Datenbank Spektrum (2023) 23:41–51
https://doi.org/10.1007/s13222-023-00439-0

SCHWERPUNKTBEITRAG

Bias Mitigation for Capturing Potentially Illegal Hate Speech

Johannes Schäfer[1]

Received: 22 October 2022 / Accepted: 11 February 2023 / Published online: 29 March 2023

Abstract
Hate speech is a persistent issue in social media. Researchers have analyzed and developed detection methods for hate speech on the basis of example data, even though the phenomenon is only rather vaguely defined. This paper provides an approach to identify hate speech in terms of German laws, which are used as a basis for annotation guidelines applied to real world data. We annotate six labels in a corpus of 1,385 German short text messages: four subcategories of illegal hate speech, offensive language and a neutral class. We consider hate speech expressions as illegal if the linguistic content could be interpreted in a given context possibly violating a specific law. This interpretation and a check by lawyers would be the next step which is not yet included in our annotation. In this paper, I report on strategies to avoid certain biases in data for illegal hate speech. These strategies may serve as a model for building a larger dataset. In experiments, I investigate the capability of a Transformer-based neural network model to learn our classification. The results show that this multiclass classification is still difficult to learn, probably due to the small size of the dataset. I suggest that it is crucial to be aware of data biases and to apply bias mitigation techniques when training hate speech detection systems on such data. Data and scripts of the experiments are made publicly available.

Keywords Hate speech detection · Hate speech data · Illegal hate speech · Identity term bias

1 Introduction

Online social media services are a major device for information and communication in the developed world, indispensable in everyday life. Their interactive design and strong incentive for user participation as well as limited regulation make them vulnerable for misuse. This can manifest itself in the form of hate speech with the intent to harm people. For example, the EU Code of conduct on countering illegal hate speech online[1] considers illegal hate speech as "all conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin." While this definition states specific groups which can be targets of hate speech, it does not state equally clearly when a message counts as hateful – which leaves room for interpretation.

In previous work [19], we defined illegal hate speech with four different subcategories on the basis of German laws and developed annotation guidelines. We consider a message to contain illegal hate speech if the linguistic content (such as offensive language) could be interpreted in a certain context as possibly violating specific laws, most of which concern public messages directed against groups of people or individuals. We do not distinguish illegal hate speech from (potentially legal) hate speech as we do not assess the severity of a possible offense. However, we differentiate between illegal hate speech, undirected offensive language (e.g. use of swearwords) and neutral comments. Thus, we consider in total six annotation labels.

The focus of the present work lies on our methods for data acquisition and on the annotation process for illegal hate speech. I report on our strategies to avoid bias in the data as well as on the novel annotation of identity term mentions.

[1] https://ec.europa.eu/newsroom/just/document.cfm?doc_id=42985.

This paper contains examples of offensive language. Data and code: https://github.com/Johannes-Schaefer/ihs.

✉ Johannes Schäfer
johannes.schaefer@uni-hildesheim.de

[1] Institute for Information Science and Natural Language Processing, University of Hildesheim, Hildesheim, Germany

1.1 Related Work

In research, broad offensive phenomena are often analyzed and subcategorized, which include more general forms of hate speech. For example, Ruppenhofer et al. [18] consider subcategories of offensive language as *Profanity*, *Insult* and *Abuse*. The focus in their work lies on the linguistic content of messages, thus, their definition is substantiated by an annotated dataset of example messages. This is a common research practice where, in general, the task of hate speech detection is considered as a short text classification task for a concrete dataset. Methods are then to be developed on such data that predict whether an unseen, new message contains hate speech or not, and in particular which subcategory of hate speech a given short text can be ascribed to. A survey on general approaches for hate speech detection is given by Schmidt and Wiegand [20]. Recent approaches are mostly based on models using the Transformer architecture [22]. For example, the pre-trained language model BERT [8] has proven to show promising results when being applied to hate speech detection [13, 16, 17, 23].

The detection of hate speech or of similar phenomena has been extensively explored in several shared tasks, e.g. in the following: EVALITA 2018 Hate Speech Detection Task [3], GermEval 2018 Shared Task on the Identification of Offensive Language [24], SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval) [27], SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter [1], GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language [21], HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages [14], HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages [15]. The usual approach is to focus on the binary case, i.e. to distinguish hate speech from acceptable content. Therefore, the definition of hate speech has to cover a wide range of variants of the phenomenon and remains rather vague. In work by Mandl et al. [15] the annotation almost entirely relies on the subjective assessments of the annotators who are only provided with a very vague conceptual definition of hate speech. This, however, results in categories which are often very difficult to interpret, which in turn makes the differences of the performance of detection systems on such data difficult to interpret. Thus, the explainability of a systems' performance is limited and a qualitative evaluation usually does not reveal clear patterns.

1.2 Our Method

In order to reduce the vagueness of hate speech definitions, we proposed in previous work [19] an approach to define more well-founded subcategories of illegal hate speech and to develop a dataset on the basis of applicable German laws. At the same time in the project DeTox [7], Demus et al. also consider the criminal relevance of comments amongst other categories. In the present paper, I focus on the process of creating our German dataset for illegal hate speech as well as on different bias avoidance steps we took into account.

The description of hate speech in the EU code of conduct mentioned above already points at certain groups of people who can be the target of hate speech. These groups are often explicitly mentioned in the messages and can be detected on a lexical level as shown by ElSherief et al. [9]. Based on how data have been sampled, different biases can be observed in the resulting dataset, such as topic bias or identity term bias [25]. Researchers have investigated components of this issue, for example racial bias [6]. We followed different techniques to avoid bias or at least to measure it in our dataset for illegal hate speech as presented in the following.

The remainder of this paper is structured as follows. In Sect. 2, a dataset for illegal hate speech is presented for which we gathered examples and annotated them on the basis of guidelines developed in earlier work [19]. In this section, I also discuss the applied bias avoidance techniques and give results from the annotation of illegal hate speech categories and identity term mentions. Subsequently, Sect. 3 reports on experiments to learn the annotated categorization using a Transformer-based neural network model. Sect. 4 discusses the results of the experiments and serves to asses the value of the developed dataset. I conclude in Sect. 5 with prospects for future endeavors to capture illegal hate speech online.

2 Data for Illegal Hate Speech

This section introduces our dataset for **i**llegal **h**ate **s**peech (iHS) which consists of 1,385 instances of German short text messages from Twitter. The data is annotated to distinguish four subcategories of illegal hate speech from offensive language and from a neutral class. Additionally, mentions of identity terms are marked for each message. The corpus data is stored in an XML format to allow several different types of annotations on message level and to maintain efficient manual editing and computational processing. In this section, I present how we created this dataset and discuss techniques to avoid biases.

In order to describe the phenomenon of illegal hate speech in a manner as clearly and concisely as possible, we decided to base our annotation guidelines on German laws and court rulings related to hate speech. We describe the steps of our method in more detail in our technical report [19].

We noticed that the following laws from the German criminal code ("Strafgesetzbuch (StGB)") are regularly applied in cases that deal with hate speech:

- § 111 StGB Public incitement to commit offenses ("Öffentliche Aufforderung zu Straftaten"),
- § 130 StGB Incitement of masses ("Volksverhetzung"),
- § 185 StGB Insult ("Beleidigung"),
- § 186 StGB Malicious gossip ("üble Nachrede"),
- § 187 StGB Defamation ("Verleumdung"),
- § 241 StGB Threatening commission of serious criminal offense ("Bedrohung").

While transcripts of verdicts sometimes contain examples of published hateful messages, the verdict itself usually never relies on single posts. Such posts are instead evaluated against a broader background. Thus, it proved impossible to create an annotated corpus solely on such examples. A better option was to use commonly applied laws to develop guidelines for patterns of expressions which could be interpreted as illegal and to annotate social media data based on those.

It should also be pointed out that this annotation of illegality is limited. The respective messages in our dataset are not necessarily illegal as this can only be decided by a court. We only consider them to be potentially illegal. They could be interpreted as illegal in a certain context – which we do not have access to. The annotation is based on patterns in the linguistic content only, with as little as possible interpretation, e.g. we do not detect irony.

Each Twitter message in the iHS dataset has been annotated by one trained annotator.[2] In the remainder of this section, I discuss our methods to gather data for this corpus and describe the annotation process.

2.1 Data Acquisition

Empirically collecting hate speech data is not a straightforward process, due to certain peculiarities of the phenomenon. If one simply collected a random sample of social media data, only a very small proportion of the messages would actually contain hate speech. The annotation would be unfeasible as the effort would be too high to produce a considerable number of positive examples. Instead, there are different approaches to automatically filter data first. Here, it is crucial to choose the method carefully in order not to introduce too much bias in the data.

Wiegand et al. [24] gather data by using seed accounts of users who have been observed to post hate speech. They widen their search given these accounts and manage to find more examples of hate speech with sufficient precision. A desired advantage of this approach is that it does not impose limitations on the lexical content which is collected. However, for our data we decided not to follow this method as it is difficult to avoid an author bias here and to avoid getting stuck in echo chambers of specific user groups when collecting data.

Table 1 Most frequent keywords in the iHS dataset with the percentage of the respective instances annotated to contain illegal hate speech (% HS)

Keyword	# instances	% HS
"dumm" (*stupid*)	71	41
"Sau" (*pig*)	67	18
"übel" (*foul*)	52	10
"Problem" (*problem*)	48	15
"Müll" (*trash*)	40	12
"Frauen" (*women*)	40	23
"fuck"	36	19
"behindert" (*disabled*)	35	23
"Partei" (*party*)	35	40

Instead, we find examples using a set of keywords[3]. We use Twitter as a source due to the easy accessibility of their API and the popularity of their platform. Social media platforms such as Twitter do filter for illegal content themselves. Nevertheless, we assume that we find a sufficient number of relevant cases, for the following reasons: When querying the API, we explicitly include recently published posts which presumably have not been checked by the platform. While some of the instances considered for our data might since have been reported by other users and have been deleted due to being hate speech, they are still in our corpus. Additionally, we also consider less severe cases for our categories which show patterns similar to illegal examples. However, these cases are possibly not excluded by the stricter filter which has been used by the platform.

A keyword-based collection method naturally has the effect that collected messages always contain at least one of the keywords and thus are biased towards these keywords. In order to mitigate this issue, we include words from several different categories (in total 82 keywords) to get some variety in the results. We choose words which can also be used in neutral contexts like certain swear words, e.g. *pig* ("Schwein"); words referring to frequently mentioned groups, e.g. *police* ("Polizei") or certain action verbs, e.g. *hit* ("schlagen"). We investigated the results and found that there is an intended bias towards hate speech in the data – we have a higher precision than random sampling. However, the results still contain a substantial amount of neutral

[2] Training of the annotator consisted of intensive study of the annotation guidelines with multiple rounds of discussions on the basis of annotated examples and unclear cases. However, no lawyers were involved in the process.

[3] The full list of our keywords is available online at https://github.com/Johannes-Schaefer/ihs/blob/main/keywords.

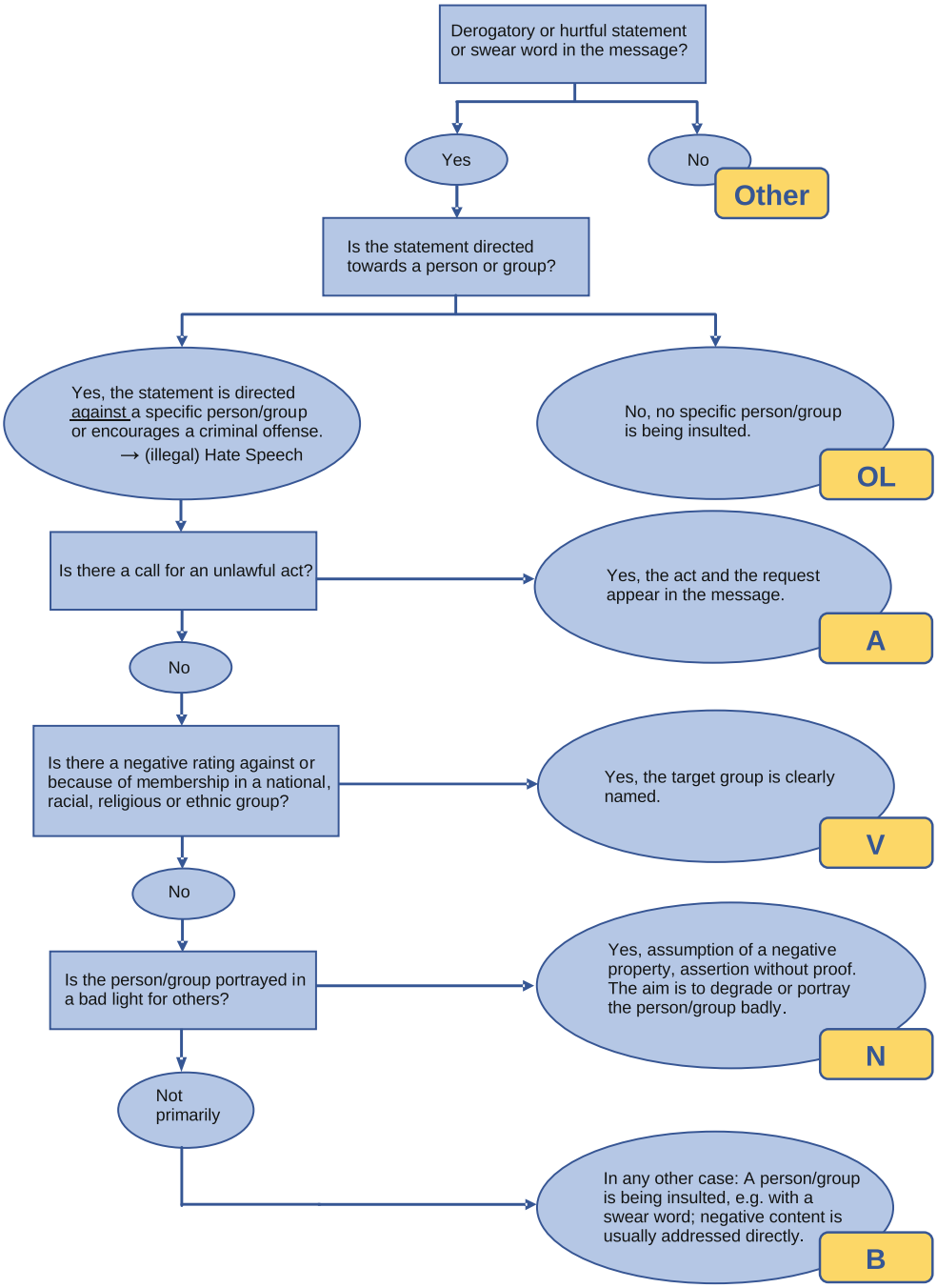

Fig. 1 Decision tree for the annotation of illegal hate speech [19]

messages. Table 1 shows the top keywords which are mentioned most frequently in instances in the iHS dataset with the percentage of the respective instances which are annotated to contain illegal hate speech. For all of these keywords the resulting instances are found to contain illegal hate speech in substantially fewer than 50% of the cases. Thus, a detection system cannot necessarily infer from the presence of a keyword that a message contains hate speech.

Another form of data bias can be a specific topic of the messages. This can occur, for example, when data is collected in a short time period after popular public events. In such cases many social media posts will discuss the popular topic, thus resulting in a dataset being biased towards it. To avoid this issue, we run queries in random intervals over a larger time period from 05/05/2021 to 11/19/2021. We assume this method covers a wide set of topics in posts by different user types. We did not conduct a comprehensive analysis of the topics discussed in our data, however, after screening a sample of the data, we are convinced that there is no strong topic bias. In total, we collected 274,347 Twitter messages using our keyword queries. We randomly sampled instances from the collected messages and anno-

tated 1,385 cases. The resulting iHS dataset comprises all these annotated instances.

2.2 Annotation of Illegal Hate Speech

The text messages contained in the iHS dataset are annotated with six different labels: illegal hate speech categories (four subclasses) vs. offensive language vs. a neutral class.

The annotation guidelines for illegal hate speech [19] were developed on the basis of the abovementioned six laws, but the annotation categories distinguish only four types of illegal hate speech (*A*, *V*, *N*, *B*). Here, § 241 StGB Threatening commission of serious criminal offense ("Bedrohung") was disregarded as in public Twitter messages it is an extremely rare phenomenon and we did not manage to gather a substantial amount of examples where this law might be applicable. Additionally, we combined § 186 StGB Malicious gossip ("üble Nachrede") and § 187 StGB Defamation ("Verleumdung") into a single category (*N*) as an assignment to one of these cannot be decided based on the linguistic content only. The label *A* marks cases related to § 111 StGB Public incitement to commit offenses ("Öffentliche Aufforderung zu Straftaten"), Label *V* marks cases related to § 130 StGB Incitement of masses ("Volksverhetzung") and the label *B* marks cases related to § 185 StGB Insult ("Beleidigung").

We differentiate illegal hate speech from (untargeted) offensive language (*OL*) and neutral content (*Other*). While hate speech itself often contains offensive words, we exclude cases where the offense is not targeted towards a certain individual or group of persons and annotate such cases as offensive language. We believe that it is important to distinguish such cases since this difference is crucial for assessing whether a message is illegal or not. Furthermore, we annotate offenses merely directed at non-human things or swearing as *OL*, to differentiate this category from neutral messages. Social media platform operators might consider such cases of offensive language as unacceptable content.

We performed two early annotation experiments to refine the guidelines [19]. A first annotation experiment employed 39 students as annotators which had no prior experience with research on hate speech. However, the quality of the resulting annotation is insufficient as the annotators did not strictly follow the guidelines (average Fleiss' kappa [10] of approximately 0.38, considered as fair agreement according to Landis and Koch [12]). In a second experiment with five annotators (experts of research on hate speech) on a dataset of 100 Tweets, we achieved a moderate inter-annotator agreement (Fleiss' kappa [10] of 0.50, interpretation according to Landis and Koch [12]). The highest Cohen's kappa [5] of 0.59 was achieved between the two annotators with the most experience with the guidelines (the authors of the guidelines). We present further details of these experiments in our previous report [19].

Table 2 Number of annotated German short text messages in our illegal hate speech dataset

Illegal hate speech	Label	# examples
+	*A*	16
+	*V*	11
+	*N*	106
+	*B*	178
-	*OL*	224
-	*Other*	850
	Total	1,385

After discussions of deviating cases and of the annotators' feedback from using the guidelines, we revised the annotation guidelines to be more precise. As a result, we also developed a decision support schema for annotation (see Fig. 1). This schema is modeled as a decision tree and meant to serve as a guiding tool during annotation. An annotator should first decide whether a specific message contains any offensive content, and second, if the offense is targeted at an individual or a group. These two steps cover the necessary decisions to distinguish the labels *Other* and *OL* from illegal hate speech. Subsequently, if the instance does contain a targeted offense, the annotator should evaluate other features which are significant for the four illegal hate speech categories to conclude the annotation of the given message.

Given the refined annotation guidelines with detailed label descriptions and example cases, we assume that even a layperson in law can be trained to annotate cases of the mentioned concept of illegal hate speech. However, the actual illegality of individual cases can only be interpreted by lawyers taking all factors from contexts into account. Therefore, we limit our analysis to an annotation of cases which can potentially be illegal based on similarities of patterns of the linguistic content.

We annotated the iHS dataset using the described schema (one trained annotator). The resulting class distribution for the different labels is shown in Table 2. Out of the total 1,385 instances, almost two thirds (850 instances) belong to the neutral *Other* class. These are counter-examples and neither contain hate speech nor offensive language. Approximately 16% (224 instances) of the messages are to be considered as offensive language and 22% (311 instances) are annotated to fall into one of the four illegal hate speech categories. However, the distribution of the four illegal hate speech categories *A*, *V*, *N* and *B*, is imbalanced. The labels *A* and *V* are annotated only 16 and 11 times, respectively, whereas the labels *N* and *B* are considerably more frequent as they are annotated 106 and 178 times. The labels in Table 2 are not sorted by their frequency in the data but

Table 3 Examples of annotated identity term mentions from the iHS corpus

Example message	Identity terms
"@mariamdbs @AhadunAhad02 wurde in der 11. auch mal aus der Klasse geschmissen weil meine pgw Lehrerin so aus dem nichts meinte dass Frauen im Islam weniger wert sind und ich versucht habe sie aufzuklären. Ging bisschen in die Hose hahaha" (*@mariamdbs @AhadunAhad02 was also kicked out of the class in the 11th grade because my pgw teacher said out of nowhere that women are less valuable in Islam and I tried to enlighten her. Was a bit of a flop hahaha*)	"@mariamdbs", "@AhadunAhad02", "Lehrerin" (*teacher*), "Frauen" (*women*), "Islam"
"@Speckshidoo Warum bist du so rechtsradikal speck" (*@Speckshidoo Why are you so radical right-wing speck*)	"@Speckshidoo", "rechts" (*right-wing*), "speck"

rather in decreasing order by our estimation of the severity of the respective offense. We can thus conclude that the more severe cases of illegal hate speech are more rarely to be found in social media.

2.3 Annotation of Identity Term Mentions

In addition to the annotation of the hate speech labels, identity term mentions are also annotated for each message. We consider identity term mentions often to be separable from expressions which constitute hate speech in a message. The decision whether a message is to be considered as hate speech or not should be in general made independently of the specific target (group) identity. An incitement of masses should not change its status as being hate speech if the mentioned, e.g. religious, group is replaced by another one.

For annotating identity terms in our data, no specific guideline document has been created. Instead, we discussed a possible definition of what counts as an identity term. We deliberately decided against annotating only a predefined set of entities as we intend to include all terms that mention any identities. We empirically gather identity terms on data and only then intend to group them into categories. In recent publications, the focus often lies on selected groups of identity terms. For example, Davidson et al. [6] investigate racial bias by considering the categories *black* and *white*. A larger set of different identity terms consisting of 24 discrete classes is annotated in the Civil Comments dataset [2]. However, a general analysis of identity term mentions in such data is limited by definition as they disregard all other identity term mentions which do not fall into the predefined categories. In contrast, we annotate strings from the messages directly and do not rely on a specific set of identity term categories, which allows for a general analysis.

In this annotation, we only consider expressions which are used to specify the identity of an individual or a group of people. We do not consider named entities referring to non-human objects or abstract things, e.g. names of geographical places or branded products. Here, however, the distinction is in some cases not trivial. Names of geographical places can sometimes be mentioned to refer to the people living or working in that area, in which case we would consider the item as an identity term. The following question is essential for the annotation: Is an expression used to address a message to a specific individual or a group by mentioning a specific identity of that individual or group?

In order to demonstrate the annotation procedure, two example messages are shown in Table 3. Consider the first example for which a total of five identity term mentions are annotated. @-marked user name mentions are always annotated automatically as they unambiguously refer to the identities of specific Twitter users. The term "Frauen" (*women*) denotes a specific gender identity, the term "Islam" denotes a specific religion and the term "Lehrerin" (*teacher*) denotes a specific professional group. It may perhaps be surprising that we consider specific occupation groups as identity terms since they do not refer to a specifically protected group; however, during annotation we also ask the question whether a mentioned group could possibly be a target of hate speech. If this is the case, a bias towards such group could appear in the hate speech data. Equally, we also consider, for example, the group of COVID-vaccinated people as an identity term as they can be targeted by hate speech and we want to avoid bias towards such a group when training systems.

The second example given in Table 3 contains a sub-word identity term "rechts" (*right-wing*) which is annotated as part of the word "rechts-radikal" (*radical right-wing*). Here, "radikal" (*radical*) is part of the hate speech expression, however, the direction of the political conviction "rechts" (*right-wing*) is a personal identity. Thus, we only annotate the sub-word character sequence which refers to the specific identity.

Using these criteria, identity term mentions were identified in each message in our dataset by one annotator. So far, we did not conduct an agreement study for this annotation.

A question for discussion remains how to proceed with the knowledge of the presence of possibly biased mentions of identity terms. Our goal is to annotate these terms to assess a possible bias in the dataset as well as to grant a possibility to automatically learn on a de-biased dataset by, for example, masking such occurrences. However, masking identity terms or anonymizing proper names is a double-edged sword. On the one hand, this step might seem necessary to avoid learning biases, e.g. by learning spe-

cific identity terms as features of hate speech, which is undesired. On the other hand, one might remove important information by modifying the input, which makes hate speech detection more difficult or even impossible. Certain expressions are only to be considered as hate speech if specific identity terms are targeted. The abovementioned label *V*, Incitement of masses ("Volksverhetzung"), requires that a specific group is mentioned. Nevertheless, we annotate such cases in our dataset and leave further processing options open.

Strings which refer to identity terms are identified as character sequences from each message and then listed for each message in the XML corpus. In the iHS dataset, there are 3,217 identity term mentions annotated, i.e. approximately 2.3 identity terms per message on average. 1,910 of the annotated identity term mentions are automatically identified @-marked Twitter username mentions. The remaining 1,307 annotations are manually identified strings from the messages referring to identities of individuals or groups.

The average number of identity terms for messages with different labels in the corpus is displayed in Table 4. Here, the number substantially varies for the different labels from 1.3 identity terms per message for label *A* to 4.0 for label *N*. As expected, messages with labels that do not necessarily require a mentioned target person or group contain a lower number of identity terms – i.e. label *A* which is based on § 111 StGB Public incitement to commit offenses ("Öffentliche Aufforderung zu Straftaten") and label *OL* which marks (untargeted) offensive language.

Table 5 shows all identity terms in the iHS dataset which are mentioned in more than ten messages. For each term, we additionally display the proportion of cases which are annotated belonging to any one of the four categories of illegal hate speech.

Here, a keyword bias as a result of our data acquisition method (see Sect. 2.1) can be observed. We find that seven different identity terms were also used as keywords for data sampling: "behindert" (*disabled*), "Flüchtling" (*refugee*), "Frauen" (*women*), "homosexuell" (*homosexual*), "Nazi", "Männer" (*men*) and "schwul" (*gay*). In total, these constitute for 157 identity term mentions out of the 1,307 manually identified strings. For future work, we suggest to not include identity terms in keyword-based sampling strategies.

Table 4 Average number of identity terms (idts) in messages for the different labels in the iHS dataset

Label	Avg. # idts
A	1.3
V	3.0
N	4.0
B	2.2
OL	1.5
Other	2.4
Overall	2.3

Besides, we observe that several frequent identity terms refer to politicians ("@ABaerbock") or political parties ("AfD", "CDU", "@Die_Gruenen"). Furthermore, terms referring to sexual, gender or national identities are also mentioned frequently. Many identified strings are specific Twitter usernames or abbreviations where it is not obvious which identity category they could belong to. Thus, for some identity terms we tried to deduce the full name and included it as an attribute for each identified identity term mention in the XML corpus. We have recently started to assign a category to some identity terms (e.g. *politician*), however, this categorization is still at an early stage.

3 Experiments

The presented iHS dataset is intended as a basis for automatic systems which are able to detect distinct classes of illegal hate speech. I conduct an experiment to learn this classification using a Transformer-based neural network model predicting whether a given message contains illegal hate speech.

The pre-trained German language model GBERT [4] has shown good results for German text classification tasks. It also serves as a basis for the task at hand to encode the text messages.[4] The pooled output of GBERT for each message is further processed with a classification head to predict the abovementioned six labels (see Table 2). As the iHS dataset is rather small it seems not to be sufficient to train the entire network on it. Thus, during fine-tuning on the iHS data, only the classification head is trained while the GBERT encoder is fixed in its pre-trained state. Therefore, I construct this classification head in a slightly more complex way: a dropout layer and two linear layers (the first one with 300 neurons, the second one with 6) with a RELU activation function in between.

I split the shuffled dataset into training and test data by means of stratified sampling due to the low number of instances for some classes. Models are trained on 80% of the iHS data using the Adam optimizer [11]. The performance of the trained model is evaluated by calculating accuracy and F1 score (for each class separately as well as a macro-average) on a test set consisting of the remaining 20% of the iHS data. The label distribution in the subsets is shown in Table 6. Hyperparameter optimization is implemented by a grid search for the dropout value, the number of training

[4] I use the model `deepset/gbert-base` available on https://huggingface.co/deepset/gbert-base.

Table 5 Most frequent identity terms in the iHS dataset with the proportion of the respective instances which are annotated to contain illegal hate speech

Identity term	Frequency	% HS
"Deutschland" (*Germany*)	40	33
"behindert" (*disabled*)	34	24
"Frauen" (*women*)	33	24
"Nazi"	31	32
"Männer" (*men*)	28	11
"AfD"	26	46
"CDU"	21	71
"Israel"	18	50
"Grünen"	17	53
"@welt"	15	40
"homosexuell" (*homosexual*)	13	8
"schwul" (*gay*)	12	0
"@ABaerbock"	11	27
"@Die_Gruenen"	11	64

epochs and the learning rate.[5] To deal with the imbalanced class distribution in the iHS dataset, I use class weights, which forces the model to optimize the prediction for all classes equally.

Performance results of the best[6] model from this first experiment are shown in the first row of scores in Table 7 (model "GBERT"). It can be observed that a rather low performance is achieved for this multiclass classification problem with a macro-averaged F1 score of 37%. The highest F1 score for a specific class annotated in the iHS dataset is achieved for the neutral class *Other* with 74%. The model does not manage to predict correctly any of the two instances in the test data labeled as *V*, which is the most infrequent class.

In a second experiment, the same training and evaluation procedure was applied, however, on data where all annotated identity term mentions have been masked by the token `*Name*`. Performance results of the best[7] model from this second experiment are shown in the second row of scores in Table 7 (model "GBERT [masked idt]"). The results show slightly lower scores for this model in comparison to the first experiment as the macro-average F1 score drops to 36%.

[5] I tested dropout values {0.01, 0.1, 0.2, 0.3}, number of training epochs {3, 5, 7, 10} and learning rates of {0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001} and report the best results based on a sum of accuracy and macro-averaged F1 score. It should be noted that this optimization was carried out on test data. Due to the small size of the dataset, holding back another portion of the dataset for optimization is not feasible. Thus, the results are to be taken with a grain of salt.

[6] The best model in the first experiment used the hyperparameters dropout = 0.01, number of epochs = 10 and learning rate = 0.001.

[7] The best model in the second experiment used the hyperparameters dropout = 0.1, number of epochs = 10 and learning rate = 0.0005.

Table 6 Label distribution in training and test set

	# instances	
Label	Training	Test
A	13	3
V	9	2
N	85	21
B	142	36
OL	179	45
Other	680	170
Total	1,108	277

Additionally, given the rather bad performance on the individual illegal hate speech classes, I investigate the ability of the models to detect illegal hate speech at all by aggregating the four subcategories. The aggregated class *HS* constitutes in total 311 instances in the dataset, split into 249 instances in the training set and 62 instances in the test set. Table 8 shows the performance of the respective models for this classification with three classes. Here, the GBERT model predicts these coarse-grained classes with a macro-average F1 score of 53%. Additionally, a consistent drop for all measures can also be observed for the model using masked identity terms.

4 Discussion

The experiment to detect illegal hate speech with a state-of-the-art Transformer model (see Sect. 3) shows that this multi-class classification task is challenging. None of the trained models achieves an F1 score of above 50% for any of the defined classes of illegal hate speech. The best macro-averaged F1 score over all six classes is 37%, which is not sufficient for an application to unseen data. In comparison, in the Shared Task on the Identification of Offensive Language (GermEval Task 2, 2019 [21]) the best system [16] achieved a macro-averaged F1 score of 53.59% at the fine-grained 4-way classification task. These results can be explained by considering that the detection of illegal hate speech is more challenging as it considers a higher number of classes and considerably fewer training data instances were available. However, we intended to develop a clearer categorization, consisting of classes which should be identifiable using specific patterns. This apparently cannot be shown with the current experiment.

Additionally, the rather small dataset and experimental setup limits the reliability of the results. A cross-validation sampling would possibly lead to more reliable results. Also, more data is needed, especially for the underrepresented classes *A* and *V*. Given more data, a more adequate data split into a larger training, test and validation set could be considered, which usually is suitable for deep learning

Table 7 Performance of models at illegal hate speech detection (fine-grained classes) on the iHS test data

			Class F1 scores					
Model	Accuracy	Macro-F1	*Other*	*OL*	*B*	*N*	*A*	*V*
GBERT	0.59	0.37	0.74	0.40	0.33	0.36	0.40	0.00
GBERT (masked idt)	0.53	0.36	0.68	0.34	0.37	0.38	0.36	0.00

Table 8 Performance of models at illegal hate speech detection (coarse-grained classes) on the iHS test data

			Class F1 scores		
Model	Accuracy	Macro-F1	*Other*	*OL*	*HS*
GBERT	0.61	0.53	0.73	0.42	0.44
GBERT (masked idt)	0.59	0.52	0.72	0.41	0.43

methods. Nevertheless, the neural network manages to predict the neutral class *Other* with an F1 score of 74%, which is acceptable. Thus, we can assume that it can differentiate such cases from offensive language (F1 42%) and illegal hate speech (F1 44%), while it struggles to find the correct subcategories.

The second experiment showed that masking identity terms leads to an overall drop in performance in comparison to the approach using the entire input. This result hints at the importance of identity terms for hate speech detection in some cases. However, the performance drop is only marginal. Thus, we can conclude that the model does not absolutely rely on the mentions of targets as features, since the masking of identity terms does not reduce the performance on all classes. In cases where hate speech can only be identified when considering the specific target, masking identities is detrimental. Hence, a different method might be more suitable to avoid a focus on identities which can lead to a model bias.

5 Conclusion

I suggest that there is a fine line when distinguishing a bias (e.g. an identity term bias) from actual features being relevant for illegal hate speech detection. The decision whether a message contains illegal hate speech or not should be made independently of mentioned personal identities of individuals or groups. However, completely masking the mentions of identity terms would be too radical as the mentioned group can play an important role in the evaluation of potential hate speech. A de-biased model is desired to differentiate features of hate speech in embeddings of targets from (for the detection of hate speech) irrelevant identity-specific information which could constitute a bias. Research on a larger scale using a larger dataset (e.g. the Civil Comments dataset [6]) is needed. Bias avoidance techniques applied during training (e.g. adversarial correction [26]) already tend to show promising results to deal with bias in hate speech datasets.

This paper provides a dataset which can be used as a starting point for further investigation of the phenomenon of illegal hate speech in online social media. Our strategies to avoid bias in the data may serve as a model for building a larger dataset. The conducted experiments show that the iHS dataset can be used to learn some features for the categorization while it does not solve the problem completely. During the data acquisition process, several bias mitigation approaches were implemented. Annotations of six categories in the data allow a more precise analysis of different hate speech features. In addition, the bottom-up annotation of mentions of identity terms in the iHS dataset can help to investigate possibilities to measure and mitigate bias. Future work should also include data augmentation methods to increase the number of examples, especially for the rarer subcategories of illegal hate speech.

Acknowledgements Special thanks to Kübra Boguslu for fruitful discussions while we cooperatively developed the annotation guidelines and annotated text messages for illegal hate speech categories and identity term mentions. Thanks as well to Hanna Westermann for her work on annotating identity term mentions and drafting a preliminary categorization.

Conflict of Interest
There are no competing financial or non-financial interests that are directly or indirectly related to this work.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Basile V, Bosco C, Fersini E, Nozza D, Patti V, Rangel Pardo FM, Rosso P, Sanguinetti M (2019) SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp 54–63 https://doi.org/10.18653/v1/S19-2007
2. Borkan D, Dixon L, Sorensen J, Thain N, Vasserman L (2019) Nuanced metrics for measuring unintended bias with real data for text classification. In: Companion proceedings of the 2019 world wide web conference, pp 491–500 https://doi.org/10.1145/3308560.3317593
3. Bosco C, Felice D, Poletto F, Sanguinetti M, Maurizio T (2018) Overview of the EVALITA 2018 hate speech detection task. In: EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, CEUR, vol 2263, pp 1–9. http://ceur-ws.org/Vol-2263/paper010.pdf. Accessed 23 Mar 2023
4. Chan B, Schweter S, Möller T (2020) German's next language model. In: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), pp 6788–6796 https://doi.org/10.18653/v1/2020.coling-main.598
5. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46. https://doi.org/10.1177/001316446002000104
6. Davidson T, Bhattacharya D, Weber I (2019) Racial bias in hate speech and abusive language detection datasets. In: Proceedings of the Third Workshop on Abusive Language Online, Florence, Italy, pp 25–35 https://doi.org/10.18653/v1/W19-3504
7. Demus C, Pitz J, Schütz M, Probol N, Siegel M, Labudde D (2022) Detox: A comprehensive dataset for German offensive language and conversation analysis. In: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Seattle, Washington (Hybrid), pp 143–153 https://doi.org/10.18653/v1/2022.woah-1.14
8. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Long and Short Papers. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol 1, pp 4171–4186 https://doi.org/10.18653/v1/N19-1423
9. ElSherief M, Kulkarni V, Nguyen D, Wang WY, Belding E (2018) Hate Lingo: A target-based linguistic analysis of hate speech in social media. In: Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018), pp 42–51
10. Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378–382. https://doi.org/10.1037/h0031619
11. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. https://arxiv.org/abs/1412.6980. Accessed 23 Mar 2023
12. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174. https://doi.org/10.2307/2529310
13. Liu P, Li W, Zou L (2019) NULI at SemEval-2019 Task 6: Transfer learning for offensive language detection using bidirectional transformers. In: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp 87–91 https://doi.org/10.18653/v1/S19-2011
14. Mandl T, Modha S, Shahi GK, Jaiswal AK, Nandini D, Patel D, Majumder P, Schäfer J (2020) 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, CEUR Workshop Proceedings, Hyderabad, India, pp 87–111. http://ceur-ws.org/Vol-2826/T2-1.pdf. Accessed 23 Mar 2023
15. Mandl T, Modha S, Shahi GK, Madhu H, Satapara S, Majumder P, Schäfer J, Ranasinghe T, Zampieri M, Nandini D, Jaiswal AK (2021) Overview of the HASOC subtrack at FIRE 2021: Hate speech and offensive content identification in English and Indo-Aryan Languages. In: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR Workshop Proceedings, India, pp 1–19. http://ceur-ws.org/Vol-3159/T1-1.pdf. Accessed 23 Mar 2023
16. Paraschiv A, Cercel DC (2019) UPB at GermEval-2019 Task 2: BERT-based offensive language classification of German Tweets. In: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019). https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/germeval/Germeval_Task_2_2019_paper_9.UPB.pdf. Accessed 23 Mar 2023
17. Risch J, Stoll A, Ziegele M, Krestel R (2019) hpiDEDIS at GermEval 2019: Offensive language identification using a German BERT model. In: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019). https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/germeval/Germeval_Task_2_2019_paper_10.HPIDEDIS.pdf. Accessed 23 Mar 2023
18. Ruppenhofer J, Siegel M, Wiegand M (2018) Guidelines for IGGSA Shared Task on the identification of offensive language. http://www.melaniesiegel.de/publications/2018_GermEval_Guidelines.pdf. Accessed 23 Mar 2023
19. Schäfer J, Boguslu K (2021) Towards annotating illegal hate speech: A computational linguistic approach. Detect Then Act (DTCT) technical report 3. https://dtct.eu/wp-content/uploads/2021/10/DTCT-TR3-CL.pdf. Accessed 23 Mar 2023
20. Schmidt A, Wiegand M (2017) A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media (SocialNLP@EACL 2017), Valencia, Spain, pp 1–10 https://doi.org/10.18653/v1/w17-1101
21. Struß J, Siegel M, Ruppenhofer J, Wiegand M, Klenner M (2019) Overview of GermEval task 2, 2019 shared task on the identification of offensive language. In: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019). German Society for Computational Linguistics & Language Technology, Erlangen, pp 354–365
22. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems, vol 30. Curran Associates, Red Hook
23. Wiedemann G, Yimam SM, Biemann C (2020) UHH-LT at SemEval-2020 Task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval@COLING 2020), pp 1638–1644 https://doi.org/10.18653/v1/2020.semeval-1.213
24. Wiegand M, Siegel M, Ruppenhofer J (2018) Overview of the GermEval 2018 shared task on the identification of offensive language. In: Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018). Österreichische Akademie der Wissenschaften, Vienna, pp 1–10
25. Wiegand M, Ruppenhofer J, Kleinbauer T (2019) Detection of abusive language: the problem of biased datasets. In: Long and short papers. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol 1. Association for Computational

Linguistics, Minneapolis, pp 602–608 https://doi.org/10.18653/v1/N19-1060

26. Yuan S, Maronikolakis A, Schütze H (2022) Separating hate speech and offensive language classes via adversarial debiasing. In: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH). Association for Computational Linguistics, Seattle, pp 1–10 https://doi.org/10.18653/v1/2022.woah-1.1
27. Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R (2019) SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval). In: Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT 2019), pp 75–86 https://doi.org/10.18653/v1/s19-2010