DeTox: A Comprehensive Dataset for German Offensive Language and Conversation Analysis

Christoph Demus[2,3], Jonas Pitz[1], Mina Schütz[1], Nadine Probol[1], Melanie Siegel[1], Dirk Labudde[2,3]

[1] Darmstadt University of Applied Sciences
Max-Planck-Straße 2, 64807 Dieburg
{jonas.pitz, mina.schuetz, melanie.siegel}@h-da.de
{nadine.probol}stud.@h-da.de

[2] Fraunhofer Institute for Secure Information Technology
Rheinstraße 75, 64295 Darmstadt
{christoph.demus, dirk.labudde}@sit.fraunhofer.de

[3] Mittweida University of Applied Sciences
Technikumplatz 17, 09648 Mittweida
{christoph.demus, dirk.labudde}@hs-mittweida.de

Abstract

In this work, we present a publicly available offensive language dataset (DeTox-dataset) containing 10,278 annotated German social media comments collected in the first half of 2021. With twelve different annotation categories annotated by six annotators, it is far more comprehensive than other datasets, and goes beyond just hate speech detection. The labels aim in particular also at toxicity, criminal relevance and discrimination types of comments. Furthermore, about half of the comments are from coherent parts of conversations, which opens the possibility to consider the comments contexts and do conversation analyses in order to research the contagion of offensive language in conversations. The dataset is available in our GitHub repository: https://github.com/hdaSprachtechnologie/detox

1 Introduction

With the increasing popularity of social networks in the last decade, people started to communicate more and more online, organising themselves in groups and social networks in general. It became easier than ever before to interact with foreign people because geographical distance played no role any more. While this is a great opportunity for our society, it does not come without risks regarding toxic and offensive language.

Whereas many research groups focus mainly on binary hate speech classification, offensive language contains several other aspects. These include insult, threat, and discrimination based on characteristics such as age, gender, ethnicity, religion, and sexual orientation.

The two main tasks to limit the amount of toxic language are detection and classification, e.g., to identify threats at an early stage or to effectively support criminal investigators in their work. The basis to train algorithms that can support such assessments are labelled datasets of high quality as well as quantity.

Such datasets exist only for a few languages, e.g. Vidgen et al. (2021) provided the Contextual Abuse Dataset with fine grained labels for English language. For many languages, including German, this is a limiting research factor. Therefore, motivated by the concrete application to assist a fine granular classification of offensive comments in a reporting centre for hate comments[1] of the German state government, we present a new German dataset that aims among others at hate speech, toxicity, sentiment, target, but also at criminal relevance (regarding German law) and threat. It contributes in three main aspects: (1) With 10,278 annotated comments it provides a new valuable resource for the German hate speech community. (2) Having twelve different labels per comment opens broad research and application options beyond basic hate speech detection and (3) the inclusion of whole conversation threads being partly annotated allows to make use of comments contexts as well as other supervised and unsupervised conversation analyses.

[1] https://hessengegenhetze.de/

Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 143 - 153

Dataset	Source	# Comments	Tasks
Bretschneider and Peters (2017)	Facebook	5,600	binary hate speech and intensity (moderate or clearly)
Ross et al. (2017)	Twitter	470	binary hate speech and intensity (scale 1-6)
GermEval 2018 and 2019 (Wiegand et al., 2018; Struß et al., 2019)	Twitter	15,567	coarse: offense, other fine: abuse, insult, profanity
	Twitter	2,888	implicit, explicit
HASOC 2019 (Mandl et al., 2019)	Twitter, Facebook	4,669	coarse: binary offense fine: hate, offensive, profane
GermEval 2021 (Risch et al., 2021)	Facebook	4,188	toxic, non-toxic engaging comments fact-claiming comments

Table 1: **Overview of Public German Datasets with Hate Speech Related Annotations.**

2 Related Work

2.1 German Datasets

In the last years, shared tasks played a major role for research in the hate speech detection field as they were accompanied with appropriate annotated datasets for the German language (Tab. 1). The largest dataset was created by the organizers of the GermEval 2018 and 2019 shared tasks (Struß et al., 2019), with the dataset of 2019 being an extended version of the data in 2018 and containing a total of over 15,000 comments with offensive language annotations. In the following version of GermEval in 2021 (Risch et al., 2021) a new dataset with slightly different tasks was published. The dataset of the HASOC 2019 (Mandl et al., 2019) has similar annotations to those of the GermEval 2018 and 2019 datasets. Bretschneider and Peters (2017) focused on detecting hate against foreigners. All presented datasets contain data from social networks, which represent typical online conversations.

2.2 Data quality

Aside from the quantity, the quality of the data in a dataset is of major importance. The data quality can be examined from three different viewpoints: Interpretability, relevance and accuracy (Kiefer, 2016). Interpretability describes whether the data is technically interpretable by the algorithm. An example would be a NLP-Model that was designed for text inputs and therefore cannot process images. Relevance describes whether the data is appropriate for the given problem that should be solved. For hate speech detection this means that the data should contain a certain amount of hate speech but also non hate speech comments, and it should ideally be unbiased. Finally, accuracy indicates, whether the data reflects the reality. All those factors influence each other.

2.3 Data Collection Strategies

As there is no perfect strategy to create a dataset that fulfils the aforementioned factors as much as possible, research groups use different methods for data collection. One main issue for hate speech collections is that the real proportion of hateful comments in social networks is too low to train models on (Schmidt and Wiegand, 2017). Therefore, it is often necessary to enrich the corpus with additional hate speech comments. Waseem and Hovy (2016) suggest to first identify frequently used swearwords and slurs, and then search for comments containing these words. For example, Zampieri et al. (2019a) created a dataset (OLID) - for the OffensEval shared task 2019 (Zampieri et al., 2019b) - using only ten keywords. Wiegand et al. (2018) concern that this strategy could lead to a missing variety of offensive terms, which could lead to hate speech detection models just learning those keywords (Schmidt and Wiegand, 2017). Therefore, for the GermEval 2018 and 2019 datasets (Wiegand et al., 2018; Struß et al., 2019), the authors first identified Twitter accounts that regularly post hate speech by using keyword lists. Then they sampled comments that were posted by these users. On the one hand, this omits the keyword search, but on the other hand, a single user might use a certain vocabulary. To counteract this problem, they separated the users for the train and test set splits. A combination of both methods was used by Mandl et al. (2019). In 2020 the HASOC organizers (Mandl et al., 2020) used preliminary datasets to train a simple SVM model with an average performance which they then used to identify possible hate speech comments on Twitter to create a new dataset. In addition, they included a small amount of random comments.

2.4 Annotation Strategies

Three main factors that contribute to a high annotation quality are (1) the selection of annotators, (2) the annotation schema and (3) the annotation process itself, including the process of quality insurance.

There are three options for who annotates the collected data (Poletto et al., 2020). In the best case, the data is annotated by selected subject-matter experts. However, this is not always possible due to the amount of work involved. Therefore, amateurs are often used for annotation. These can also be selected individuals (e.g., students) who are familiar with the subject background. The third possibility is the use of crowdsourcing platforms, where the annotators are not known in advance.

In all cases where non-experts annotate data, they should ideally go through a training process before they start the labelling process to ensure a high quality of the annotations. In the mentioned shared tasks, the first two methods were used, i.e. the data was either annotated by the authors themselves or by selected individuals.

2.5 Inter-Annotator Agreement

The inter-annotator agreement (IAA) is an important measure to assess the quality of the annotations. Depending on the number of annotators and the data type, there are several measures that can be used to evaluate the IAA. The most popular are Kappa-measures like Cohens (Cohen, 1960) or Fleiss Kappa (Fleiss, 1971) and Krippendorff's alpha (Krippendorff, 1980). The latter is especially used for datasets containing missing data values.

Gwet (2008) introduced the AC_1 (AC - Agreement Coefficient) measure for IAA. The author shows that this measure is more resistant against the paradoxes of Kappa measures, which is described in detail in Feinstein and Cicchetti (1990) and Gwet (2015) (despite the name the paradoxes are also valid for Krippendorff's alpha). Furthermore, its weighted version AC_2 (Gwet, 2014) is able to handle different scales (e.g. ordinal scale).

3 Data Collection and Description

For this dataset we used Twitter as the data source, as it grants free access to most of its tweets for research purposes, and it is possible to (automatically) extract tweets by multiple criteria via the Twitter-API. The use of Twitter text data guarantees a high interpretability (see Sec. 2.2) and thus allows algorithms to be developed using this data.

In contrast to previous related work, we aimed not only at collecting single comments but also at collecting whole conversations or parts thereof, which means tweets or comments and their reply trees. Both require different data collection strategies, which will be explained in the following sections. All collected comments and conversations are in German language and posted in the first half of 2021. The most present topics in the media during the time we crawled the data were the Corona pandemic with all its aspects, as well as politics related to the elections of the German Bundestag in September 2021.

3.1 Comments

As we intended to cover a wide range of topics, types of discrimination, and political attitudes, we manually created keyword lists for the fields we wanted to receive comments for. As keywords, we used words that we expected to occur in offensive comments as well as offensive words. Furthermore, we determined keywords with the help of Google Trends in order to capture currently much discussed topics. For example, we used "merkelmussweg" (engl. "Merkel must go", often used as a hashtag) as one keyword for political attitude and "Querdenker" (engl. "lateral thinkers"), which is a pejorative term for Corona deniers, for Corona-related hate speech, but also words like "Jude" (engl. "jew") that are neutral by its own but often used in discriminating comments. In the end, our keyword lists contained a total of 131 words. During the comment search we did not only search the comment text for the keywords but also the hashtags. With these keyword lists, we pulled 781,991 comments from 154,151 Twitter users.

In a second step - to create a smaller dataset with a higher probability to contain offensive and relevant content - we filtered these comments with two additional lists: 1) a hate word list and 2) a list containing profane words.[2] The hate word list was set up for an earlier participation at GermEval 2018 (Siegel and Meyer, 2018). It was extended on different hate speech corpora using the tf-idf mechanism. The profane word list was extracted from a website containing around 2,000 offensive and profane words in German. For our sampling strategy each of the filtered tweets needed to con-

[2] https://www.insult.wiki/schimpfwort-liste

Comments	annotated	not annotated	Total
single Com.	4,936	0	4,936
Com. of Conv.	5,342	444,300	449,642
Total	10,278	444,300	454,578

Table 2: **General Statistics of the complete Dataset:** Numbers of annotated and additional not annotated comments in the single comments and conversation part of our dataset.

Single Comments	
# hate word filtered	3,214
# unfiltered	1,722
Conversations	
# Convs	514
Mean # Com. per Conv.	873.09
Mean # Authors per Conv.	502.14

Table 3: **Additional Dataset Statistics:** Statistics regarding the composition of the single comments and the number and size of conversations. All together the datasets contains 100 conversations with more than five annotated comments those conversations contain on average 45 annotated comments (max. 463 annotated comments per conversation).

tain at least one word from each of the two lists. Finally, we took the comments for the annotation from about two thirds from the pre-filtered stream and one third from the 781,991 comments set (Tables 2 and 3).

3.2 Conversations

For the selection of conversations, we first selected parent tweets on Twitter and then pulled the whole response tree. This can be done by searching for all tweets having the same conversation ID as the parent tweet.

We expected that by involving entire conversations, the hate speech portion on Twitter would be reflected more realistically, addressing the requirement for data accuracy. But, we also expected that this dataset would contain less hate speech overall than the dataset with individual comments, thus leading to a problem of relevancy. To counteract this problem, we selected a total of 25 Twitter pages containing content of politicians, scientists related to the Corona pandemic, conspiracy theorists and influencers. The selection was based on those figures often being a catalyst for controversial discussions in recent media. This resulted in 4,698 conversations containing 637,027 comments. For annotation, we intended to select coherent conversation parts that may contain - with a high probability - hate speech. Therefore, in a first step we selected comments from these conversations, which have 10 to 199 direct replies but, to avoid major biases, are not posts of the owners of the crawled twitter pages. This resulted in 1,665 comments that were annotated. As it is known that offensive comments trigger other users to post offensive responses (Cheng et al., 2017; Almerekhi et al., 2020) we used this knowledge to find offensive passages in the conversation trees. Therefore, in a second step, we noted 57 comments from 49 conversations that were annotated as hate speech (majority voting) or toxic (averaged toxicity annotation > 2.5). Finally, we extracted these comments' parent comments and all their successor comments (the whole conversation after each selected comment). This resulted in 5,342 annotated comments belonging to captured conversations (Tables 2 and 3). Next to the annotated comments belonging to conversations, we also included all not annotated comments of conversations where at least one comment was annotated from as these comments could be useful for unsupervised analyses.

4 Data Annotation

The annotation scheme was established to best support the specific task of building models for fine grained classification in the mentioned reporting office for hate comments but also with a view to future research. This resulted in a comprehensive annotation schema with twelve different categories at two levels. Furthermore, various metadata such as annotation time and duration were logged to leave the possibility not only to use the dataset to train models but also for future analyses like the Inter-Rater-Agreement-Learning described by Hanke et al. (2020) that uses annotation metadata to compute the reliability of annotators.

4.1 Annotation Schema

An overview over the annotation schema is given in Figure 1. Initially, comments that could not be (fully) understood, i.e. because of missing context, could be labelled as "Incomprehensible" which made further annotations to a comment voluntary. If this was not the case, the other main categories had to be annotated. "Sentiment" refers to the assumed emotional state of the comment's author when writing the comment: negative, neutral or positive. "Expression" describes whether the author expressed its message in an implicit or explicit manner. With the "Target" of a comment, we refer

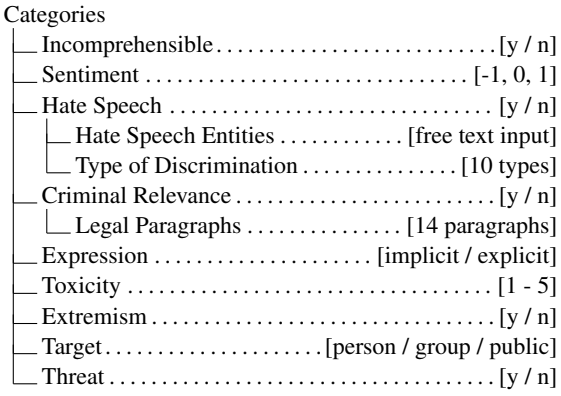

Figure 1: **Overview of the Annotation Schema:** The categories and their respective labels ("y" - yes, "n" - no). Categories in second order depend on their parent category.

to who is addressed, as this is of importance for hate speech contagion in conversations (Kwon and Gruzd, 2017). The comment can be addressed to a single or multiple separate persons, a group or groups of people, or it can have no specific target (public). With the category "Threat" we address comments that invoke or announce acts of violence and therefore pose a direct danger or threat to the public.

While "Toxicity" and "Hate Speech" are closely related, they are not interchangeable and can even occur independently of each other. To distinguish between the two categories, we used the following definitions:

Toxicity: Toxicity indicates the potential of a comment to "poison" a conversation. The more it encourages aggressive responses or triggers other participants to leave the conversation, the more toxic the comment is. We introduced a scale of 1 (not toxic) to 5 (very toxic) to be able to model the impact of toxic comments on the conversation more accurately.

Hate Speech [3] **:** Hate speech is defined as any form of expression that attacks or disparages persons or groups by characteristics attributed to the groups. Discriminatory statements can be aimed at, for example, political attitudes, religious affiliation or sexual identity of the victims.

In a free text input form, the annotators could submit words or phrases that were pivotal in their decision to label the comment as "Hate Speech". The type of discrimination could also be specified. The following types of discrimination were available for selection (zero, one or multiple selections were possible):

Types of Discrimination: Job; Political Attitude; Personal Engagement and Interests; Sexual Identity; Physical, Psychological or Mental Characteristics; Nationality; Religion; Social Status; World View; Ethnicity.

The category "Criminal Relevance" indicates whether a comment can be considered as relevant under German criminal law. If a comment was selected to be criminally relevant, annotators had to further specify the legal paragraphs that were applicable. This was one of the most difficult tasks for the annotators, as they did not have a legal background. The following paragraphs were considered to be applicable to online comments:

Legal Paragraphs (StGB[4]**):** § 86, § 86a, § 111, § 126, § 130, § 131, § 140, § 166, § 185, § 186, § 187, § 189, § 240, § 241.

4.2 Annotation Disagreements

Labelling hate speech data relates a lot to personal beliefs, experience and demographic properties (Sap et al., 2021). As our main goal was to train models for classification, we applied a prescriptive annotation standard, meaning we aimed at having clear decisions regarding to annotation guidelines and not surveying personal annotator beliefs (Röttger et al., 2021). Nevertheless, also the use of detailed annotation guidelines cannot reach full objectivity. As a result disagreements between the annotators will necessarily appear and can be handled in multiple ways. Common strategies are majority voting for classification on a nominal (incl. binary) scale and averaging for classification on an ordinal scale. Majority voting has the property, that underrepresented opinions get likely voted out, in particular if the number of annotators is high (Davani et al., 2022). If this is good or bad depends on the specific application. To avoid this behaviour other approaches model annotators separately and even make it possible to estimate uncertainty which could be used to make no decision if uncertainty is high (Davani et al., 2022). We used majority voting and averaging for this work but also included the single annotations of each annotator in the dataset to allow other approaches.

[3] Based on the definition of the United Nations: https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf

[4] StGB (engl. German Criminal Code): https://www.gesetze-im-internet.de/englisch_stgb/index.html

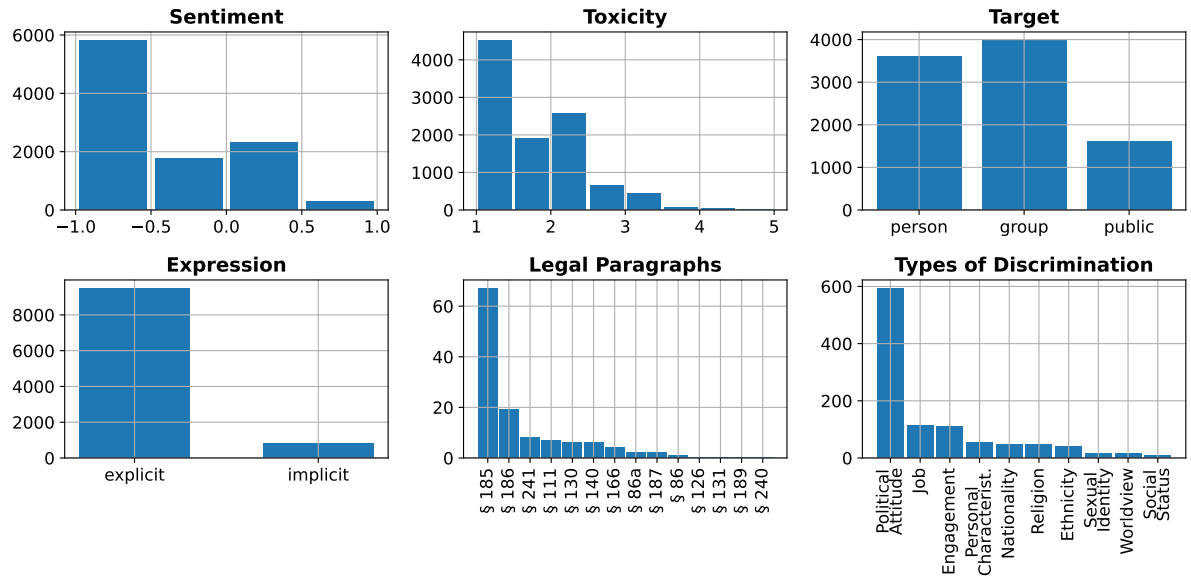

Figure 2: **Frequency Distributions for the Labels Sentiment, Toxicity, Target, Expression, Legal Paragraphs and Types of Discrimination:** Sentiment values reach from -1 (negative) to +1 (positive), Toxicity values reach from 1 (not toxic) to 5 (most toxic). The categories Legal Paragraphs and Type of Discrimination are multi label classes related to the labels Criminal Relevance and Hate Speech respectively. The paragraphs meant to be paragraphs in the German "Strafgesetzbuch (StGB)" (engl. German Criminal Code).

4.3 Annotation Process

The group of annotators consisted of six students, four of them studying Information Science and two studying General and Digital Forensics. The complete annotation process was permanently monitored. First, we introduced the annotators for the task, including an explanation of the annotation guidelines. Afterwards, we started a training phase, where each annotator annotated 200 comments split in two sets. After each set, we identified controversial annotated comments and discussed them. Then, the annotators were split in two groups of three persons each according to their annotations in the training phase. That means, annotators were divided such that the annotators are equally distributed on how offensive they labelled comments on average. The goal was to avoid that all annotators, who tend to label comments more toxic than others, are in the same group, as this would bias the annotations. In the next step, we began with the annotation phase, which run over about five months. In this phase, the data was annotated in 8 batches. To cut it short, finally every comment was annotated by three annotators and every annotator annotated 5139 comments (half of the data). In the datasets, single Twitter comments and comments from conversations were mixed and roughly equally distributed. During this phase the inter annotator agreement was permanently monitored and every one to three weeks unclear comments were extracted and discussed, also under consideration of the agreement in all categories.

For subsequent analyses on consistency (Sec. 5.2) of the annotations, each annotator had to annotate 20 randomly selected comments per annotation set twice. This results in 123 to 138 twice annotated comments per annotator (it is not $8 \cdot 20 = 160$ as we additionally had some other non-public comments in the annotation datasets).

5 Results and Discussion

In this section, we will first outline general specifications like the frequencies of the annotated labels of the dataset. In the second part the annotation quality containing measures for the IAA and the annotator consistency are presented and finally a closer look at the conversations is taken. As the dataset is very comprehensive, we can only show selected, most important statistics and results.

5.1 General

After the annotation process, the dataset contains 31,327 Annotations for 10,278 Twitter comments. The 141 comments for which the annotations differed the most (difference of > 7 annotations, with 12 being the maximum) were re-examined by the

Category	Com.	Conv.	Total
Incompre-hensible	117 2.39 %	214 4.01 %	332 3.23 %
Hate Speech	880 17.83 %	235 4.40 %	1,115 10.85 %
Criminal Relevance	99 2.01 %	32 0.60 %	131 1.27 %
Extremism	55 1.11 %	27 0.51 %	82 0.80 %
Threat	11 0.22 %	1 0.02 %	12 0.12 %

Table 4: **Frequencies of the Annotation Categories:** Absolute and percentage values for the frequencies of the binary annotation categories separated in single comments (Com.) and comments from conversations (Conv.).

Category	Group A	Group B
Incomprehensible	0.7982	0.9343
Sentiment	0.7744	0.8785
Hate speech	0.7286	0.8056
Criminal Relevance	0.9368	0.9364
Expression	0.9625	0.9515
Toxicity	0.8584	0.9159
Target	0.7281	0.7701
Extremism	0.5441	0.6086
Threat	0.9987	0.9997
Mean	**0.8144**	**0.8667**

Table 5: **Inter-Annotator Agreement:** IAA for all labels and both groups of annotators containing three annotators each. Sentiment and Toxicity values are computed with the AC_2 measure, all others with AC_1.

authors.

For the analysis, we separated the labels in binary ones (Tab. 4) and non-binary ones (Fig. 2). For the binary labels, except for the category Threat, we did a majority voting to achieve a gold standard. As the category Threat is much less represented than the other categories, we lowered the border and assumed a comment as Threat, if at least one annotator labelled it as Threat. In real applications, one would likely do the same, not to miss any comments that pose a threat. The analysis of the binary label frequencies shows that the hate speech proportion of the complete dataset is 10.85 % (1,115 comments) and 1.27 % (131 comments) are labelled as criminally relevant. The categories Extremism and Threat were only in less than 1 % of the comments labelled as true. In contrast to that, 3.23 % (329 comments) were annotated as incomprehensible by the majority of the annotators, which means the sense of the comment could not fully be understood, and therefore was not (completely) labelled. Table 4 also shows that the single comments part of the dataset contains a higher proportion of offensive comments. This is visible most clearly in the category Hate Speech, where the proportion in the single comments part is 17.83 % but in the conversations part only 4.40 %. The reason is that each comment in the single comments part was selected only by its own properties, in particular by keyword search. In contrast to that, in the conversations part not the comments were selected but whole conversations with all comments. Therefore, this is an expected observation.

Figure 2 shows the frequency distributions for the labels Sentiment, Toxicity, Target, Expression, Legal Paragraphs, and Type of Discrimination. It is noticeable that most of the comments have a toxicity of less than 2.5, although the sentiment of the majority of the comments is negative (-1 is the most negative). Nevertheless, the percentage of toxic comments is with 9.63 % just a little lower than the hate speech proportion in the dataset. Concerning the target of the comments, it shows that specific persons and groups are almost equally addressed, and it is rare that a comment addresses no specific target.

The categories Legal Paragraphs and Type of Discrimination differ from the others as they are connected to other categories (Criminal Relevance and Hate Speech respectively) and they are multi-label categories. As before also for the paragraphs and the types of discrimination a majority voting was done. The most often annotated paragraph is by far § 185 ”Beleidigung” (engl. insult) followed by § 186 ”Üble Nachrede” (engl. malicious gossip). Regarding the Type of Discrimination, the dominating category is ”Political Attitude” which suggests, that most of the hate speech comments seem to be offensive towards the political view of people.

5.2 Inter-Annotator Agreement and Consistency

To assess the quality of the annotations, we measured the IAA and the consistency of the annotators using Gwets agreement coefficients (AC_1, AC_2), as they are resistant against the paradoxes of Kappa-measures and resulted in more realistic values here (see Sec. 2.5). AC_1 (with a nominal scale) was used for all classes except Sentiment and Toxicity. As they have an ordinal scale, we used AC_2 for them. For both, the IAA and the consistency, we did not evaluate the categories ”Legal Paragraphs” and ”Type of Discrimination” here, as they depend

Category	A1	A2	A3	B1	B2	B3	Mean
Incomprehensible	0.94	0.96	0.80	0.97	0.98	0.93	**0.93**
Sentiment	0.86	0.94	0.69	0.92	0.94	0.94	**0.88**
Hate speech	0.90	0.95	0.77	0.90	0.89	0.93	**0.89**
Criminal Relevance	0.95	0.99	0.83	0.95	0.95	0.91	**0.93**
Expression	0.89	0.85	0.67	0.77	0.89	0.86	**0.82**
Toxicity	0.94	0.96	0.97	0.97	0.95	0.94	**0.95**
Target	0.82	0.73	0.62	0.80	0.76	0.76	**0.75**
Extremism	0.99	1.00	0.78	0.94	0.95	0.98	**0.94**
Threat	1.00	1.00	0.80	0.96	0.98	0.97	**0.95**
Mean	**0.92**	**0.93**	**0.77**	**0.91**	**0.92**	**0.91**	**0.89**

Table 6: **Annotator Consistencies:** Every annotator labelled around 130 comments twice. From these duplicate annotations, the agreements for every annotator and every category were computed using Gwets AC_1 and AC_2 (for Sentiment and Toxicity) measures.

	Random Selection	Answers of offensive Comments
Toxic	1.97 %	5.97 %
Hate Speech	2.81 %	6.24 %

Table 7: **Proportion of Offensive Comments in Conversations:** Random Selection shows the proportion of toxic and hate speech comments in 1,673 random selected comments from conversations. The second column shows the proportion of toxic and hate speech comments in 881 answers to comments that were labelled as toxic / hate speech.

on the categories "Criminal Relevance" and "Hate Speech" respectively which would require more complex analyses to get reliable results.

The IAA (Tab. 5) was measured over all comments for each of the two groups of annotators. The table shows that the mean agreement of group A is about 0.05 lower than the agreement of group B. Still, both groups have mean values over 0.8 which indicates a very good agreement. Looking at the IAA of each label category, it is visible that the category "Extremism" has by far the lowest agreement in both groups (0.54 and 0.61) and "Threat" has with over 0.99 the highest agreement. The latter is caused by the fact that there are just 12 comments in the whole dataset that are labelled threatening at all.

In addition, we analysed the consistency of the annotators using the duplicate annotations of each annotator (see Sec. 4.3). An ideal annotator would annotate the same comment always the same (high consistency) but in reality this is not the case.

The analysis (Tab. 6) shows that five of the six annotators have - with an agreement over 0.80 in their twice annotated comments - a very good consistency, which indicates that they labelled the same comment both times almost equally. Annotator A3 has with a value of 0.77 a lower consistency but still being good (> 0.6). This could also be one reason, why Group A has a lower average IAA. In contrast to the IAA, the consistency of the category Extremism is, except of annotator A3, very good (over 0.90). This shows that there might have been different interpretations of these category as it would have lead to a better IAA otherwise.

5.3 Conversations

A main question in the conversation analysis related to hate speech is, what impact an offensive comment to the following conversation has. In our analysis, we define all comments as offensive that are labelled as hate speech (majority voting) or toxic (averaged toxicity annotation > 2.5). Then we compared the proportion of offensive comments in the random selection (from the first annotation step, see Sec. 3.2) with the proportion of offensive comments in direct answers to offensive comments (annotated in step 2). The results in Table 7 show, that the proportion of toxic comments in answers to offensive comments is with almost 6 % three times higher than in the random selection. For hate speech it is with 6.24 % even a bit higher. This observation indicates that offensive comments trigger users to answer with offensive speech.

6 Baseline Models

The categories hate speech, toxicity and sentiment were selected to train simple baseline models on. We did not make use of comments contexts so far, this will be done in later work. Even though toxicity and sentiment are regression tasks, we used classification models for them as this heavily improved the performance for the underrepresented classes (high toxicity and positive Sentiment).

We used a multi layer perceptron (MLP) with an

	MLP			SVM			GBert			XLM-R		
Category	**Prec**	**Re**	**F1**	**Prec**	**Re**	**F1**	**Prec**	**Re**	**F1**	**Prec**	**Re**	**F1**
Hate Speech	0.67 (0.85)	0.54 (0.89)	0.56 (0.85)	0.65 (0.90)	0.79 (0.80)	0.67 (0.83)	0.78 (0.89)	0.67 (0.91)	0.71 (0.89)	0.53 (0.79)	0.58 (0.89)	0.55 (0.83)
Toxicity	0.28 (0.53)	0.27 (0.54)	0.27 (0.53)	0.35 (0.66)	0.41 (0.61)	0.35 (0.62)	0.41 (0.67)	0.37 (0.68)	0.39 (0.67)	0.56 (0.67)	0.56 (0.66)	0.54 (0.65)
Sentiment	0.60 (0.62)	0.44 (0.63)	0.45 (0.60)	0.58 (0.71)	0.63 (0.70)	0.59 (0.70)	0.66 (0.71)	0.55 (0.71)	0.58 (0.71)	0.64 (0.72)	0.64 (0.71)	0.63 (0.71)

Table 8: **Performance measures of our baseline models on the given labels:** The values are macro averaged and in round brackets the weighted values are given.

additional embedding layer with a vocabulary size of 15,000 and softmax function, an SVM model that uses an 200 dimensional Fasttext feature vector as input as well as GBert and XLM-R Transformer models. We did a stratified train-test-split (80 % training, 20 % test) and evaluated the results (Tab. 8) using macro and weighted (values in round brackets) precision, recall and F1-score. The bigger the difference between the macro and weighted value, the bigger is the difference of the recall scores of the classes.

In most categories the SVM outperforms the MLP and the transformer models slightly outperform the SVM. In particular, the macro recall scores of the SVM, which are relevant for detection of offensive language, are higher compared to the MLP and on a same level as the transformer models. Overall the MLP tended to have a higher performance on the majority class but a much lower performance on the minority class. In the other models this gap was mostly smaller. GBert produced better results for Hate Speech detection while XLM-R had better macro average scores for Toxicity.

7 Limitations of the Dataset

Even though the data collection and annotations were done as properly as possible, the dataset has some limitations. Twitter as a data source has some disadvantages: First, it is just one of many social networks. Every network brings its own properties and influences therefore the people, their writing style and communication standards. Second, comments on Twitter are moderated and therefore offensive language might have already been removed before our data collection. A more general problem, which is partly but not exclusively caused by the method of keyword search, is the presence of selected topics limiting generalizability. Regarding the annotations, even with three annotations per comment the number is relatively small resulting to a high influence of possible biased annotations or annotators. The comprehensive annotation schema is complex to annotate, and the definitions of hate speech and toxicity naturally leave a lot of room for personal interpretations. Further, the annotations for the legal paragraphs should be treated carefully, as no annotator (and also no one of the authors) has a legal background. Finally, the annotations for legal paragraphs are specific to German legislation.

8 Conclusion

Modern machine learning methods require sufficient amounts of annotated data that are of high quality and at the same time, the annotations must be granular enough that the learned models can be used effectively in real applications.

For this dataset the data was carefully selected and biases were avoided as much as possible. The annotation schema was developed together with first-hand users from a reporting office for offensive comments in Germany. During the annotation process, the quality was systematically monitored and adjusted. Parts of the data are available together with their conversation contexts (i.e. its parent comments and replies). We have conducted initial statistical data analyses with the annotated data, which we will continue in the future and trained baseline models on selected categories.

The dataset gives the possibility to train models on high quality annotated data that go beyond binary classification tasks. Moreover, it can be used to build more complex algorithms which may take the comments' context into account and even conversation analyses and analyses regarding the spread of offensive language are possible.

9 Acknowledgements

This work is enhanced by the Darmstadt University of Applied Sciences, which supported this work with the research in Information Science (https://sis.h-da.de/), in collaboration with the Fraunhofer Institute for Secure Information Tech-

nology. Additionally, this contribution has been funded by the project "DeTox" (Cybersecurity research funding of the Hessian Ministry of the Interior and Sports) which is a collaboration with the Hessian CyberCompetenceCenter (Hessen3C).

References

Hind Almerekhi, Haewoon Kwak, Joni Salminen, and Bernard J. Jansen. 2020. Are these comments triggering? predicting triggers of toxicity in online discussions. In *Proceedings of The Web Conference 2020*. ACM.

Uwe Bretschneider and Ralf Peters. 2017. Detecting offensive statements towards foreigners in social media. In *Proceedings of the 50th Hawaii International Conference on System Sciences (2017)*. Hawaii International Conference on System Sciences.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1217–1230, New York, NY, USA. Association for Computing Machinery.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

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

Kilem Gwet. 2015. On krippendorff's alpha coefficient.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Kilem Li Gwet. 2014. *Handbook of Inter-Rater Reliability*, chapter Agreement Coefficients for Ordinal, Interval, and Ratio Data. Advanced Analytics, LLC.

Kai-Jannis Hanke, Andy Ludwig, Dirk Labudde, and Michael Spranger. 2020. Towards inter-rater-agreement-learning. In *The Tenth International Conference on Advances in Information Mining and Management*.

Cornelia Kiefer. 2016. Assessing the quality of unstructured data: An initial overview. In *LWDA*, pages 62–73.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology*, chapter 12. Sage Publications, Beverly Hills.

K. Hazel Kwon and Anatoliy Gruzd. 2017. Is offensive commenting contagious online? examining public vs interpersonal swearing in response to donald trump's YouTube campaign videos. *Internet Research*, 27(4):991–1010.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 29–32, New York, NY, USA. Association for Computing Machinery.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC track at FIRE 2019. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*. ACM.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, pages 1–47.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Workshop on the Identification of Toxic, Engaging, and Fact-Claiming Comments : 17th Conference on Natural Language Processing KONVENS 2021*.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication (Bochum), Bochumer Linguistische Arbeitsberichte, vol. 17, sep 2016, pp. 6-9*.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B. Pierrehumbert. 2021. Two contrasting data annotation paradigms for subjective nlp tasks.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Melanie Siegel and Markus Meyer. 2018. h_da submission for the GermEval shared task on the identification of offensive language. In *Proceedings of the GermEval 2018 Workshop*, Vienna, Austria. Austrian Academy of Sciences.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 354–365, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of the GermEval 2018 Workshop*, Vienna, Austria. Austrian Academy of Sciences.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.