Natural Language Engineering (2023), **29**, pp. 1561–1585
doi:10.1017/S1351324923000098

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

The problem of varying annotations to identify abusive language in social media content†

Nina Seemann*, Yeong Su Lee, Julian Höllig and Michaela Geierhos

Research Institute CODE, University of the Bundeswehr Munich, Neubiberg, Germany
*Corresponding author. E-mail: nina.seemann@unibw.de

(Received 31 August 2021; revised 8 July 2022; accepted 24 July 2022; first published online 29 March 2023)

Abstract

With the increase of user-generated content on social media, the detection of abusive language has become crucial and is therefore reflected in several shared tasks that have been performed in recent years. The development of automatic detection systems is desirable, and the classification of abusive social media content can be solved with the help of machine learning. The basis for successful development of machine learning models is the availability of consistently labeled training data. But a diversity of terms and definitions of abusive language is a crucial barrier. In this work, we analyze a total of nine datasets—five English and four German datasets—designed for detecting abusive online content. We provide a detailed description of the datasets, that is, for which tasks the dataset was created, how the data were collected, and its annotation guidelines. Our analysis shows that there is no standard definition of abusive language, which often leads to inconsistent annotations. As a consequence, it is difficult to draw cross-domain conclusions, share datasets, or use models for other abusive social media language tasks. Furthermore, our manual inspection of a random sample of each dataset revealed controversial examples. We highlight challenges in data annotation by discussing those examples, and present common problems in the annotation process, such as contradictory annotations and missing context information. Finally, to complement our theoretical work, we conduct generalization experiments on three German datasets.

Keywords: Natural Language Processing; Abusive Language; Dataset Analysis

1. Introduction

During the ongoing pandemic and the resulting stay-at-home orders, the social media usage increased globally by 13.2% from 2020 to 2021, according to DataPortal (Kemp 2021). Back in April 2020, Facebook representatives stated that the "usage growth from COVID-19 is unprecedented across the industry, and we are experiencing new records in usage almost every day" (Schultz and Parikh 2020). Unfortunately, there are people who use online platforms to utter profanities, insult and/or disparage other people or even denigrate entire groups. In research, this kind of user-generated content has many names, for example, "offensive language," "abusive language," or "hate speech." Throughout this paper, we will refer to all related names consistently with "abusive language" unless we are referring to other works. From a juridical point of view, the legal treatment of abusive language varies quite significantly across countries. On one hand, the U.S. Supreme Court has repeatedly ruled that most of what would be considered as abusive language

†This research is partially funded by dtec.bw—Digitalization and Technology Research Center of the Bundeswehr within the project MuQuaNet. dtec.bw is funded by the European Union—NextGenerationEU.

in other Western countries is legally protected as free speech under the First Amendment, as long as it is not intended to directly incite violence (American Bar Association 2017). On the other hand, European countries do have laws against abusive language on online platforms, but so far there is no established mechanism to remove such content. Yet, European countries rely on voluntary cooperation and transparency on the part of online networks. In 2016, Facebook, Microsoft, Twitter, and YouTube signed a voluntary code of conduct against illegal hate speech. Over the years, other companies followed (European Commission 2020). Meanwhile, the Federal Government of Germany has taken it a step further, passing a law in 2021 that establishes new rules and harsher penalties for the accused. Furthermore, starting in February 2022, social networks will not only have to delete threats of murder, rape, and other serious hate crimes but also report them to the Federal Criminal Police Office. Even with a solid legal basis, removing abusive social media content will remain a tedious and time-intensive task if done manually, which is why the development of automatic detection systems would be desirable. From a research perspective, the classification of abusive social media content can be solved automatically with the help of machine learning. The basis for successful development of machine learning models is the availability of consistently labeled training data. Hereby, a crucial barrier is the diversity of terms and definitions of abusive language, for which related terms are used interchangeably. For instance, Davidson *et al.* (2017) discuss the differences between hate speech and offensive language, stating that the two terms are often conflated. Furthermore, popular social media platforms such as Facebook, YouTube, and Twitter support critical differences in definitions of abusive language (Fortuna and Nunes 2018). Different definitions and indistinct assessments between related terms of abusive language make it difficult to create uniformly annotated datasets. Consequently, they cannot be shared or harmonized in order to develop machine learning models that generalize well on different data sources.

Effective identification of abusive language using machine learning methods depends on the dataset and its annotation scheme. In this work, we analyze the annotation quality of five English and four German datasets designed for the detection of abusive language. We show that there are major differences between the definitions of abusive language, the annotation scheme, and we identify common issues in the annotation process. To support our findings, we provide examples from the datasets. For completeness, we perform experiments on three German datasets showing that even minor differences across datasets hinder generalization.

The remainder of this paper is organized as follows. Section 2 discusses the current state of research. Then, we provide a detailed overview of the datasets in Section 3, while we show the findings of our analysis of a sample of the datasets in Section 4. In Section 5, we present our experimental setup and results. Finally, we conclude in Section 6.

2. Related work

In the following, we present related work on abusive language detection which points out the challenges encountered from both theoretical and practical perspectives.

2.1 Analytical work

Vidgen *et al.* (2019) describe significant challenges and unaddressed limitations, and how these issues constrain abusive content detection research and limit its impact on the development of real-world detection systems. The authors identify two major challenges. First, there is the research challenge which consists of three parts. Part one concerns the categorization of abusive content. For one, there should be more clarity in subtasks. Research that purports to address the generic task of abuse detection is actually addressing something much more specific. This can often be seen in the datasets, which may contain systematic biases toward certain types and targets of abuse. Hence, the authors propose that subtasks are further disambiguated by the direction

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

of abuse. Another issue is clarity of terminology. Clarifying terminology will help delineate the scope and goals of research and enable better communication and collaboration. Some of the main problems are (i) researchers use terms that are not well-defined, (ii) different concepts and terms are used across the field for similar work, and (iii) the terms used are theoretically problematic. In particular, three aspects of the existing terminology have significant social scientific limitations: speaker's intent, effect of abuse, and audience sensitivity. The authors advocate ignoring the speaker's intent, not conflating two distinct types of abuse in terms of effect, and not making strong assumptions about the audience. The second part deals with the recognition of abusive content. The authors identify five linguistic difficulties that complicate the detection of abusive content: humor/irony/sarcasm, spelling variants, polysemy, extensive dependencies, and language changes. The last part concerns the context, which must be taken into account. Meaning is inherently ambiguous and depends on the subjective perspective of both the speaker and the audience, as well as the specific situation and power dynamics. The second challenge concerns the community. Creating appropriate datasets for training abusive language detection systems is an important but time-consuming task. Currently available datasets have several limitations. For many datasets, including those from Twitter, content cannot be shared directly, but IDs are shared and the dataset is recreated each time. This can lead to significant degradation in the quality of datasets over time. To address this issue, the authors suggest working more with online platforms to make datasets available. Furthermore, annotation is a notoriously difficult task, which is reflected in the low level of inter-annotator agreement reported in most publications, especially for more complex multi-class tasks. Few publications provide details on their annotation process or annotation guidelines. Providing such information is the norm in social scientific research and is considered integral to verifying others' findings and robustness. Another issue is the quality, size, and class balance of datasets, which varies considerably. Understanding the decisions underlying the creation of datasets is crucial for identifying the biases and limitations of systems trained on them. The authors propose that datasets could be curated to include linguistically difficult instances, as well as "edge cases": content that is non-abusive but very similar to abuse. The authors note that most research on abusive content detection focuses on text, but not on multimedia content. Additionally, the implementation of a system should be fair, explainable, and efficient. Finally, researchers should ensure that systems can be used in a variety of settings.

In a comprehensive study, Vidgen and Derczynski (2020) analyze 63 datasets designed for the development of abusive language classifiers. The subjects of the analysis are the classification tasks and corresponding target taxonomies of the datasets, the contents and annotations schemes of the datasets, as well as the source of the data and the motivations for creating datasets. Furthermore, the authors present papers that address the broad spectrum of online abusive language detection and its associated challenges, ranging from data collection, bias in the training data, varying definitions of abusive language, different target labels, the role of online platforms, the need for machine learning support, and annotation methods. In addition to the contributions of their paper, the authors are creating and maintaining the website https://hatespeechdata.com, which provides basic information about the datasets, such as a link to the dataset and the corresponding publication, size and language of the data, and the source of the content. The website is intended as a starting point for a data-sharing platform in this domain. The analysis shows, among other things, that datasets of abusive language come from heterogeneous sources (YouTube, Twitter, Wikipedia, Facebook), that they aim at different target classifications, that the level of expertise of the annotators varies, and that annotation schemes are often not transparent or loose. In addition to a comprehensive analysis of datasets for abusive language detection and a website to promote data sharing, the authors derive guidelines for creating new datasets for this purpose. The guidelines include instructions for the task, the actual dataset creation, the annotation process, and the documentation of best practices.

In a systematic review, Poletto *et al.* (2021) analyze 64 datasets including eight dictionaries for hate speech detection and point out their heterogeneous character. In particular, they try to

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

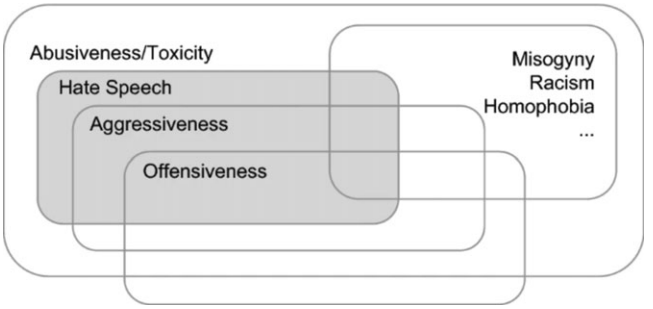

Fig. 1: Relations and boundaries between hate speech and related concepts according to Poletto *et al.* (2021).

illustrate the different concepts related to hate speech in the literature and datasets. Figure 1 (Poletto *et al.* 2021) shows both the boundaries and the relationship of inclusion and exclusion between hate speech and the various related terms. One of the main goals of the study is to create a method for developing robust hate speech detection resources. Therefore, they investigate the resources found in the research papers according to five different dimensions: type, topical focus, data source, annotation, and language. Among these, topical focus and annotation make hate speech recognition confusing and difficult. As noted in Figure 1, they point out that the boundaries between hate speech and other concepts overlap and each concept can be interpreted differently depending on cultural background and individual perception. Moreover, they note that some papers do not even provide a clear definition of the investigated concepts. Besides the topical focus, the annotation scheme and the skills of the annotators are essential aspects for the creation of suitable datasets. But annotation schemes are often not clearly described by the dataset creators. Annotations can be done either by experts, non-experts like crowdsourcing, or automated classifiers. Organizers often use a measure of inter-annotator agreement, with values ranging from very high (Cohen's $\kappa = 0.982$) to extremely low (Krippendorff's $\alpha = 0.18$) implying the difficulty of the classification task. They also argue that the collected data are not representative due to the collection methodology. Most data collections are made up by keyword searches on social media, especially Twitter. In their experiment with the English HatEval dataset, the evaluation showed that this leads to biased data collection. On the one hand, there are unclear boundaries regarding hate speech and its related concepts. On the other hand, the datasets are biased and sometimes too specific, which can lead to overfitting. These two issues are the main problems why an optimal result cannot be achieved in automatic abusive language detection. To overcome the challenges related to biased datasets, they suggest an in-depth error analysis that investigates the results of systems trained on a given dataset. For shared tasks in particular, they emphasize that thorough error analysis by the organizers helps to identify critical problems in the dataset scheme and annotation.

Although the main contribution of Risch *et al.* (2021) is a software tool that provides easy access to many individual toxic comment datasets, the authors also analyze the datasets. For the creation of the collection, the authors consider all publicly accessible comment datasets that contain labels that are subclasses of toxicity (e.g., offensive language, abusive language, and aggression). The broad definition of toxicity as a higher-level concept builds a bridge between the different lower-level concepts (e.g., sexism, aggression, and hate). During collection, the different dataset formats are converted into the same standardized csv format. Overall, the collection contains comments in thirteen different languages—obtained from twelve platforms—with 162 distinct class labels. The total number of samples is currently 812,993, but according to the authors, they are constantly adding more datasets. It is noteworthy that the collection does not contain the datasets themselves, but that the tool retrieves these datasets from their source and converts them into the unified format. According to the authors, this gives easy access to a large number of datasets and the ability to

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

filter by language, platform, and class label. The authors also address the problem of class imbalance in many datasets. While most datasets show a bias toward "normal" comments, the dataset by Davidson *et al.* (2017) shows a bias toward "offensive" comments. However, the latter class distribution is not representative of the underlying data in general. Additionally, some datasets are collected by filtering comments by a list of keywords, for example, "muslim," "refugee," or hashtags (#banislam, #refugeesnotwelcome). But such filtering introduces a strong bias because all hateful tweets in the created dataset contain at least one of the keywords or hashtags. Hence, the data are not a representative sample of all hateful comments, and models trained on this data might overfit the list of keywords and hashtags.

In contrast to the presented works, our contribution is based on participation in shared tasks and the practical challenges encountered. While we reach the same conclusions in many aspects (inadequate annotation process, heterogeneous motivations, data sources, and content), we provide a deeper analysis on a smaller selection of datasets. This includes the quantitative analysis of nine datasets where we present key data including data size, data collection, annotation guidelines, and best system performance. For the qualitative analysis, we manually inspect a sample of each dataset and present common categories of controversial annotations.

2.2 Experimental work

Arango *et al.* (2019) closely examine the experimental methodology used in previous work and find evidence for methodological problems by analyzing two models that performed best on the dataset of Davidson *et al.* (2017). One of the experiments conducted by Arango *et al.* (2019) tests the generalization to other datasets. For this purpose, they train both models of Badjatiya *et al.* (2017) and Agrawal and Awekar (2018) on the complete dataset of Waseem and Hovy (2016) and use the HatEval dataset (Basile *et al.* 2019) as test set. Both models achieve a macro-averaged F1 score below 50%, indicating that generalization to other datasets is not given. Another issue is the impact of user distribution on model generalization. For example, in the dataset by Waseem and Hovy (2016), one user is responsible for 44% of all "sexist" comments, while another user is responsible for 96% of "racist" comments. The authors hypothesize that user distribution is another source of overestimation of the performance of state-of-the-art classifiers. To prove this, they conduct another set of experiments in which the authors ensure that posts from the same user are not included in both training and test sets. This leads to a further drop in F1 scores for both models, and Arango *et al.* (2019) conclude that the models are susceptible to what they call "user-overfitting," that is, the models learn to discriminate between users rather than comment content. Indeed, Arango *et al.* (2019)'s results help to understand why state-of-the-art models that show impressive performance in their original test sets do not generalize well to other datasets, even when new data come from the same domain. They also conduct experiments aimed at alleviating state-of-the-art problems and improving model generalization. Their proposed solutions show improved performance across datasets and better estimation of model performance, leading to a more accurate state-of-the-art. In addition, they conclude that it is important to pay more attention to experimental evaluation and generalization of existing methods. In particular, the authors believe that there is an urgent need to obtain more information about the distribution of users in existing datasets or, even better, to create datasets that do not contain significant user biases.

Moreover, MacAvaney *et al.* (2019) identify and explore the challenges faced by automated online approaches to hate speech detection in texts. The authors identify the varying definitions of hate speech as one problem. Another problem is the limitation of publicly available data for training and testing. Many of the available datasets come from Twitter, but because of the character limit, these texts are short and concise. Therefore, models trained on such data cannot be generalized to texts from other sources. In addition, there are not many datasets that more accurately distinguish between hateful, aggressive, sexist, and similar content. Furthermore, datasets vary significantly in terms of their size and the characteristics of the hate speech considered. The

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

authors use a multi-view SVM model and evaluate its performance on four datasets. Their experiments show that this approach can outperform existing systems but the authors also note there is need for more research on both technical and practical aspects given the remaining challenges.

Kovács *et al.* (2021) find the automatic detection on social media posts difficult because they contain paralinguistic signals (e.g., emoticons and hashtags), and their linguistic content contains a lot of poorly written text. The authors encounter several challenges. First, there is word obfuscation, which is the intentional or unintentional misspelling of words. In addition, there are expressions that are considered offensive only in context. Another challenge is the use of ethnic slurs by in-group and out-of-group speakers, as the former are not considered offensive. Furthermore, the authors state that there is no universally accepted definition of hate speech and that even people have difficulty annotating hate speech using annotation guidelines. In their experiments, the authors study the effect of leveraging additional data on the HASOC dataset. They use several classical machine learning methods as well as different deep learning models. Their results show that additional data is useful in most cases, even if the data comes from another collection.

3. Annotated datasets

To identify the following datasets, we searched for shared tasks, read additional literature on shared tasks, searched the Internet for keywords, and studied https://hatespeechdata.com/. Finally, we chose to use datasets from shared tasks because they have been used by many researchers to develop relevant machine learning applications. Additionally, we chose HateBaseTwitter (Davidson *et al.*, 2017) because this dataset has an interesting feature: it contains more abusive than normal content.

In the following sections, we analyze five English and four German datasets. We describe the tasks and the systems that achieved the best results, how the data were collected, and what annotation guidelines were given. Please note that we report the best systems only for subtask A for each dataset, since we focus in our analysis only on these coarse-grained annotations (see Section 4).

3.1 OLID

Zampieri *et al.* (2019b) present the Offensive Language Identification Dataset (OLID) as part of the SemEval 2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval).

Subtasks. As described by Zampieri *et al.* (2019b), the OLID is annotated for three subtasks. Subtask A deals with the classification of offensive ("OFF") and non-offensive ("NOT") tweets, subtask B with classifying the type of offensive content (targeted vs. untargeted), and subtask C with identifying the target of the tweet. In the OffensEval shared task, offensive language includes both covert and direct targeted offense, such as insults and threats, as well as any profane language. Zampieri *et al.* (2019b) reported that the best model for subtask A was a BERT-based uncased model with standard hyperparameters, a maximum sequence length of 64, and trained with two epochs developed by Liu *et al.* (2019).

Data collection. Zampieri *et al.* (2019a) collected tweets using the Twitter API. They started with a trial annotation of 300 tweets extracted with 13 keywords/filters potentially related to offensive content. Three of these were removed for final annotation because the amount of offensive content was low. Most keywords/filters were either politically motivated or contained typical phrases of direct speech (e.g., "you are"). The highest percentage of offensive content was found by filtering tweets that were marked as "not safe" by Twitter. The trial annotation was used to test the annotation scheme and create a gold standard for the final annotation. URLs and user mentions in the dataset are masked (Zampieri *et al.* 2019a).

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

Annotation guidelines. As described in Zampieri *et al.* (2019a), the dataset was created following a three-level annotation scheme:

- Identification of offensive language
- Type of offensive language (targeted/untargeted)
- Target (individual/group/other)

Here, offensive language is defined as any threat, insult, or profanity, both implicit and explicit. Offensive language is further distinguished between those that target individuals or groups and those that do not. In step three, the targeted offensive tweets are classified by target type, that is, either individuals, groups, or other targets such as institutions. Annotation was done via crowdsourcing on the Figure Eight platform. Experienced annotators are tested before accepting their annotations, and each tweet was labeled by a majority vote of multiple annotators. The level of agreement among annotators on 21 tweets in the trial set was relatively high (Fleiss' kappa coefficient of 0.83).

3.2 HatEval

The shared task HatEval describes the detection of hate speech against immigrants and women in Spanish and English tweets extracted from Twitter at SemEval 2019 (Basile *et al.*, 2019).

Subtasks. Subtask A is a coarse-grained binary classification for the detection of hate speech. Hereby, hate speech has to be targeted at either women or immigrants. Subtask B is a fine-grained binary classification regarding both the aggressive attitude and the harassed target of a tweet classified as hate speech by subtask A. Indurthi *et al.* (2019) have the best-performing system for subtask A using a SVM model with RFB kernel without the need of external data, exploiting sentence embeddings from Google's Universal Sentence Encoder as features, achieving the macro-averaged F1 score of 65.10%.

Data collection. The organizers collected tweets in the period from July to September 2018, except the data where women are targeted. The largest part of the training set targeting women was taken from a previous collection. To collect the data, the following approaches were applied: (i) monitoring potential victims of hate accounts, (ii) downloading the history of identified haters, and (iii) filtering Twitter streams with keywords. For keywords, the organizers used both neutral keywords, derogatory words against the targets, and highly polarizing hashtags to collect a corpus that also reflected the subtle but important differences between hate speech, offensiveness, and stance (Basile *et al.*, 2019).

Annotation guidelines. The annotators were given a set of guidelines by the organizers, including the definition for hate speech toward the two targets. For the immigrant target group, the annotators are asked to mark a particular tweet as hate speech only if two questions are answered with yes: (i) the content of the tweet must have immigrants/refugees as main target, or even a single person but considered due to his/her membership in that category (rather than individual characteristics), and (ii) the given tweet must deal with a message that spreads, incites, promotes, or justifies hatred or violence toward the target, or a message that aims at dehumanizing, hurting, or intimidating the target group. For target group of women, annotators should decide whether a particular tweet is misogynous or not. A tweet is classified as misogynous if it specifically expresses women hating in the form of insults, sexual harassment, threats of violence, stereotyping, objectification, and negation of male responsibility. The annotation was done by untrained contributors on the crowdsourcing platform Figure Eight Basile *et al.* (2019).

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

The main contribution and challenge of the HatEval task is the limited scope of target of hate speech. For example, the following domains are not considered as hate speech in this task:[a]

- hate speech against other targets
- offensive language
- blasphemy
- historical denial
- overt incitement to terrorism
- offense toward public servants and police officers
- defamation

3.3 TRAC-2020

The shared task on Aggression and Gendered Aggression Identification was organized as part of the Second Workshop on Trolling, Aggression, and Cyberbullying (TRAC-2) at the 12th Language Resources and Evaluation Conference (LREC 2020; Kumar *et al.* (2020)).

Subtasks. Based on a dataset by Bhattacharya *et al.* (2020) in Hindi, Bengali, and code-mixed English, participants had to classify YouTube comments according to aggression types, which were defined in two subtasks. Subtask A required classification into three levels of aggression: "Overtly Aggressive" (OAG), "Covertly Aggressive" (CAG), and "Non-aggressive" (NAG). Subtask B required classification into gendered aggression and non-gendered aggression. In subtask A, the best model was a collection of fine-tuned BERT models that were combined into a bagging classifier (Risch and Krestel 2020). The ensemble method reduced the high variance of individual, complex models (5% for BERT in the TRAC-2020 shared task) on small datasets. The developers used up to 15 BERT models with the same hyperparameters and trained each with differently seeded training data and weight initialization in the classification layer. The final prediction for a sample was decided upon the highest sum of probabilities per class. The performance increased by 2% compared to the individual BERT models. The implemented BERT model is a BERT base model with a batch size of 48, a learning rate of $5e^{-5}$, and a maximum sequence length of 200. For preprocessing, the developers inserted whitespaces around emojis so that BERT would tag them as several "unknown" words rather than just one. For regularization, a 10% dropout was applied at the word embedding level.

The information on the data collection process and annotation guidelines given in the following two subsections was retrieved from Bhattacharya *et al.* (2020).

Data collection. The collection of YouTube comments was limited to "conversations" with at least 100–150 comments, as a high number of comments potentially increases the occurrence of aggressive content and indicates higher user engagement. Furthermore, only YouTube videos on specific topics where misogynistic comments were suspected were selected.

Annotation guidelines. The dataset was annotated at the comment level using a two-level annotation scheme:

- Aggression level (Overtly/covertly/non-aggressive)
- Misogyny (gendered/non-gendered)

The authors defined misogyny as aggression directed against stereotypical gender roles and sexuality. They instructed the annotators to pay particular attention to comments that accept,

[a]See annotation guidelines at https://github.com/msang/hateval/blob/master/annotation_guidelines.md

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

endorse, or propagate bias toward these topics. There were no strict annotation guidelines; instead, annotators were encouraged to share their perspectives in personal discussions. A total of four annotators were involved in the creation of the dataset, with each instance annotated by two annotators and disagreements resolved by a third annotator and internal discussions. All annotators had undergraduate or graduate degrees in linguistics. Disagreements among annotators were resolved using the counterexample method, majority staff votes, and discussions. The inter-annotator agreement was measured with a Krippendorff's alpha coefficient of 0.75.

3.4 HateBaseTwitter

Davidson *et al.* (2017) present their approach on automated hate speech detection while addressing the problem of offensive data. Unlike other work, the authors do not conflate hate speech and offensive speech, but distinguish between the two. The corpus was compiled by the authors themselves.

Task. Their task is a tertiary classification of "HATE," "OFFENSIVE," and "NEITHER." As a final model, the authors used logistic regression with L2 regularization. Here, a separate classifier is trained for each class, and each tweet is assigned the class label with the highest prediction probability of all classifiers. The best-performing model has an overall precision of 0.91, recall of 0.90, and F1 score of 0.90.

Data collection. Using the Twitter API, the authors searched for tweets that contained specific words and phrases from a lexicon. This lexicon was compiled from words and phrases identified as hate speech by `Hatebase.org`. For the resulting tweets of 33,458 Twitter users, they extracted each user's timeline and then randomly sampled 25,000 tweets.

Annotation guidelines. Davidson *et al.* (2017) "[. . .] define hate speech as language that is used to express hatred toward a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group. In extreme cases, this may also be language that threatens or incites violence, but limiting our definition only to such cases would exclude a large proportion of hate speech." Importantly—and quite interestingly, in our opinion—the "[. . .] definition does not include all instances of offensive language because people often use terms that are highly offensive to certain groups but in a qualitatively different manner." The exact definitions given to the annotators are not disclosed, but the authors explain that "they were provided with our definition along with a paragraph explaining it in further detail." The annotations were created manually by CrowdFlower staff. The authors instructed the annotators to judge a post not only by the words it contained, but also to consider the context in which they were used. Furthermore, they were instructed that the presence of a particular word—however offensive—did not necessarily mean that a tweet was hate speech. Ultimately, each tweet was annotated by at least three people, and the intercoder agreement score determined by CrowdFlower is 92%. The majority vote for each tweet was used to assign a label. Interestingly, some tweets were not assigned labels because no majority vote could be reached.

3.5 HASOC

Mandl *et al.* (2019) describe the identification of hate speech and offensive content in Indo-European languages, that is, English, Hindi, and German, at FIRE'19 (Majumder *et al.* 2019).

Subtasks. There were three subtasks: Subtask A was the coarse-grained binary classification into Hate & Offensive ("HOF") and regular content ("NOT"). If a post was classified as "HOF," it was processed further in subtask B, where a fine-grained classification into either hate speech, offensive, or profane had to be made. Subtask C dealt with the targeting or non-targeting of individuals,

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

groups, or others when a post was classified as "HOF." For English, Wang *et al.* (2019) had the best performance with a macro F1 score of 78.82% on subtask A. The model used was an Ordered Neurons LSTM with an attention layer. The authors used *k*-fold cross-validation and accumulation averaging to obtain the final output from the *k* results. The authors used the OLID dataset (Zampieri *et al.* 2019a) as additional data. Furthermore, the authors generated a development set from HASOC and OLID and ensured that it contained a 1:1 ratio of "HOF" and "NOT" examples. On the German dataset, the best performance was obtained with a macro F1 score of 61.62% by Saha *et al.* (2019). Here, a Light Gradient Boosting Machine was used as model. As input, the authors compute the sentence embeddings from BERT and LASER and feed the model with the concatenated sequences.

Data collection. Mandl *et al.* (2019) used heuristics to search for typical hate speech in social media platforms, thereby identifying topics where many hate posts are expected. Different hashtags and keywords were used for all three languages. The authors recorded the ID for some users and collected the timeline for some of these users. For diversity reasons, the most recent posts of the users were crawled. The HASOC dataset was then sampled from Twitter and partially from Facebook for all three languages using hashtags and keywords that contained offensive content. For the annotations, the authors relied on several juniors for each language.

Annotation guidelines. Annotators were given the following brief guidelines.

- Hate Speech: Attributing negative characteristics or defects to groups of individuals (e.g., all poor people are stupid). Hateful comments toward groups based on race, political opinion, sexual orientation, gender, social status, health condition, or similar.
- Offensive: Degrading, dehumanizing, or insulting a person. Threat of violence.
- Profanity: Unacceptable language in the absence of insults and abuse. This usually involves the use of swear words ("shit, fuck," etc.) and cursing ("Go to hell! Damn!" etc.).
- Other: Normal content, statements, or anything else. If the utterances are considered to be "normal" and not offensive to anyone, they should not be labeled. This could be part of youth language or other language registers.

Mandl *et al.* (2019) point out that the annotation process is highly subjective and that even after discussions of controversial posts, agreement was often not reached. The authors used a subset for which two annotations were obtained and calculated the inter-annotator agreement, resulting in $\kappa = 0.36$ for English and $\kappa = 0.43$ for German.

3.6 GermEval 2018

Wiegand *et al.* (2018) present the pilot edition of the "GermEval Shared Task on the Identification of Offensive Language." This shared task deals with the classification of German tweets from Twitter.

Subtasks. Task 1 was coarse-grained binary classification into two classes, "OFFENSE" and "OTHER." Task 2 was a fine-grained classification into four classes including the class "OTHER." The class "OFFENSE" identified by Task 1 was further divided into three fine-grained classes "PROFANITY," "INSULT," and "ABUSE." The best-performing system by Montani and Schüller (2018) employed an ensemble classification model of traditional feature-based supervised learning and achieved an F1 score of 76.77%.

Data collection. According to the organizers, they paid great attention to data collection. The data were not collected on the basis of queries. The organizers considered tweets from about 100 different users and their timelines. The data collected should meet the following criteria:

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

- Each tweet had to be written in German.
- Each tweet had to contain at least five ordinary alphabetic tokens.
- No tweet was allowed to contain any URLs.
- No tweet was allowed to be a retweet.

Annotation guidelines. According to the annotation guidelines (Ruppenhofer *et al.* 2018), the class "OFFENSE" includes three subclasses: "INSULT," "ABUSE,", and "PROFANITY." "INSULT" is the attribution of negatively evaluated qualities or defects or the labeling of persons as unworthy or worthless, while "ABUSE" is a special kind of disparagement. "PROFANITY" is assigned when an utterance uses unacceptable language, even if it does not imply insult or abuse. Although more feature annotations are related to the task, the following labels are not considered in the shared tasks: Ephithets, typical targets of abusive speech such as feminists, black people, Muslims, refugees, etc., stereotypes, curses, threats, calls to action, disguises. Annotators are asked to tag tweets if they are unsure of which label to choose. In addition, the annotators are instructed that . . .

- they should not annotate utterances if they are so unacceptable that they should be automatically filtered.
- abusive language and insults can be directed against people, whether they are dead or alive.
- an utterance is considered an insult or abuse regardless of whether it was reported by the person or group of people against whom it is directed.
- they should always carefully consider the given context of a potentially offensive utterance.
- abusive language may also occur in sentence types other than plain assertion.

The three organizers annotated the tweets themselves. They are all German native speakers. To ensure a certain quality level of their annotations, 300 tweets were selected and annotated by all three annotators to measure their agreement. They removed all tweets that were annotated with "HUNH" (i.e., tweet is incomprehensible to the annotator) or with "EXEMPT" (i.e., retweet) by at least one annotator. For the remaining 240 tweets, an agreement of $\kappa = 0.66$ was measured. All other tweets were annotated by one annotator.

3.7 GermEval 2019

Following GermEval 2018, Struß *et al.* (2019) presented the second edition of the "GermEval Shared Task on the Identification of Offensive Language." The topics and political orientation were expanded to include left-wing extremism, right-wing extremism, and antisemitism.

Subtasks. Besides subtask A for coarse-grained binary classification and subtask B for fine-grained classification, subtask C described the explicitness of "OFFENSE." According to Struß *et al.* (2019), most of the better-performing systems employed some form of the transformer-based language model BERT. For example, the best-performing system by Paraschiv and Cercel (2019) used a BERT model and achieved an F1 score of 76.95% for subtask A. Word embeddings were pre-trained using 6 million German tweets.

Data collection. Most of the training data for GermEval 2019 consisted of the collection from the previous year. In addition, the organizers collected new data by heuristically identifying users who regularly post offensive tweets. While the data from GermEval 2018 mainly consisted of the far-right spectrum and the dominant theme of migration, the organizers added timelines of users from the far-left spectrum, and antisemitism to increase theme variance. Although the collected data represent certain political parties and some of their representatives, the organizers tried to

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

collect the data across the political spectrum. They also took care to split the data into training and test sets, making sure that a user's complete tweets were not assigned to either the training set or test set, and that training and test sets had a balanced distribution.

Annotation guidelines. For subtasks A and B, the same annotation guidelines were given as in the previous year. For subtask C, they define implicit offensive language as a form of offensive language in which an explicitly or implicitly given target has to be inferred from the attribution of (hypothetical) target properties that are insulting, degrading, offending, humiliating, and so on. Offensive tweets that use figurative language such as irony or sarcasm, or punning, are also considered implicit. Additionally, inappropriate slang that addresses a serious topic is subsumed under implicit offensive language.

As for GermEval 2018, the annotations were done by the organizers. Again, the organizers sampled 300 tweets and annotated them for quality measurement. As in the year before, tweets that were annotated as "HUNH" and "EXEMPT" were removed. For the remaining 206 tweets, an agreement of $\kappa = 0.59$ was measured. It can be considered moderate. All other tweets were annotated by one of the four annotators.

3.8 GermEval 2021

The GermEval 2021 workshop on "Identification of Toxic, Engaging, and Fact-Claiming Comments" was organized as part of KONVENS 2021 (Conference on Natural Language Processing). The following descriptions are either taken from the official workshop website,[b] or from our own dataset analysis unless otherwise noted.

GermEval workshops are usually organized informally by groups of interested researchers. The GermEval 2021 workshop is endorsed by the IGGSA (Interest Group on German Sentiment Analysis), a special interest group within the GSCL (German Society for Computational Linguistics).

Subtasks. GermEval 2021 consists of three binary classification tasks: Classification of toxic, engaging, and fact-claiming comments; (i) subtask 1: binary classification of toxic and non-toxic posts, (ii) subtask 2: binary classification of engaging and non-engaging posts, and (iii) subtask 3: binary classification of fact-claiming and non-fact-claiming posts. Here, the term "toxic" is not further specified, but the organizers point out that its definition is the same as in previous GermEval workshops. In this context, GermEval 2019 defined offensive language as insulting, profane, or abusive (Struß *et al.*, 2019). "Engaging" is defined as respectful, rational, and reciprocal posts that can lead to better quality and quantity of discussion. The "fact-claiming" subtask aims to identify potential fake news in order to initiate a manual fact checking.

Data collection. The collected data come from the Facebook page of a political talk show of a German TV station. The organizers collected discussions between February and July 2019. The host, the users, and the show itself were anonymized by masked tokens for the respective text mentions. The test and training data were taken from comments on different broadcasts to avoid thematic bias. More detailed information about the data collection process was not disclosed at this point of time. User mentions are masked in the dataset.

Annotation guidelines. The annotation scheme is based on an excerpt from a preliminary codebook (Wilms *et al.* 2021). The codebook served as guide for the annotators, who additionally participated in a coding training. The data source is a Facebook page with a principal setup: there is a top-level post by the talk show staff. Users refer to these posts by writing comments below

[b] https://germeval2021toxic.github.io/SharedTask/

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

Table 1. Overview of datasets used and their corresponding sizes. The percentage of abusive content is calculated by considering all types of abuse (e.g., hate, offense, aggression)

Language	Dataset	# Train	# Test	% Abusive	Type
English	OLID	13,240	860	0.33	Offense
	HatEval	9,000	3,000	0.40	Hate
	TRAC-2020	5,354	1,175	0.25	Aggression
	HateBaseTwitter	24,802	—	0.83	Hate & Offense
	HASOC	5,852	1,153	0.40	Hate & Offense
German	GermEval 2018	5,009	3,532	0.34	Offense
	GermEval 2019	3,995	3,031	0.32	Offense
	GermEval 2021	3,245	944	0.35	Toxic
	HASOC	3,819	850	0.24	Hate & Offense

them. There can also be comments about comments. The individual comments are the only text sample that goes into the machine learning models. However, the annotators are asked to also consider the staff's posts and other higher-level comments when evaluating a comment. This means that the context of a comment played an important role in the annotation process. The comment was then evaluated according to the following characteristics:

- shouting: indicated by words with capital letters and multiple punctuation marks
- profanity
- insult
- sarcasm/cynicism: when used to disparage the reference object
- discrimination
- denigration
- accusation of lying or deception

If any of these characteristics applied, the annotators were asked to mark the corresponding comment as "TOXIC." Detailed information on each feature was provided to the annotators and is listed in the codebook, which can be obtained from the authors.

3.9 Summary of datasets

In Table 1, we provide an overview of the datasets by specifying the name, language, number of training and test samples, percentage of abusive samples, and type of abusive content. The datasets vary significantly in size, which risks overfitting deep learning models on the smaller datasets. The percentage of abusive samples is calculated by dividing the number of abusive samples by all samples. For the HateBaseTwitter dataset, this was done by combining "HATE" and "OFFENSIVE," and for the TRAC-2020 dataset, this was calculated by combining "OAG" and "CAG." The percentage of "HATE" is 6%, that of "OFFENSIVE" is 77%, that of "OAG" is 12.78%, and that of "CAG" is 12.16%. The resulting numbers show that most of the datasets are imbalanced.

From the information in the previous sections, the values for the inter-annotator agreement are not very high in most of the datasets. The degree of agreement for English and German HASOC is minimal to weak, while it is weak to moderate on GermEval 2018 and GermEval 2019. The level of agreement for HatEval is considered high, but was calculated using only 21 tweets, so this value is not statistically significant. The value for TRAC-2020 is moderate, while the authors give a very

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

Table 2. Overview of the systems that achieved the best results on each dataset

Language	Dataset	Best System	Macro F1
English	OLID	BERT (Liu *et al.* 2019)	82.90
	HatEval	SVM with RFB Kernel (Indurthi *et al.*, 2019)	65.10
	TRAC-2020	BERT ensemble learner (RischandKrestel 2020)	80.20
	HateBaseTwitter	Logistic Regression (Davidson *et al.*, 2017)	90.00
	HASOC	Ordered Neurons LSTM (Wang *et al.*, 2019)	78.82
German	GermEval 2018	Ensemble Classification (Montani and Schüller, 2018)	76.77
	GermEval 2019	BERT model (ParaschivandCercel 2019)	76.95
	GermEval 2021	BERT/ELECTRA ensemble (Bornheim *et al.* 2021)	71.75
	HASOC	Light Gradient Boosting Machine (Saha *et al.*, 2019)	61.62

high intercoder agreement score for HateBaseTwitter, without disclosing its calculation. These values show that people have difficulty agreeing on what constitutes abusive language.

In terms of data collection, only HateBaseTwitter and GermEval 2018 provided information on user distribution, that is, from how many different users' tweets/posts were collected. As pointed out by Arango *et al.* (2019), this is crucial information to ensure that the classification is not a distinction between users, but based on content. While the user distribution of HateBaseTwitter is very good, this is not the case for GermEval 2018 (only 100 users). Future dataset creators should use a broad base of users and also provide at least an internally generated ID for each user.

In terms of annotation guidelines, there is great variation. HatEval is limited to hatred against two specific targets, women and immigrants. OLID defines offensive language as threatening, insulting, or profane. This is almost the same as GermEval 2018 and GermEval 2019, whose definition of offensive includes insult, abuse, or profanity. There seems to be a discrepancy between OLID's threats and GermEval's abuse. HateBaseTwitter's definition seems to only include hate speech toward a group or member of a group. They also exclude instances of abusive language uttered by in-group speakers. A model without knowledge of a user's ethnic background cannot correctly distinguish these cases. HASOC's definition of hate speech also refers to negative characteristics toward a group or person of a group, while their definition of offensive language refers to individuals. Interestingly, the authors consider threats to be offensive but not hateful, which is in contrast to OLID, HatEval, and HateBaseTwitter. GermEval 2021 provides guidelines for toxic language, which seems to be a fairly broad definition covering six different topics. The most interesting case is TRAC-2020, where no strict guidelines were given. Despite the lack of clear instructions, a Krippendorff's $\alpha = 0.75$ was reached. One factor that many of the guidelines have in common is the instruction to pay attention to context when annotating. This means that when training models on such data, context should also be considered and that context should be provided in the dataset.

In Table 2, we present the best-performing systems for each dataset evaluated using the macro-averaged F1 score, with the exception of TRAC-2020, which was evaluated using the micro-averaged F1 score. As the scores suggest, the datasets differ in difficulty. While the classification of HatEval and the German HASOC appears to be difficult in terms of reported F1 scores, this is not true for OLID, TRAC-2020, and HateBaseTwitter. The scores for the English HASOC, GermEval 2018, and GermEval 2019 are in the middle range. It is also worth noting that different models were used. While traditional machine learning models were applied on the smaller datasets, deep learning models can be used on the larger datasets.

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

4. Analysis of the annotated datasets

In order to review the quality of the datasets described in Section 3, we manually examined a sample for each dataset. To this end, we randomly selected a total of 500 instances (250 abusive and 250 non-abusive) per dataset. When performing the analysis, we strictly followed the annotation guidelines of each dataset when available. First, each of the authors analyzed a subset of the samples. Then, the identified controversial annotations were discussed in a group session. Particular attention was paid to finding common annotation issues across the datasets in order to support improved guidelines for future annotation processes.

4.1 OLID

When we manually reviewed the data, we found a total of 68 controversial annotations, a rate of 13.6%. Many of them are insults that were marked as non-offensive. For instance, in example (1), the author calls someone stupid in a distasteful way and (2) suggests that the referred person will die because of his obesity. Although these examples are undoubtedly wrongly annotated, insults are generally very subjective. We identified a number of borderline cases that we could not label without a doubt, neither intuitively nor by following the annotation guidelines. The examples (3)–(5) illustrate this problem. Another common source of controversial annotations are tweets marked as offensive, which are not offensive in themselves, but in context. Although the OLID dataset does not contain contextual information, we identified some cases where we suspect that context "makes" the tweet offensive, for example, examples (6)–(8). In example (7) is "kids" used here as a pejorative? Or is this user more defending young people fighting for gun control?

(1) *@USER [brain emoji] damages she is*—NOT
(2) *@USER 2020? Michael Moore is a cheeseburger away from not making it to 2019!*—NOT
(3) *@USER @USER No he is in some other world he dreamed of when he was on drugs in college*—NOT
(4) *@USER Take a stand against all violence! Talk to your ANTIFA friends.*—NOT
(5) *@USER This all I get [sad smiley] damn these haters deleting you IG for no reason. . . URL*—OFF
(6) *@USER She should ask a few native Americans what their take on this is.*—OFF
(7) *@USER @USER @USER gun control! That is all these kids are asking for!*—OFF
(8) *@USER @USER What a waste of one's short life.[shrugging emoji][U.S. flag emoji]*—OFF

4.2 HatEval

From the random samples we examined, only 21 annotations are controversial (4.2%). These include thirteen posts targeted at "women" and eight directed at "immigrants." It is important to keep in mind that the tasks specifically target women or immigrants, and all other abusive posts targeting others are not classified as "hate speech" (HS). For example, the sentences in the examples (9) and 10 are not classified as "HS" because they are not targeted at women or immigrants. We assume that a model trained on such data does not generalize well to other target tasks related to abusive language. The examples (11) and (12) are labeled as "HS," but the annotator needs background knowledge and/or contexts to infer this classification. The examples (13) and (14) show contradictory cases. We found many cases where facts were asserted that would need to be verified to determine the correct class. If the claimed facts in example (15) are false, then this tweet is hateful; otherwise, it is not.

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

Table 3. Top five hashtags in HatEval

Hashtag	Hate Speech	NOT	Total
#buildthatwall	333	0	333
#maga	218	26	246
#buildthewall	201	12	217
#immigration	86	61	147
#trump	109	35	145

(9) *Im smack if out one of you pussy boys*—NOT

(10) *ALL YOU PUSSY BOYS FALL BACK!*—NOT

(11) *Let's build that wall.#BuildThatWall*—HS

(12) *Thank You #ICE !! #BuildThatWall !*—HS

(13) *Dat Whore Karma ah Bitch*—NOT

(14) *Ewk man this bitch be with mad whores*—HS

(15) *It is way overdue to admit Europe cannot take all of the migrants which are trying to get there. Nor can places like Canada. @Europarl_EN @PopulationIC @PopnMatters @BBCNews @dwnews @CBC Please look at this dramatic graphic demonstration... https://t.co/wZ3v1J3Uqh*—HS

During our analysis of this dataset, we noticed that many tweets did not contain any hateful content other than the hashtag content. For this reason, we extracted all hashtags from the whole dataset. The top five hashtags (lowercased) and their frequency per class are shown in Table 3. The values in Table 3 confirm our expectation that hashtags play an important role in classification. For example, the tweets with the hashtag #buildthatwall are all classified as HS, while the tweets with the more neutral hashtag #immigration compete for each class.

4.3 TRAC-2020

A total of 54 controversial annotations were found, resulting in an error rate of 10.8%. Due to the loose annotation guidelines, the difference between overtly aggressive (OAG) and covertly aggressive (CAG) is not further specified, leaving the judgment to the annotators' intuition. From our point of view, covertly aggressive means aggression expressed only through content, while overt aggression is enhanced by tone of voice, swear words, and aggressive emojis. Considering this definition, we identified many comments which should be labeled as OAG but were labeled as CAG, such as the examples (16)–(18). We also found many comments that are labeled as CAG even though they do not contain aggression, for example, the examples (19)–(22). This indicates that the CAG label was used in cases where the annotators were unsure whether the comment should be considered aggressive, rather than a label with clear characteristics. It may be necessary to define the annotation scheme more clearly and establish new labels as appropriate. As in the other datasets, we found many controversial annotations that lacked contextual information. In example (23), the author might be referring to another comment with inappropriate content—should the comment still be considered aggressive? In example (24), the annotators seemed to consider the comment as sarcastic, but depending on the context, it could have been a positive comment.

(16) *Arundathi roy has no right to live in India, if she instigates violence through spreading her dangerous ideology. We Indians should not fall for these kind of TRAITORS. PLEASE*

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

BEWARE OF THESE KIND OF PEOPLE. SPREAD AWARENESS. WE NEED TO STOP THESE NON SENSE.—NAG

(17) *We will not going to show any documents to govt.. We will see what they can do.. The mother f**ker BJP ?*—NAG

(18) *This is the fucking true which today's generation pretend. . .[angry emoji]*—NAG

(19) *@Rishita Pandey these guys don't know meaning of feminism. In simple words it means equality. We are not property of anyone. Thank you.*—CAG

(20) *You're watching the movie in a completely negative way. I get it that it does have a negative impact, but you're not giving a balanced analysis.*—CAG

(21) *worst video*—CAG

(22) *Feminist mean the equal right of women and men, why people take it wrong way*—CAG

(23) *Shame. . . . . .shame. . . . . .*—CAG

(24) *Brother, you're a Savior for this generation*—CAG

4.4 HateBaseTwitter

Overall, the random sample is well annotated. There are only a few samples that can be classified as "contradicting." Example (25) is labeled as "NEITHER" but contains a derogatory term for women. Other samples with the same word are labeled as "OFFENSIVE." The same is true for example (26) and example (27). While example (26) contains the same ethnic slur as example (27), the former example is labeled as "NEITHER," the latter is marked as "OFFENSIVE." Furthermore, we think that example (27) should rather be marked as "HATE" because it sounds like an insult to a certain group. The examples (28) and (29) contain hateful expressions that are not easy to understand without background knowledge, due to the terms "hopper" and "beaner."

(25) *You wanna find hoes on here? Just follow the chicks who reply with heart eyes under sex gifs*—NEITHER

(26) *Log off nigger RT @PoloKingBC: #relationshipgoals*—NEITHER

(27) *No way all u niggers are employees of the month*—OFFENSIVE

(28) *in real life and online I follow that. sorry folks but just cause I look like a border hopper don't mean I am one. born and raised here*—HATE

(29) *@HuffingtonPost im American, don't give 2 shits about the beaners, ship them beaners back*—HATE

Upon inspection, we noticed that almost every HOF instance contained an instance of a small set of abusive words. To validate our observation, we preprocessed the whole dataset by first locating all "OFFENSIVE" and "HATE" instances and removing stop words and numbers. We then computed the fifteen most frequent terms listed in Table 4. The frequencies of all abusive words amount to 27,465. Out of the 15 most frequent terms, eight are offensive/hateful. Therefore, models trained on such data cannot be generalized to texts from other sources because they will likely focus on these terms.

4.5 HASOC

German. In the German sample, we found ten cases of controversial annotations, a rate of 2%. In example (30), the author generalizes—based on one incident—that all Muslims are criminals. This should be annotated with "HOF" according to the guidelines. Example (31) contains a swear word and is not annotated as such, contradicting other tweets that contain the same word. Example (32)

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

Table 4. Fifteen most frequent terms (after preprocessing) in HateBaseTwitter

Top #1–#5	Count	Top #6–#10	Count	Top #11–#15	Count
bitch	11,414	pussy	2,240	lol	962
hoe	4,282	ass	1,573	faggot	829
nigger	3,202	got	1,471	know	720
fuck	2,643	get	1,289	love	616
like	2,477	shit	1,282	one	576

is labeled as "NOT," but this is only correct if there was a rioting incident. Otherwise, the author is trying to incite hatred against a group of people. Finally, example (33) is a case of missing context. It is not clear why this is considered "HOF."

(30) *Dafür werden die Moslems vom Steuerzahler alimentiert. Junge Männer verletzen Polizeibeamtin in Stendal . . .*—NOT "For this, the Muslims are alimented by the taxpayer. Young men injure (female) police officer in Stendal . . ."

(31) *Diejenigen, welche einen solchen Scheiss rauslassen könnte man noch verkraften. Aber dass man damit Millionen blinde Waehler findet ist das Schlimme und sollte zu denken geben.*—NOT "One can cope with those who talk such shit. But that this kind of talk gives you million of blind voter is bad and makes one pensive."

(32) *Oh nein, #Tatort wieder total realitätsfern und mit Erziehungsgedanken: "Reichsbürger"! Hey liebe @Tatort Autoren, könntet ihr mal bitte zum Hauptbahnhof fahren (wahlweise: #Gelsenkirchen), wo vorhin gerade Türken+Syrer randaliert haben, um mal echte deutsche Tatorte zu sehen?*—NOT "Oh no, #Tatort again totally unrealistic and with educational thoughts: "Reichsbürger"! Hey dear @Tatort authors, could you please go to the main station (alternatively: #Gelsenkirchen), where just now Turks + Syrians have rioted, to see real German crime scenes?"

(33) *Ich wusste nicht, dass #Mohammed ein Volk ist? #Volksverhetzung*—HOF "I didn't know #Mohammed is a nation? #Volksverhetzung"

English. In the English sample, we found 125 controversial annotations, a rate of 25%. Example (34) does not contain any swear words or other offensive/hateful content, but is labeled as "HOF." Example (35) does contain the same derogatory word as example (36), but the former is labeled as "HOF" while the latter is not. Similarly, the examples (37) and (38) contain the same offending hashtag but are marked differently. This type of contradictory annotation is quite common.

(34) *@labourpress Great parliamentary quotes of our times. #BorisJohnsonShouldNotBePM*—HOF

(35) *@GovHowardDean #JackieO was an American treasure who inspired millions of Americans and made us proud. Melania is a Russian whore. #Putin's left over. #TrumpIsATraitor #JohnMcCainDayJune14 @FLOTUS*—HOF

(36) *I've met #JackieO, you're no Jackie Kennedy. She had class and dignity. She married a war hero, not a coward or Soviet mole. She wasn't a #Russian whore. #TrumpIsATraitor @FLOTUS @CNN*—NOT

(37) *@lbcbreaking @chunkymark @LBC @dodgyknees1 really #dickhead*—HOF

(38) *@HamAnalysis Shame on him! #dickhead*—NOT

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

4.6 GermEval 2018

From the samples we manually reviewed, we found only four controversial annotations, a rate of 0.8%. The examples (39) and (40) involve a politician and are tagged as "OTHER," but we would classify them as abusive. The examples (41) and (42) are probably cases of "missing context" because they do not contain obvious abusive content.

(39) *Das EU Experiment soll einer kosmetischen Behandlung unterzogen werden, nur taugt die Kosmetikerin Merkel in der Wurzel nichts [Winking Face with Tongue] ÜBERFORDERT*—OTHER "The EU experiment is to be subjected to a cosmetic treatment, but the beautician Merkel is fundamentally useless [Winking Face with Tongue] OVERLOADED"

(40) *Merkel lernte u. studierte in der DDR für "Null" > ihre wichtigste Erkenntnis, halte deine Wähler DUMM, denn dann keine "dummen Gedanken" [Winking Face with Tongue]*—OTHER "Merkel learned and studied in the GDR for 'free' > her most important insight, keep your voters STUPID, because then no 'stupid thoughts' [Winking Face with Tongue]"

(41) *Liebe Freunde, viel zu lange geschwiegen!Was hat diese Person verbrochen?*—OFFENSE "Dear friends, silent for far too long! What has this person done wrong?"

(42) *@Sakoelabo @Padit1337 @SawsanChebli Du bist jung und schön*—OFFENSE "@Sakoelabo @Padit1337 @SawsanChebli You are young and beautiful"

4.7 GermEval 2019

From the sample, only eight posts are controversial, a rate of 1.6%. They are related to lack of context or the vagueness between factual aspect and abusive attitude. As discussed with other datasets, some posts cannot be clearly classified as such without the proper context. The example (43) is annotated as "OFFENSE," but it is difficult to understand this decision without the proper context. The word "Mist" (en: crap) could imply an offensive stance, but it is a very common word in German that can be used without insulting intent. Example (44) is marked "OTHER," but the person "Wagenknecht" and the party "Die Linke" are humiliated in this example. In example (45), the author considers #bolsonaro a fascist and the post is labeled as "OTHER." Even though Bolsonaro is often criticized for his authoritarian dictatorial tendencies, it is debatable whether this post should rather be labeled as "OFFENSE." In general, criticism of politicians is widely accepted, even if it is often abusive. Example (46) is classified as "OFFENSE," but if the claimed facts in this example are true, then this tweet is not offensive.

(43) *@maxihome27 für Ohne Worte . . . schreibste ganz schön viel . . .Mist! =)*—OFFENSE "@maxihome27 for without words . . . you write a lot . . . crap! =)"

(44) *@JaMa08768523 @b_riexinger @dieLinke . . . weil es die Wagenknecht-Partei ist. . .was erwartest du denn? [face with rolling eyes]*—OTHER "@JaMa08768523 @b_riexinger @dieLinke . . . because it's the Wagenknecht party . . . what do you expect? [face with rolling eyes]"

(45) *@SPIEGELONLINE #bolsonaro ist ein Faschist sehr geehrtes Nachrichtenmagazin*—OTHER "@SPIEGELONLINE #bolsonaro is a fascist dear news magazine"

(46) *Nicht nur Poroschenko verdient gut,auch Deutsche Politiker verdienen gut an das Schwarzgeschäft Rüstung*—OFFENSE "Not only Poroshenko earns well, German politicians also earn well from the black market of armaments"

4.8 GermEval 2021

When we manually reviewed the data annotations, we found a total of 41 controversial annotations, a rate of 8.2%. One problem with the annotation, which also occurs with other datasets, is

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

the lack of context. Specifically for the creation of the GermEval 2021 dataset, context is known to be considered in the annotation process, but it is not provided in the dataset. Many comments such as the examples (47)–(49) do not in themselves satisfy any of the features in the annotation guidelines. However, veiled toxic features such as discredit, discrimination, or sarcasm could apply in these cases if the context annotation and/or higher-level annotations are taken into account. This means that such samples have ambivalent meanings that lack crucial information, which will most likely "confuse" a classification model during training. The examples (50) and (51) show that there are borderline cases where the evaluation is subjective and labeling can easily become arbitrary. This starts getting an issue as soon as there are conflicting labels, as it can be seen in the examples (52) and (53). Example (52) was classified as "TOXIC," most likely due to the word "kotzen" (en: puke) being profane. Example (53) was classified as "NOT TOXIC," even though it contains the same word. In both examples, there are no other indicators for abusive language.

(47) *@USER Oh weh das möchten wir uns garnicht vorstellen.*—TOXIC "@USER Oh dear we do not want to imagine that at all"

(48) *@USER hat eine Ähnlichkeit mit ihnen*—NOT TOXIC "@USER has a resemblance with you"

(49) *Nach mir die Sintflut [Smiling emoji]*—TOXIC Difficult to translate. It is a mostly negatively connotated German proverb to express one's indifference about a situation.

(50) *@USER weil man nicht über Islmaistengewalt öffentlich reden will wächst die Afd!!*—TOXIC "@USER the Afd grows because no one publicly talks about Islamist violence!!"

(51) *Klatscht das Publikum für jeden Mist*—NOT TOXIC "The audience claps for every crap"

(52) *wo kommt man live in die Sendung mit den nachrichten !!!! ich könnt kotzen*—TOXIC "which show takes such news live on air !!!! this sucks"

(53) *Und wieder weg – es geht um diese linksgrünen und jedes 2 Wort AfD zum kotzen*—NOT TOXIC "And away again—it's about these left-green and every 2 word AfD this sucks"

4.9 Summary

The analysis revealed that the percentage of controversial annotations varies significantly from dataset to dataset. As a result of our analysis, we define the following main classes of controversial annotations: (i) *Contradictory:* content with the same abusive words is labeled differently, (ii) *missing context*: the abusiveness of a content cannot be determined if there is no context, (iii) *fact check:* the decision on abusiveness requires verification of the facts presented in the content, and (iv) *borderline cases:* content cannot be easily annotated.

OLID contains a lot of insulting instances that have not been labeled as such. This is mainly due to the fact that insults are perceived subjectively. The problem with HatEval arises from the specific targets, that is, even very abusive content is not labeled as such if it is not directed against women and/or immigrants. The loose definition of "OAG" and "CAG" in TRAC-2020 leads annotators to rely on their intuition. As a result, we believe that some comments are wrongly classified. In the HateBaseTwitter sample, we found hardly any controversial annotations but it became obvious that the data were collected by using abusive keywords. The same is true for the German HASOC part, where hardly any controversial annotation was found. In contrast, we found 25% of the English HASOC part to be controversial. In opposition to GermEval 2018 and GermEval 2019, where hardly any controversial annotations were found, GermEval 2021 contains particularly many controversial annotations. Mostly, this is attributed to a lack of context. Posts do not sound abusive per se, but in a given context they obviously are. Table 5 gives an overview of which dataset falls into which classes.

To achieve higher-quality annotations, we recommend the following actions for the respective classes. For *missing context*, we recommend providing context information, for example, the

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

Table 5. Classes of controversial annotations in each dataset

Dataset/ Class	Contradictory	Missing Context	Fact Check	Borderline Cases
OLID		X		X
HatEval	X		X	
TRAC-2020		X		
HateBaseTwitter	X			X
HASOC (English)	X			
GermEval2018		X		
GermEval2019		X	X	
GermEval2021	X	X		X
HASOC (German)	X	X		

preceding and following post. For *borderline cases*, we recommend providing a more distinguishable annotation scheme and clear definitions for each target class. The other two classes would require additional resources in the annotation process. *Contradictory* cases could be prevented by having repeated discussions between the annotators. *Fact* check would require additional time to research the truthfulness of claimed facts.

5. Experimental setup

To keep this paper within a reasonable scope, we will present only a limited computational part. The full set of experiments will be published elsewhere (Seemann *et al*. 2023). In this paper, we focus on the three datasets of the GermEval series, since they have almost identical categories and are based on similar annotation guidelines. We want to answer the question whether generalization effects can be achieved where they are most likely to be expected.

5.1 Machine learning models

Three machine learning classifiers are used in the experimental setup, that is, Logistic Regression (LG) from the linear model series, Complement Naive Bayes (CNB) from the Naive Bayes model series, and Linear Support Vector Classifier (LSVC) from the SVM model series of the scikit-learn library.[c] Features are obtained from the default CountVectorizer with case-insensitive word representation. During preprocessing, we replaced "&" by "und" (en: *and*) and ">" by "folglich" (en: *consequently*; users use this character to indicate a conclusion), masked user mentions by "user" and URLs by "http" (if not already done by the organizers), replaced emojis with their literal equivalents using the emoji library,[d] and segmented hashtags into their word components using the current state-of-the-art hashformer library[e] (a German GPT-2 model[f] was chosen as segmenter model, but not a reranker model) in all three datasets. We keep the original training and test data split as provided by organizers of the shared tasks.

[c] https://scikit-learn.org/stable/
[d] https://pypi.org/project/emojis/
[e] https://github.com/ruanchaves/hashformers
[f] https://huggingface.co/dbmdz/german-gpt2

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

Table 6. Results of the intra-dataset experiments that serve as our baseline (top) and the results of the generalization experiments (bottom). All values shown are macro-averaged F1 scores

Type	Training Data	Test Data	LG	CNB	LSVC
Baseline	GermEval2018	GermEval2018	60.88	66.72	63.14
	GermEval2019	GermEval2019	61.75	61.28	61.90
	GermEval2021	GermEval2021	56.72	59.81	57.50
Generalization	GermEval2018	GermEval2019	62.32	61.20	61.98
		GermEval2021	55.91	51.86	56.57
	GermEval2018 + 2019	GermEval2018	64.41	68.28	65.58
		GermEval2019	65.80	65.01	65.24
		GermEval2021	54.46	53.58	54.71
	GermEval2018 + 2019 + 2021	GermEval2018	65.20	68.19	65.67
		GermEval2019	66.02	63.50	65.51
		GermEval2021	54.72	57.58	56.12
	GermEval2021	GermEval2018	53.94	57.16	55.00
		GermEval2019	56.07	55.37	57.81

5.2 Intra-dataset results

To put our generalization results into perspective, we first perform intra-datasets experiments as to obtain a baseline. These experiments are not designed to achieve state-of-the-art results. Table 6 shows the performance results for these models.

5.3 Cross-dataset results

For our generalization experiments, we start with GermEval2018 as training data and evaluate on the other test sets. Then, we incrementally add the other GermEval training sets and evaluate on each test set. The generalization performance results of our experiments are also shown in Table 6.

The GermEval2018 dataset shows a decent generalization ability on the GermEval2019 test set with a plus of 0.19% in F1 score on average across all models compared to the GermEval2019 baseline results. For GermEval2021, there is an average loss of –3.23% compared to the baseline models.

Since the GermEval2018 and GermEval2019 datasets are the most similar, we expected to see an increase in performance when combining the data. As the results show, this expectation was confirmed with an increase of 2.51% on the GermEval2018 test set and 3.7% on GermEval2019 on average across all models. However, there is an average loss of –3.76% for the GermEval2021 test set compared to the GermEval2021 baseline.

Combining all GermEval training sets does not further improve generalization ability, but produces similar results as combining GermEval2018 and GermEval2019. Adding the GermEval2021 training data helps with the test set, but still results in a loss of –1.87%.

In summary, the generalization from GermEval2018 to GermEval2019 worked significantly better than to GermEval2021. We assume that three factors are responsible for this. First, the source of data collection differs: GermEval2018 + 2019 was sourced from Twitter and GermEval2021 was from Facebook. Second, an important difference in the annotation criteria was

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

that context was explicitly considered in the labeling of GermEval2021, but not in GermEval2018 and GermEval2019. Third, the abusive categories in GermEval2021 contain more types of abusive subcategories than the other datasets. Since all experiments resulted in a decrease in F1 score on GermEval2021, we tested the generalization properties of the GermEval2021 training data on the other test sets (see Table 6). The results show an average loss of –8.22% on GermEval2018 and –5.23% on GermEval2019, hence confirming that the differences between GermEval2018 + 2019 and GermEval2021 play a role in generalization. The GermEval2018 and Germeval2019 datasets can be successfully used in combination for training of more effective models when tested with either test set. The differences in political opinions (left-wing vs. right-wing) do not seem to affect the models.

6. Conclusion

Abusive content exists across different domains and platforms, so tackling this issue requires a standard data basis in order to develop effective and general solutions. Ideally, researchers should work on drawing universally applicable conclusions and on developing machine learning systems that generalize well across different domains. The two most essential components for such solutions are large amounts of labeled data and a common annotation scheme based on consistent concepts of abusive content. However, as shown in this article, in reality there are overlapping types of abusive content with fuzzy boundaries, such as hate speech, aggression, cyberbullying, or offensive language. Moreover, academia, institutions, and social media providers seem to have different definitions for these terms. For this reason, there are no reusable models, training sets, and benchmarks, which are urgently needed to pool the knowledge of the scientific community and address this serious problem. Our computational experiments have confirmed that generalization abilities are significantly impaired even when there are only small differences between datasets, as it is the case for GermEval2021 and GermEval2018 + 2019.

When analyzing the datasets presented in this article, we repeatedly encountered annotations that did not seem to conform to the annotation guidelines associated with the dataset. We manually examined a part of each dataset and found controversial annotations, which in turn can be categorized into four classes. We presented some comparative examples to illustrate where conflicting assessment criteria may have been applied. However, we suspect that in some of these cases, contextual information may have been critical to the annotation decision. For example, the content of a Facebook comment could be harmless on its own, but take on an offensive meaning when viewed in light of the post to which it refers. This type of content goes beyond profanity detection, which is necessary and included in most shared tasks. Therefore, context assessment is very important for evaluation, but none of the datasets provided context information for modeling. This is a serious problem because necessary information for modeling is missing and conflicting examples exist.

The main conclusion of our work is the call for a consistent definition of abusive language, at least in research, including definitions for related concepts such as hate speech, aggression, and cyberbullying. Annotation guideline writers should adhere to these definitions to ensure consistently annotated datasets that could be used as benchmarks for analyses in the future. Although a consistent definition of abusive language is not tackled in this paper, the research of Fortuna *et al.* (2021) implies a possible direction for the practical aspect. For example, they point out that the coarse-grained categories like "toxicity," "offensive," and "abusive" can be applied to cross-dataset generalization models if these category labels are standardized when merging datasets. Following this idea, one solution to the lack of a definition standard might be to simply combine many datasets with specific subtypes of abusive labeling into large datasets. Since deep learning models are nowadays becoming more complex and adaptive, it might be feasible to train and share models capable of capturing all different aspects of the abusive language contained in the combined dataset. Such a model could then still be evaluated on test sets labeled with specific

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

types of abusive language. A second approach could be to tackle the problem from the opposite side and redefine the term abusive language. For example, from the datasets examined, it can be inferred that most annotation guidelines related to the labels of "toxicity," "offensive," and "abusive" contain an instance of "insult" that implies the concept of a negative evaluation of a person or a group. In the same way, the minimum common similarities among a group of subtypes of abusive language could be established as standard definition and used to create new datasets.

Competing interests declaration

The authors declare none.

References

Agrawal S. and Awekar A. (2018). *Deep learning for detecting cyberbullying across multiple social media platforms. IN*, Pasi G., Piwowarski B., Azzopardi L. and Hanbury A. (eds.), *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26–29, 2018, Proceedings*. Springer, Proceedings, vol. **10772**, pp. 141–153.

American Bar Association (2017). Hate speech—ABA legal fact check. Available at: https://abalegalfactcheck.com/articles/hate-speech.html.

Arango A., Pérez J. and Poblete B. (2019). *Hate speech detection is not as easy as you may think: A closer look at model validation*. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19*, New York, NY, USA: Association for Computing Machinery, pp. 45–54.

Badjatiya P., Gupta S., Gupta M. and Varma V. (2017). *Deep learning for hate speech detection in Tweets*. In *Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion*, Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, pp. 759–760.

Basile V., Bosco C., Fersini E., Nozza D., Patti V., Rangel Pardo F. M., Rosso P. and 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*, Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 54–63.

Bhattacharya S., Singh S., Kumar R., Bansal A., Bhagat A., Dawer Y., Lahiri B. and Ojha A. K. (2020). *Developing a*multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, Marseille, France: European Language Resources Association (ELRA), pp. 158–168.

Bornheim T., Grieger N. and Bialonski S. (2021). *FHAC at GermEval 2021: Identifying German toxic, engaging, and fact-claiming comments with ensemble learning*. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, Duesseldorf, Germany: Association for Computational Linguistics, pp. 105–111.

Davidson T., Warmsley D., Macy M. and Weber I. (2017). *Automated hate speech detection and the problem of offensive language*. In *Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM '17*, pp. 512–515.

European Commission (2020). The EU code of conduct on countering illegal hate speech online. Available at: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

Fortuna P. and Nunes S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys* **51**(4), 85.

Fortuna P., Soler-Company J. and Wanner L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management* **58**(3), 102524.

Indurthi V., Syed B., Shrivastava M., Chakravartula N., Gupta M. and Varma V. (2019). *FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter*. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 70–74.

Kemp S. (2021). *Digital 2021: Global Overview Report*. Available at: https://datareportal.com/reports/digital-2021-global-overview-report

Kovács G., Alonso P. and Saini R. (2021). Challenges of hate speech detection in social media. *SN Computer Science* **2**(2), 1–15.

Kumar R., Ojha A. K., Malmasi S. and Zampieri M. (2020). *Evaluating aggression identification in social media*. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, Marseille, France: European Language Resources Association (ELRA), pp. 1–5.

Liu P., Li W. and 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*, pp. 87–91.

https://doi.org/10.1017/S1351324923000098 Published online by Cambridge University Press

MacAvaney S., **Yao H.-R.**, **Yang E.**, **Russell K.**, **Goharian N. and Frieder O.** (2019). Hate speech detection: challenges and solutions. *PloS One* **14**(8), e0221152.

Majumder P., **Mitra M.**, **Gangopadhyay S. and Mehta P.** (eds.) (2019). *FIRE '19: Proceedings of the 11th Forum for Information Retrieval Evaluation*, New York, NY, USA: Association for Computing Machinery.

Mandl T., **Modha S.**, **Majumder P.**, **Patel D.**, **Dave M.**, **Mandlia C. and Patel A.** (2019). *Overview of the HASOC Track at FIRE 2019: Hate Speech and offensive content identification in Indo-European*languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation, FIRE 2019*, New York, NY, USA: Association for Computing Machinery, pp. 14–17.

Montani J. P. and Schüller P. (2018). *TUWienKBS at GermEval 2018: German abusive Tweet detection*. In *Proceedings of the GermEval 2018 Workshop, 14th Conference on Natural Language Processing (KONVENS 2018)*, pp. 45–50.

Paraschiv A. and Cercel D.-C. (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)*, pp. 398–404.

Poletto F., **Basile V.**, **Sanguinetti M.**, **Bosco C. and Patti V.** (2021). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation* **55**(2), 477–523.

Risch J. and Krestel R. (2020). *Bagging BERT models for robust aggression identification*. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 55–61.

Risch J., **Schmidt P. and Krestel R.** (2021). *Data integration for toxic comment classification: Making more than 40 datasets easily accessible in one unified format*. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, Online. Association for Computational Linguistics, pp. 157–163.

Ruppenhofer J., **Siegel M. and Wiegand M.** (2018). Guidelines for IGGSA shared task on the identification of offensive language.

Saha P., **Mathew B.**, **Goyal P. and Mukherjee A.** (2019). HateMonitors: language agnostic abuse detection in social media, *FIRE'19*.

Schultz A. and Parikh J. (2020). Keeping our services stable and reliable during the COVID-19 outbreak. Available at: https://about.fb.com/news/2020/03/keeping-our-apps-stable-during-covid-19/

Seemann, **N.**, **Lee**, **Y.**, **Höllig**, **J. and Geierhos**, **M.** (2023). Generalizability of Abusive Language Detection Models on Homogeneous German Datasets. *In Datenbank-Spektrum*. doi: 10.1007/s13222-023-00438-1.

Struß J. M., **Siegel M.**, **Ruppenhofer J.**, **Wiegand M. and 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, Nürnberg/Erlangen, pp. 354–365.

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

Vidgen B., **Harris A.**, **Nguyen D.**, **Tromble R.**, **Hale S. and Margetts H.** (2019). *Challenges and frontiers in abusive content detection*. In *Proceedings of the Third Workshop on Abusive Language Online*, Florence, Italy: Association for Computational Linguistics, pp. 80–93.

Wang B., **Ding Y.**, **Liu S. and Zhou X.** (2019). YNU_Wb at HASOC 2019: ordered neurons LSTM with attention for identifying hate speech and offensive language, *FIRE'19*

Waseem Z. and Hovy D. (2016). *Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter*. In *Proceedings of the NAACL Student Research Workshop*, San Diego, California: Association for Computational Linguistics, pp. 88–93.

Wiegand M., **Siegel M. and 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)*, S.A., Vienna, Austria, pp. 1–10.

Wilms L., **Heinbach D. and Ziegele M.** (2021). Annotation guidelines for GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments, *Excerpt of an Unpublished Codebook of the DEDIS Research Group at Heinrich-Heine-University Düsseldorf (available on Request)*

Zampieri M., **Malmasi S.**, **Nakov P.**, **Rosenthal S.**, **Farra N. and Kumar R.** (2019a). *Predicting the Type and Target of Offensive Posts in Social Media*, (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**(Long and Short Papers), pp. 1415–1420.

Zampieri M., **Malmasi S.**, **Nakov P.**, **Rosenthal S.**, **Farra N. and Kumar R.** (2019b). *SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval)*. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 75–86.

Cite this article: Seemann N, Lee YS, Höllig J and Geierhos M (2023). The problem of varying annotations to identify abusive language in social media content. *Natural Language Engineering* **29**, 1561–1585. https://doi.org/10.1017/S1351324923000098