractions of 0.8 into a training set and by 0.1 each into a validation and test set. The training set was then used to train each of the classifiers. Finally the classifiers were evaluated with the test set. To evaluate the predictions, the metrics Precision, Recall and F_1 were used.

The results of the baseline classification are given in Table 7, with best value of each metric depicted in bold. For each trained model, values for the neutral class were also presented. As the data show, the classes are predicted with varying quality. The tendency can be seen that the height of the class distribution, shown in Table 6, has a major impact on the classification quality. For hatespeech we achieved a F_1 value of 0.68 in contrary to 0.54 for the misogyny classifier. Looking on the results, it is clear that undersampling only lead to minor improvements for selected metrics in both of the classification tasks.

Moreover, in Fig. 4, the confusion matrices transparently present the test results for the models. Of particular interest in the classification is the reduction of false negatives, ensuring the detection of all instances of hatespeech. This can be shown through the metric Recall. While undersampling led to a performance decrease in hatespeech classification, a notable improvement with a result of 0.63 was achieved in the detection of misogynistic hatespeech. However, the improvement had a significant impact on the number of false positives, which are directly influencing the metric Precision (0.47). In order to improve the result, the model has to be optimized to better detect misogyny. Fig. 5 presents the ROC curve for the test results, demonstrating the relationship between true positive rate (TPR) and false positive rate (FPR) and underscoring the importance of a steep ascent in the curve (high TPR with low FPR). By calculating the area under the curve (AUC), the results can be made measurable, with model *Misogyny Undersampling* achieving the best value of 0.93.

Model	Metr.	Classification					
		Hatespeech			Misogynist. Hatespeech		
		0	1	mac. avg.	0	1	mac. avg.
GBERT-base	Prec.	**0.92**	0.64	**0.78**	**0.97**	**0.50**	**0.74**
	Rec.	0.88	**0.73**	**0.81**	**0.96**	0.59	0.77
	F_1	**0.90**	**0.68**	**0.79**	**0.96**	**0.54**	**0.75**
GBERT-base undersampling	Prec.	0.91	**0.65**	**0.78**	**0.97**	0.47	0.72
	Rec.	**0.89**	0.71	0.80	0.95	**0.63**	**0.79**
	F_1	**0.90**	**0.68**	**0.79**	**0.96**	**0.54**	**0.75**

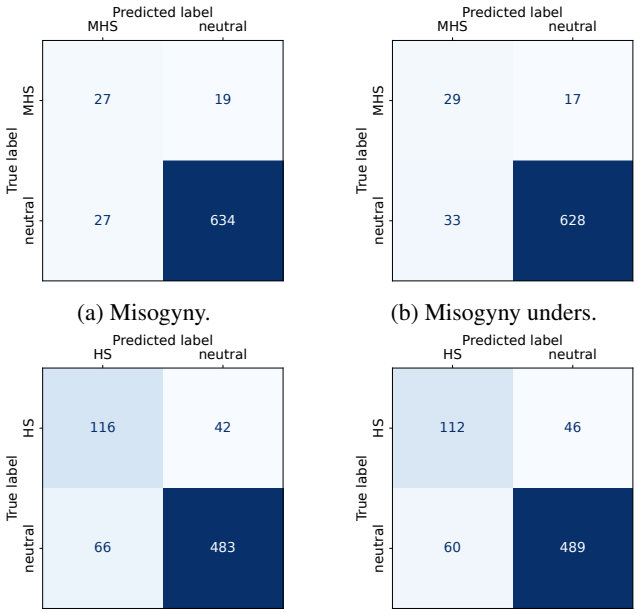
Table 7: Results of the baseline classification.

Discussion

In the study, it became evident how challenging it is to handle hatespeech. The primary difficulty arises in defining and delineating the properties of hatespeech. This definition can vary significantly depending on cultural nuances, prevailing attitudes, and the specific context. To provide a clearer understanding of the dataset's focus and utility, we extensively explained our definition of hatespeech.

Due to the increasing size of large language models (LLMs), the way corpora are used for fine-tuning may undergo significant change. These models encapsulate such vast amounts of knowledge that retraining becomes only partially necessary. However, in the present scenario, linguistic nuances can still determine how a text is classified. Since the definition of hatespeech can vary significantly, a certain form of learning or fine-tuning remains inevitable at this point in time. In addition, high-quality annotated datasets help to analyze and challenge downstream text generations of LLMs on the prevalence of hatespeech.

The limited number of annotated posts can lead to certain topics and classes being over- or underrepresented, which could introduce biases in the later stages of classification. It is important to consider and address this factor when utiliz-



(a) Misogyny.

(b) Misogyny unders.

(c) Hatespeech.

(d) Hatespeech unders.

Figure 4: Classification results.

ing the dataset for productive use in hatespeech detection.

It turned out that the detection of hate speech is a classification problem with an increased error rate. Dealing with false-positive and false-negative results can't be handled universally. Instead, an individual assessment is required for the specific application.

Ethical Considerations

Our work contributes to the training of AI to detect unethical behavior in the form of hatespeech. The privacy of all authors of tweets is preserved as all occurrences of usernames in the texts are anonymized. The tweet ID has been retained to prove the authenticity of the data.

In addition, the identities of the persons involved in the annotation were pseudonymized. The participants had the option to cancel participation or refuse to annotate posts at any time. Since the dataset is intended to detect hatespeech, it inherently and inevitably contains offensive content.

In Section Discussion, we discuss the problem that we cannot ensure that the minority classes are free of bias. This could lead to an increase in the number of false positives based on certain topics. If harmless posts are misidentified as hate speech, freedom of expression is at risk. On the other hand, if hateful posts are published, individuals may be harmed.

Similar to any positive application, the present dataset could be exploited for negative purposes, e.g. to build a system generating hatespeech texts, like spam bots.

How to Use the Corpus?

We provide the corpus presented in this paper, consisting of the texts and the annotations, to the scientific community

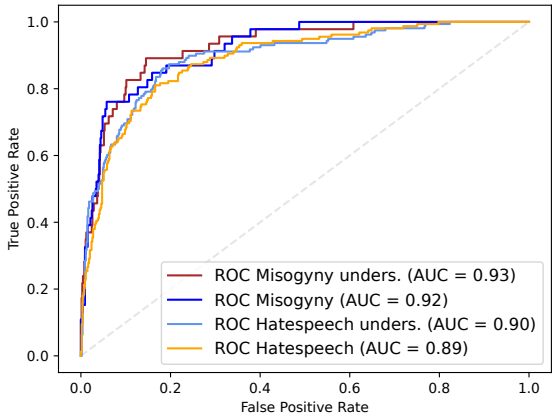


Figure 5: ROC curve.

on Zenodo (https://doi.org/10.5281/zenodo.10513452) and a website (https://ccwi.github.io/corpus-gmhp7k/), licensed under Creative Commons Attribution 4.0 International.

Conclusion

In this work a corpus was presented which consists of 7,061 posts in German language that belong to two social media platforms. The posts were annotated by volunteers according to hatespeech and misogynistic hatespeech, achieving a solid degree of agreement, as an inter-rater reliability according to Fleiss' Kappa revealed, which was 0.6409 for hatespeech and 0.8258 for misogynistic hatespeech. To evaluate the corpus, a first baseline text classification was presented, where the text of each post was used to predict hatespeech and misogynistic hatespeech. To make the corpus also usable for other applications in NLP and classification, we provide it as a dataset on German Misogynistic Hatespeech Posts (GMHP7k) to the scientific community.

References

Assenmacher, D.; Niemann, M.; Müller, K.; Seiler, M.; Riehle, D.; Trautmann, H.; and Trautmann, H. 2021. RP-Mod & RP-Crowd: Moderator- and Crowd-Annotated German News Comment Datasets. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Bouma, G. J. 2009. Normalized (pointwise) mutual information in collocation extraction. In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, 31–40.

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 (2017)*, Proceedings of the Annual Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences.

Chan, B.; Schweter, S.; and Möller, T. 2020. German's Next Language Model. ArXiv:2010.10906 [cs].

Charitidis, P.; Doropoulos, S.; Vologiannidis, S.; Papastergiou, I.; and Karakeva, S. 2020. Towards countering hate speech against journalists on social media. *Online Social Networks and Media*, 17: 100071.

Cieliebak, M.; Deriu, J. M.; Egger, D.; and Uzdilli, F. 2017. A Twitter Corpus and Benchmark Resources for German Sentiment Analysis. In Ku, L.-W.; and Li, C.-T., eds., *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 45–51. Association for Computational Linguistics.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1): 37–46.

Demus, C.; Pitz, J.; Schütz, M.; Probol, N.; Siegel, M.; and Labudde, D. 2022. A Comprehensive Dataset for German Offensive Language and Conversation Analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, 143–153. Seattle, Washington (Hybrid): Association for Computational Linguistics.

ECRI. 2016. ECRI General Policy Recommendation No.15 on combating hate speech.

Fersini, E.; Nozza, D.; and Rosso, P. 2018. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12: 59.

Fersini, E.; Rosso, P.; and Anzovino, M. E. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *IberEval@SEPLN*.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378–382.

FORCE11. 2020. The FAIR Data principles.

Founta, A.-M.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. ArXiv:1802.00393 [cs].

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92. Publisher: ACM New York, NY, USA.

Ging, D.; and Siapera, E. 2018. Special issue on online misogyny. *Feminist Media Studies*, 18(4): 515–524.

Gomez, R.; Gibert, J.; Gomez, L.; and Karatzas, D. 2019. Exploring Hate Speech Detection in Multimodal Publications. ArXiv:1910.03814 [cs].

Keller, M.-E.; Döschl, A.; and Mandl, P. 2023. AMPEL: An Approach for Machine-learning Based Prediction and Evaluation of the Learned Success of Social Media Posts. In *2023 33rd Conference of Open Innovations Association (FRUCT)*, 139–147. IEEE. ISBN 978-952-69244-9-6.

Keller, M.-E.; Forster, J.; Mandl, P.; Aich, F.; and Althaller, J. 2019. A German Corpus on Topic Classification and Success of Social Media Posts. In *Proceedings of the 25th Conference of Open Innovations Association FRUCT*, FRUCT'25. Helsinki, Finland: FRUCT Oy.

Keller, M.-E.; Stoffelen, B.; Kailer, D.; Mandl, P.; and Althaller, J. 2018. Predicting the success of posts for brand pages on Facebook. In *Proceedings of the 17th International Conference WWW/Internet 2018, Budapest, Hungary: IADIS*.

Landis, J. R.; and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1): 159.

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*, 29–32. Hyderabad India: ACM. ISBN 978-1-4503-8978-5.

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*, 14–17. Kolkata India: ACM. ISBN 978-1-4503-7750-8.

Meta. 2024. Hassrede | Transparency Center.

Narr, S.; Hülfenhaus, M.; and Albayrak, S. 2012. *Language-Independent Twitter Sentiment Analysis*. DAI-Labor, Technical University Berlin, Germany.

Pavlopoulos, J.; Sorensen, J.; Dixon, L.; Thain, N.; and Androutsopoulos, I. 2020. Toxicity Detection: Does Context Really Matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4296–4305. Online: Association for Computational Linguistics.

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. *NLP4CMC III: 3rdWorkshop on Natural Language Processing for Computer-Mediated Communication*. Publisher: DuEPublico: Duisburg-Essen Publications Online, University of Duisburg-Essen, Germany.

Schabus, D.; Skowron, M.; and Trapp, M. 2017. One Million Posts. In Kando, N.; Sakai, T.; Joho, H.; Li, H.; de Vries, A. P.; and White, R. W., eds., *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '17*, 1241–1244. ACM Press. ISBN 978-1-4503-5022-8.

Sharma, S.; Agrawal, S.; and Shrivastava, M. 2018. Degree based Classification of Harmful Speech using Twitter Data. ArXiv:1806.04197 [cs].

Shushkevich, E.; and Cardiff, J. 2019. Automatic Misogyny Detection in Social Media: A Survey. *Computación y Sistemas*, 23(4).

Sponholz, L. 2018. *Hate Speech in den Massenmedien*. Wiesbaden: Springer Fachmedien Wiesbaden. ISBN 978-3-658-15076-1 978-3-658-15077-8.

Vidgen, B.; and Derczynski, L. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS one*, 15(12): e0243300.

Waseem, Z. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In Bamman, D.; Doğruöz, A. S.; Eisenstein, J.; Hovy, D.; Jurgens, D.; O'Connor, B.; Oh, A.; Tsur, O.; and Volkova, S., eds., *Proceedings of the First Workshop on NLP and Computational Social Science*, 138–142. Association for Computational Linguistics.

Wiegand, M.; Ruppenhofer, J.; and Kleinbauer, T. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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)*, 602–608. Minneapolis, Minnesota: Association for Computational Linguistics.

Wiegand, M.; Siegel, M.; and Ruppenhofer, J. 2019. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In Ruppenhofer, J.; Siegel, M.; and Wiegand, M., eds., *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria – September 21, 2018*, 1 – 10. Vienna, Austria: Austrian Academy of Sciences. ISBN 978-3-7001-8435-5.

Paper Checklist

1. For most authors
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. Our work contributes to the training of AI to detect unethical behavior and hatespeech. Privacy of all authors of Tweets will be preserved as all occurrences of names in the texts are anonymized.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, the abstract summarizes the results of our work. Furthermore, the end of the introduction contains a list of the main contributions with references to individual sections.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, we have tried to justify all the methodologies used, e.g. we describe how we aim to ensure consistent annotations by calculating the inter-rater reliability after each iteration in phase 2 (Annotation).
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, in Corpus Statistics we analyze the correlation between the frequency of certain words and the assigned category in the dataset.
   (e) Did you describe the limitations of your work? Yes, in Discussion we describe our considerations regarding the multi-layered nature of hatespeech and the inherent complexity of automatically classifying hatespeech. We also consider that the dataset is not free from bias for certain topics, leading to a limited generalization.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, in Discussion we discuss the problem that we can't assure that the minority classes are bias free. This might result in an increase of false positives only because texts include certain topics.
   (g) Did you discuss any potential misuse of your work? Yes, statements that are criminally relevant under German law may be covered by freedom of expression under other laws. We have therefore explained our definition of hatespeech in detail (Classes to Annotate). Furthermore, examples of the negative class could be misused to set up a system for generating hatespeech texts, e.g. with spam bots (Discussion).
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. We documented the schema of the data, anonymized it and release it responsibly accompanied by detailed description in our paper. To allow reproducibility of our findings, we document the train-, test- and validation split and type as well as base model of our classifier. While the texts of the tweets were anonymized, the tweet ID was preserved to prove the authenticity of the data.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes. We have read the guidelines carefully and believe that we fulfill all points without any restrictions.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, the published dataset contains a column, with information about the split used in the machine learning experiments. Additionally, in Experiments we describe the models architecture and hyperparameters. If further details are necessary, we can share the code for training and evaluation of the model upon request. We also provide code for descriptive statistics on the dataset (Corpus Statistics).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes. We clearly describe the most important hyperparameters, including learning-rate, batch-size (Experiments). The data splits are clearly specified in a dedicated column in the provided dataset.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No. We believe that it is not necessary to specify an error, as we achieve deterministic and therefore reproducible results by specifying a single fixed seed. Since the classification primarily serves to prove the quality of the data, the focus is on the data and not not on creating the best possible model.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, in Experiments we mentions that we trained the models with an Nvidia T4 GPU with 16GB RAM. All the experiments were executed in a Google Colab Cloud environment

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, fine tuning a general BERT model on a specific task is a common method to reach first competitive results without high computational costs. Since the classes in the dataset are unevenly distributed, we decided to list the results for each class separately so that the model can be better evaluated in relation to the minority class.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, in Experiments and Discussion we discuss the trade off between a low number for false negatives and a high number of false positives.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes. Our dataset consists primarily of our own data, but was supplemented by data from existing publicly available datasets, whereby the works were cited (Data Collection)

(b) Did you mention the license of the assets? Yes, we mention the licenses with the description and citation of the datasets in Data Collection.

(c) Did you include any new assets in the supplemental material or as a URL? Yes. The main component of the dataset provided is our own new data.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, both the dataset from Charitidis et al. (2020) and the two datasets from Mandl et al. (2019) and Mandl et al. (2020) have been published for scientific purposes. By referencing their publications, we acknowledge their contribution to the creation of the datasets.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes. We have tried to anonymize the personal data to the best of our knowledge. In addition, the identities of the persons involved in the annotation were pseudonymized using unique numbers. Since the dataset is intended to detect hatespeech, it inherently and inevitably contains offensive content.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? No, we haven't created a datasheet so far.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? No, all instructions for the annotation process were communicated verbally to the participants. However, the agreed rules were described in detail in the explanations of the annotation process (Classes to Annotate).

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? No, the participants were a small group of six voluntary annotators who decided to participate after a joint discussion. There was also the option to cancel participation or refuse to annotate posts at any time.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? No, all participants took part in the annotation voluntarily and without payment.

(d) Did you discuss how data is stored, shared, and deidentified? Yes. We have tried to anonymize the personal data to the best of our knowledge. In addition, the identities of the persons involved in the annotation were pseudonymized using unique numbers.