Datenbank Spektrum (2023) 23:15–25
https://doi.org/10.1007/s13222-023-00438-1

SCHWERPUNKTBEITRAG

Generalizability of Abusive Language Detection Models on Homogeneous German Datasets

Nina Seemann[1] · Yeong Su Lee[1] · Julian Höllig[1] · Michaela Geierhos[1]

Received: 18 October 2022 / Accepted: 11 February 2023 / Published online: 10 March 2023

Abstract
Abusive language detection has become an integral part of the research, as reflected in numerous publications and several shared tasks conducted in recent years. It has been shown that the obtained models perform well on the datasets on which they were trained, but have difficulty generalizing to other datasets. This work also focuses on model generalization, but – in contrast to previous work – we use homogeneous datasets for our experiments, assuming that they have a higher generalizability. We want to find out how similar datasets have to be for trained models to generalize and whether generalizability depends on the method used to obtain a model. To this end, we selected four German datasets from popular shared tasks, three of which are from consecutive GermEval shared tasks. Furthermore, we evaluate two deep learning methods and three traditional machine learning methods to derive generalizability trends based on the results. Our experiments show that generalization is only partially given, although the annotation schemes for these datasets are almost identical. Our findings additionally show that generalizability depends solely on the (combinations of) training sets and is consistent no matter what the underlying method is.

Keywords Model Generalizability · Abusive Language Detection · Data Combination · Qualitative Analysis

1 Introduction

There is no doubt that social media usage is high and people are posting on various websites. While most of the user-generated content is harmless, there is also hateful, abusive, or offensive content. Abusive language detection aims at automatically recognizing such content. Most work focuses on improving models to achieve new levels of performance in this task on specific datasets, but there are also many works that focus on the generalizability of trained models [7, 14, 26]. The results presented in previous work show that model generalization is not easy to achieve. Various characteristics of the data play a role, e.g., different categories between the datasets and dataset size [14], dataset similarity and percentage of abusive content [26], and degree of specificity of the categories [7]. Additionally, a model with high performance on the intra-dataset classification setup has a higher generalizability potential [7]. Most of this previous work deals with heterogeneous datasets in English. In a manual review, we analyzed samples of abusive language datasets from multiple shared tasks and found that they are based on different annotation guidelines and definitions of abusive language. Thus, the fact that machine learning models trained on such datasets do not easily generalize is not surprising.

In this paper, we investigate whether generalization is given when the evaluation is performed on homogeneous datasets, i.e., coming from the same shared task series and thus are based on the same or comparable annotation guidelines. We want to answer the question of how close datasets have to be to generalize. Hereby, we focus on three German datasets from the GermEval shared task series of the years 2018, 2019, and 2021. Although continuity is not encouraged by the organizers, there is an accidental one

Nina Seemann, Yeong Su Lee and Julian Höllig contributed equally to this work.

✉ Nina Seemann
nina.seemann@unibw.de

Yeong Su Lee
yeongsu.lee@unibw.de

Julian Höllig
julian.hoellig@unibw.de

Michaela Geierhos
michaela.geierhos@unibw.de

[1] Research Institute CODE, Bundeswehr University Munich,
Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany

between the shared tasks of 2018 and 2019. We point out the similarities and differences in Sect. 3. We first train and test several models on each of these datasets to obtain intra-dataset performance scores for our generalization experiments. Then, we evaluate whether augmentation by combining these datasets improves model performance and whether these models generalize well. Therefore, we train on different combinations of the datasets and compare the performances with the intra-dataset scores. In addition, we use the HASOC dataset from FIRE'19 [21], which is annotated with similar target categories but comes from a different shared task. We want to evaluate whether models generalize on this dataset. Furthermore, we want to answer the question of whether different methods have an influence on generalizability. Hence, we use two deep learning and three traditional machine learning methods in our experiments.

This paper is organized as follows. We review related work in Sect. 2, and in Sect. 3, we present the details of the datasets used. In Sect. 4, we show our experimental setup and results, while a qualitative analysis is given in Sect. 5. Finally, we conclude and present future work in Sect. 6.

2 Related Work

Our focus is on the generalizability of abusive language classification models across different datasets. Therefore, we only present related work on this topic.

[14] use nine different datasets [9, 13, 16, 17, 28, 29, 31] in their generalization experiments[1]. It is important to note that the authors binarize the labels of all datasets into 'positive' (abusive language) and 'negative' (non-abusive language). As a result, the distinction between the original categories might be lost, making a detailed analysis of their properties and a fine-grained classification of abusive language difficult. For each dataset, an SVM is trained with unigram count-based features and then tested on the other eight datasets. For the generalization experiments, the authors use FEDA ("Frustratingly Easy Domain Adaption") as a transfer learning approach. From the results, the authors conclude that it is important to use at least some data from the target dataset as training data to achieve a good performance. Furthermore, [14] hypothesize that the differences between the categories of the datasets as well as their sizes play an important role in cross-dataset performance.

[26] show that some generalization is possible with state-of-the-art models such as BERT but that it depends heavily on the training data. The authors use four datasets [4, 8, 29, 33] for the evaluation. Like [14], the authors group the categories of the datasets into 'positive' (abusive) and 'negative' (benign), although they do not capture the same type of abusive language. According to the authors, a model generalizes better when applied to similar data. In other experiments, [26] have shown that datasets with a higher percentage of positive samples generalize better than datasets with fewer positive samples when tested against a dissimilar dataset (at least within the same platform), suggesting that a more balanced dataset is better for generalization.

[22] use four datasets [1, 11, 29, 33] in their study that cover different kinds of phenomena; e.g., [29] cover only racism and sexism, while [33] include threat, insult, and profanity. Their experiments with a linear SVM with bag-of-words and an LSTM confirm that a model trained on datasets with a broader coverage of phenomena can detect other types of abusive language not seen during training.

[20] experiment with datasets for English [18], Slovene [18], and Dutch [19]. For each language, the authors use traditional features (e.g., words, character n-grams, and their combinations) as well as stylometric (e.g., part-of-speech tags, stop words) and emotion-based (i.e., emotion words and sentiment) features. Their in-domain experiments show that these features have a positive influence in different experiments for all three languages. For cross-domain evaluation, the authors evaluate the Dutch model on the Dutch part of [21] and the English model on Ask.fm [27]. The evaluation shows a significant decrease in the F1 score for both languages. However, the evaluation also shows that their stylometric and emotion-based features still perform better than the commonly used features.

[32] evaluate three transformer-based language models (BERT, RoBERTa, and ALBERTA) fine-tuned on five datasets [1, 4, 8, 10, 29] and obtain state-of-the-art results. Additionally, the authors are able to improve performance by augmenting these datasets with large amounts of synthetic data generated by a GPT-2 model. The authors show that data augmentation produces improvements in the generalization of hate detection on unseen examples across models and datasets. Additionally, they note that large amounts of authentic task-related data are not given for fine-tuning in the case of hate speech. Their main finding is that large language models can be used for synthetic data enrichment and that they can even yield better results than related human-labeled datasets.

[7] experiment with nine datasets [1, 4, 6, 8, 10, 13, 17, 29, 33] and use BERT, ALBERT, fastText, and SVM as classifiers. Prior to training, the authors identify and merge pairs of similar categories across the datasets (e.g. class *sexism* of dataset X and class *misogyny* of dataset Y become class *sexism-misogyny*). The intra-dataset model evaluation is a binary classification, where each of the four classifiers is trained on each dataset. For the cross-dataset model evaluation, the authors train one model for each dataset and test it against the remaining models. The authors observe that

[1] [31] published three datasets.

Table 1 Overview of datasets, including their size, percentage of abusive content, tasks to be performed, and categories of abusive language

Dataset	#Train	#Test	%Abusive	Task	Categories
GermEval2018	5,009	3,532	34	Offense	Abuse, Insult, Profanity
GermEval2019	3,995	3,031	32	Offense	Abuse, Insult, Profanity
GermEval2021	3,245	944	35	Toxic	Profanity, Insult, Sarcasm, Discrimination, Accusation
HASOC	3,819	850	24	Hate & Offense	Hate, Offense, Profanity

generalization varies across models, and some of the categories (e.g., 'toxic', 'abusive', or 'offensive') perform better as training categories than others (e.g., 'hate speech'). Additional experiments were conducted with a Random Forest classifier to assess the relevance of different models and datasets and to determine which specific features affect the generalizability. The authors note that a model must already perform well in an intra-dataset scenario to generalize well, i.e., the intra-dataset model performance is the most important predictor of generalization. Besides the intra- and cross-dataset experiments, [7] also conduct cross-class experiments (e.g. classifier trained on *abuse* applied to a *sexism* test set).

3 Dataset Selection

We use the datasets from GermEval2018 [30], GermEval2019 [25], GermEval2021 [24], and HASOC [21]. All datasets have binary categories representing an abusive and a non-abusive class. The GermEval2018 and GermEval2019 datasets are based on the same annotation guidelines and include tweets. The GermEval2021 datatset is based on very similar annotation guidelines but includes Facebook comments. The HASOC dataset is from a different organization, has overlap in the annotation guidelines compared to the GermEval datasets, and contains both Twitter and Facebook content.

In Table 1, we give an overview of the dataset sizes, how much abusive content each dataset contains, which task was to be performed, and which different categories were included in each task.

3.1 GermEval2018

Task 1 is a coarse-grained binary classification into two classes, 'OFFENSE' and 'OTHER'. Task 2 is a fine-grained classification into four classes including the class 'OTHER'. The class 'OFFENSE' identified by Task 1 is further subdivided into three fine-grained classes: 'PROFANITY', 'INSULT', and 'ABUSE'. Tweets from about 100 different users and their timelines have been collected. Furthermore, the abusive content consists mainly of the far-right spectrum and the (at that time) dominant topic of migration. Finally, the data has been split into training and test sets by ensuring that a user's complete set of tweets is either exclusively in the training or test part.

3.2 GermEval2019

In this edition, Task 1 and 2 are identical to GermEval2018 and a new task regarding the explicitness of OFFENSE has been deployed. Some of the training data for GermEval2019 comes from the GermEval2018 data collection but is not part of the GermEval2018 dataset. Additional data for GermEval2019 has been collected by heuristically identifying users who regularly post abusive content. Furthermore, abusive content containing antisemitism and content from the far-left spectrum has been added to increase topic variance. The split into training and test set is identical to GermEval2018.

3.3 GermEval2021

Unlike previous editions, GermEval2021 consists only of binary classification tasks: (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. It should be noted that the annotation guidelines are different from the previous editions, but the organizers claim that they are comparable. As shown in Table 1, 'TOXIC' includes shouting, profanity, insults, sarcasm, discrimination, discrediting, and accusations of lying or deception. Not all of these categories of abusive language were included in the previous editions. Also in contrast to the previous editions, the collected data comes from the Facebook page of a political talk show on a German TV station. The training and test data have been taken from comments on different shows to avoid a topic bias. Due to the nature of the Facebook page, there are comments in response to a post by the talk show staff or even comments on comments by other users. In contrast to the other three datasets used in this paper, the annotators had to consider the entire context of a post, i.e., the previous and following posts, before assigning a category while annotating the data. Hence, the context of a comment played an important role in the annotation process. However, the dataset does not contain the context, only the post itself.

3.4 German HASOC

Subtask A is a coarse-grained binary classification into Hate & Offensive (‘HOF’) and regular content (‘NOT’). If a post has been classified as ‘HOF’, it is processed further in subtask B, where a fine-grained classification into hate speech, offense, or profanity has to be made. Subtask C deals with targeting or non-targeting individuals, groups, or others when a post is classified as ‘HOF’. Importantly, the dataset was taken from Twitter and partially from Facebook. Heuristics were used to search for typical hate speech on social media platforms to identify topics where many hate posts are expected.

4 Experimental Setup

In the experimental part of our work, we test several popular machine learning models with the datasets described in Sect. 3. We want to evaluate whether generalization effects can be achieved where they are most likely to be expected, or whether datasets created at different times and under possibly slightly different circumstances ultimately prevent generalization effects.

In the following subsections, we will explain the methods we used and present the experimental results.

4.1 Methods

Considering that one of our goals is to investigate whether the generalizability of abusive language classification models is dependent on the underlying method, we use two deep learning methods (BERT and CNN) as well as three traditional machine learning methods (Logistic Regression, Naive Bayes, and Support Vector Machines) to obtain models.

For all datasets, we performed the following preprocessing steps: Replacing ‘&’ by ‘und’ (en: *and*) and ‘>’ by ‘folglich’ (en: *consequently*), since users use this character to indicate a conclusion; masking user mentions by ‘user’ and URLs by ‘http’ (if not already done by the organizers); replacing emojis with their literal equivalents using the emoji library[2]; segmenting hashtags into their word components using the current state-of-the-art hashformer library[3] (a German GPT-2 model[4] was chosen as segmenter model, but not a reranker model). All model-specific preprocessing steps are described in the following subsections. All model-specific settings and parameters for the CNN (e.g., filter size, number of epochs) and the traditional machine learning models (word representation by CountVectorizer), were chosen due to preliminary experiments.

4.1.1 BERT

We use a pre-trained BERT [5] base cased model from Huggingface[5]. The following hyperparameter recommendations from [23] have been applied: batch size: 16, learning rate: 2×10^{-5}, epochs: 4, maximum sequence length: 128, and loss function: cross entropy. Apart from the required preprocessing steps for BERT (adding [CLS] and [SEP] tokens, and tokenization with the BERT tokenizer), no other steps have been implemented.

4.1.2 Convolutional Neural Network

For our implementation of the Convolutional Neural Network (CNN), we use the Tensorflow/Keras API[6]. As additional preprocessing, we lowercase the data and remove all punctuation and implement a character-level model with 112 features. After examining the character length in all datasets, we set the maximum sequence length to 500 per tweet or post. We train the model with the following settings: filter size = $[5, 6, 7]$, number of filters: 100, activation: ReLU, and output: sigmoid. To prevent overfitting of the model, we add a dropout layer of 0.5 within the convolution layer, another one of 0.4 after concatenation, and a batch normalization on the dense layer, followed by a final dropout layer of 0.4. The Adam optimizer [15] with default parameters is applied and the CNN is trained with a batch size of 128 for 80 epochs.

4.1.3 Traditional Machine Learning Models

We use 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[7] and obtain features from the default CountVectorizer with case-insensitive word representation.

4.2 Experimental Results

The organizers of the shared tasks provide training and test sets that we use in our experiments. For the deep learning models, we additionally split a 20 percent sample from the training set for development.

[2] https://pypi.org/project/emojis/.

[3] https://github.com/ruanchaves/hashformers.

[4] https://huggingface.co/dbmdz/german-gpt2.

[5] https://huggingface.co/transformers/model_doc/bert.html#tfbertforsequenceclassification.

[6] https://www.tensorflow.org/api_docs/python/tf/keras.

[7] https://scikit-learn.org/stable/.

Table 2 Results of the intra-dataset (a) and the results of the cross-dataset experiments (b)

Training Data	Test Data	BERT	CNN	LG	CNB	LSVC
(a) Results of the intra-dataset experiments. All values are macro-averaged F1 scores (for BERT, the standard deviation for three independent experiments is given).						
GermEval2018	GermEval2018	67.56 (6.74)	65.64	60.88	66.72	63.14
GermEval2019	GermEval2019	68.59 (0.55)	64.56	61.75	61.28	61.90
GermEval2021	GermEval2021	58.58 (6.6)	53.12	56.72	59.81	57.50
HASOC	HASOC	46.69 (1.0)	47.07	50.59	49.16	52.60
(b) Results of the cross-dataset experiments. All values are macro-averaged F1 scores (for BERT, the standard deviation for three independent experiments is given).						
GermEval2018	GermEval2019	67.12 (6.11)	60.15	62.32	61.20	61.98
	GermEval2021	56.93 (1.98)	55.19	55.91	51.86	56.57
	HASOC	57.60 (4.82)	54.59	58.30	50.22	56.11
GermEval2018 + 2019	GermEval2018	72.62 (0.41)	68.47	64.41	68.28	65.58
	GermEval2019	70.77 (0.21)	66.17	65.80	65.01	65.24
	GermEval2021	46.71 (14.4)	53.35	54.46	53.58	54.71
	HASOC	58.41 (1.83)	53.68	61.27	50.47	59.23
GermEval2018 + 2019 + 2021	GermEval2018	72.74 (1.13)	63.60	65.20	68.19	65.67
	GermEval2019	71.50 (1.33)	61.54	66.02	63.50	65.51
	GermEval2021	57.21 (2.54)	41.27	54.72	57.58	56.12
	HASOC	57.64 (1.07)	50.91	59.28	54.77	57.90
HASOC augmented	HASOC	45.46 (2.05)	44.19	44.05	44.38	44.01

All model results are measured as a macro-averaged F1 score. For the BERT models, the average and standard deviation of three independent F1 scores are taken because the values can fluctuate quite a bit due to random weight initialization of the classification layer and random batch order.

4.2.1 Intra-Dataset Performance

Before we begin the cross-dataset experiments, we train models for each dataset to obtain intra-dataset scores, i.e., what results can be expected from our models. They are not designed to achieve state-of-the-art results.

Table 2a shows the performance results for these models. However, it can be seen that traditional machine learning (ML) and deep learning (e.g., BERT and CNN) models do not show huge differences in terms of F1 scores between all datasets. We assume that this is due to the limited amount of training data. Surprisingly, the results for HASOC are also much lower than the results for all other datasets. The performances of the individual classes in the test set show that the 'HOF' class received an F1 score of 0 for the deep learning models and nearly 0 for the machine learning models. We suspect that this is due to the strong class imbalance (11.74 percent abusive samples) in the training set. For this reason, we augment the abusive class by uniformly sampling abusive samples from the GermEval datasets to balance the classes for our experiments on the cross-dataset performance. Since the GermEval2018 model as well as the combination models obtained higher F1 scores than the HASOC baseline, this seems like a reasonable choice. The assumption is that solving the class imbalance problem increases the performance on its test set.

4.2.2 Cross-Dataset Performance

To evaluate the generalizability of the datasets, we first train our models on a single dataset, and then we incrementally add the other GermEval datasets.

For the training on a single dataset, we use GermEval2018 because it provides the most training samples. We expect a good cross-dataset score on the GermEval2019 test set since the annotation guidelines and categories are identical for GermEval2018 and GermEval2019. Moreover, the data comes from the same social media platform (Twitter) and the content selection criteria (political discussions) are very similar. We also expect a reasonable cross-dataset score on the GermEval2021 test set, as the data was collected as part of the same shared task series and the organizers claim that the annotations are comparable. The HASOC dataset has similar categories compared to the GermEval datasets but was not part of the GermEval shared task series, so we expect the lowest cross-dataset score for this test set.

For training on the combined datasets, we assume that more data should lead to higher performance if the datasets

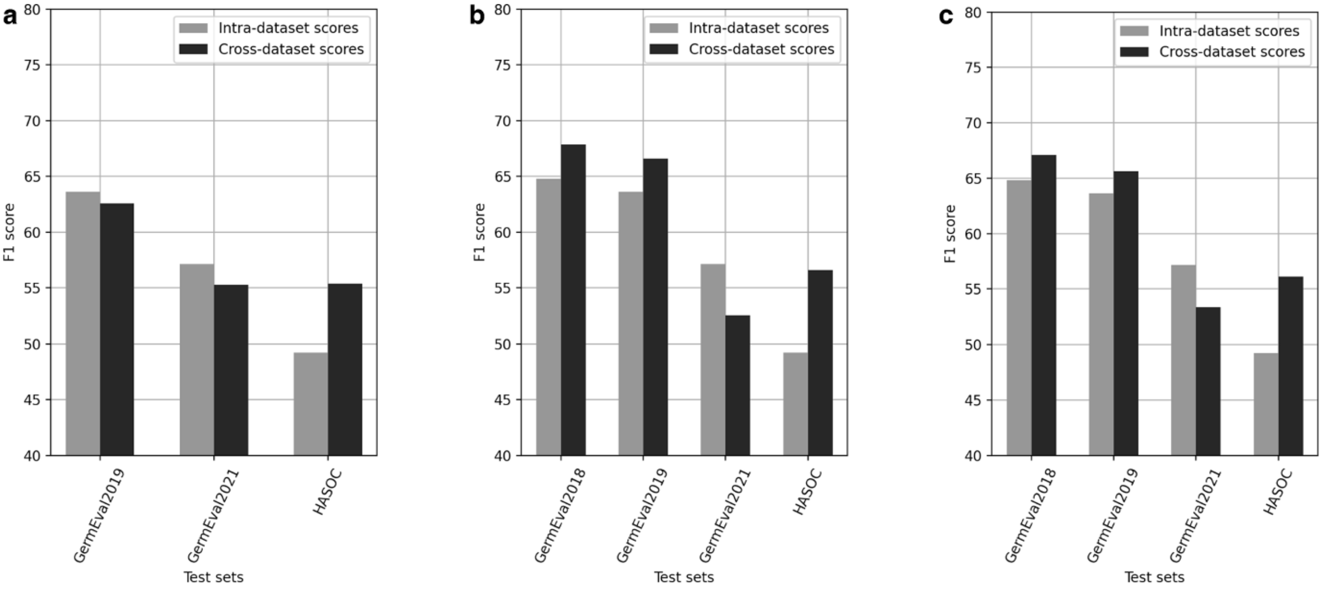

Fig. 1 Generalizability (cross-dataset scores) of the models trained on the three different training set(s) (combinations) compared to the intra-dataset results. The subtitles **a–c** indicate the datasets on which the models were trained. All F1 scores are averaged across the five models described here. **a** GermEval2018, **b** GermEval2018 + 2019, **c** GermEval2018 + 2019 + 2021

have high generalizability. Again, we test our models trained on these combinations on each individual test set.

The performance results of our cross-dataset experiments are shown in Table 2b. The GermEval2018 model shows decent generalizability on the GermEval2019 test set with a loss of 1.07 percent points in F1 score on average across all models (see first row in Table 2b) compared to the GermEval2019 intra-dataset results (see the second row in Table 2a). The generalization effect on the GermEval2021 test set is smaller. Surprisingly, the performance on the HASOC test set is higher than the scores obtained by the HASOC model itself. Fig. 1a shows these differences where the F1 scores of all five models are averaged.

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 for GermEval2018 + 2019 in Fig. 1b show, this expectation was confirmed with an increase of 3.08 percent points on the GermEval2018 test set and 2.94 percent points on the GermEval2019 test set on average across all models. However, there is no increase for the GermEval2021 test set compared to the GermEval2021 model itself and the GermEval2018 model. Similarly, the generalization performance on the HASOC test set is not much better in comparison to the GermEval2018 model (compare Fig. 1a and b).

The combination of all GermEval training data (see GermEval2018 + 2019 + 2021 in Table 2b) does not further increase the generalizability, but slightly decreases it compared to the combination GermEval2018 + 2019 (compare Fig. 1b and c).

Under the assumption that solving the class imbalance problem of the HASOC dataset increases the performance on its test set, we augmented it with abusive samples from the other datasets (see Sect. 4.2.1). Although the metrics show a large increase in performance for the formerly underrepresented class, the overall performance was not better than for the models trained with the original imbalanced HASOC training set (see HASOC augmented in Table 2b).

5 Qualitative Analysis

In the following, we present a qualitative analysis of the cross-dataset results of the GermEval series. For almost all experiments presented in Sect. 4.2.2, the BERT models obtained the highest F1 scores. So we limit our analysis to the results obtained by these models.

5.1 Error Analysis

In Fig. 2, we present the confusion matrices for the predictions on each of the three GermEval test sets as obtained by the BERT models trained with different (combinations of) training data. On the GermEval2018 test set (see Fig. 2a), all three models predict the true negatives well but not the true positives. Additionally, all models predict a lot of false negatives which is only slightly remedied with augmented training data. Similar observations can be made for the GermEval2019 test set (see Fig. 2b). Interestingly, no model (combination) is able to predict more true positives than false negatives as opposed to the results of the model com-

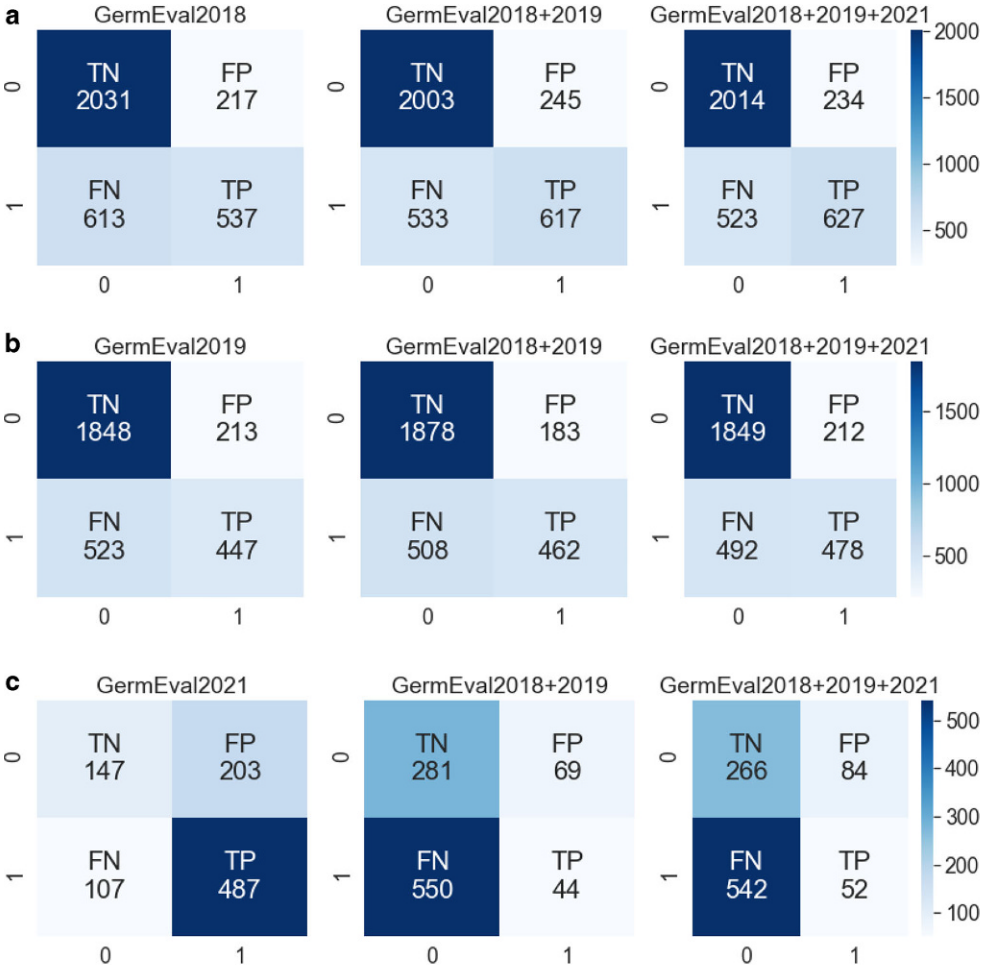

Fig. 2 Confusion matrices for the predictions of the models obtained by (combinations of) training set(s) on the three test sets. The abusive class is represented by 1', while the non-abusive class is represented by 0'. The subtitles **a–c** indicate the test set to which the models were applied. **a** GermEval2018 test set, **b** GermEval2019 test set, **c** GermEval2021 test set

binations on the GermEval2018 test set. A completely different picture is painted by the results of the GermEval2021 test set (see Fig. 2c). While the GermEval2021 model predicts many true positives, the combination models can predict hardly any. Contrary, the combination models tend to predict many false negatives but are at least able to correctly predict more true negatives than the GermEval2021 model.

Additionally, we computed the Matthews correlation coefficient (MCC) [3]. MCC is a statistical measure that produces a high score only if the prediction obtained good results in all of the four confusion matrix categories, proportionally both to the size of positive samples and the size of negative samples in the dataset. MCC ranges in the interval $[-1,+1]$, with -1 reached in case of perfect misclassification and +1 reached in case of perfect classification. An MCC = 0 is the expected value for the coin-tossing classifier. The results in Table 3 confirm the findings from the confusion matrices. The intra-dataset models of GermEval2018 and GermEval2019 as well as both combination models obtain a reasonable MCC score between 0.40 and 0.48 on all three test sets. But the GermEval2021 intra-dataset model reaches only an MCC of 0.26 and both combination models obtain negative MCC scores.

5.2 Manual Inspection

We manually analyze the test samples of the GermEval series that are mispredicted by the combination models but

Table 3 MCC for each model (combination) on the three test sets

Model/Test Set	GermEval2018	GermEval2019	GermEval2021
GermEval2018	0.42	–	–
GermEval2019	–	0.40	–
GermEval2021	–	–	0.26
GermEval2018+2019	0.47	0.44	–0.18
GermEval2018+2019+2021	0.48	0.43	–0.21

Table 4 Number of remaining false positive (FP) and false negative (FN) test samples mispredicted by the combination models but correctly predicted by the intra-dataset models

Model/Test Set	GermEval2018		GermEval2019		GermEval2021	
	FP	FN	FP	FN	FP	FN
GermEval18 + 19	144	92	84	98	44	460
GermEval18 + 19 + 21	130	80	105	94	58	464

correctly predicted by the corresponding intra-dataset models. We restrict our analysis to the models that produced the most interesting results, which are GermEval2018 + 2019 and GermEval2018 + 2019 + 2021. There are overlapping wrong predictions between the intra-dataset models and the combination models but we are more interested in the wrong predictions that were only made by the combination models. Table 4 shows these mispredictions for each GermEval test set. Also, both the GermEval2018 + 2019 and the GermEval18 + 19 + 21 models have overlapping wrong predictions. We made sure to inspect different samples per model. Overall, we inspected 600 samples, i.e., we analyzed 50 false positives and 50 false negatives samples per test set for each model.

5.2.1 GermEval2018/GermEval2019 Test Set

The manual inspection of the GermEval2018 and GermEval2019 test sets resulted in the same findings, which is why they are presented in the same subsection.

We have identified three possible reasons why both combination models generate a high number of false positives. First, there are samples that contain abusive terms but the samples are not abusive *per se*. Examples Ch1.e1 and Ch1.e2 contain terms that are often used in the 'OFFENSE' class but are not meant in a derogatory way here. Second, the test set contains samples that have – in our opinion – controversial annotations. Adhering to the annotation guidelines, we would have annotated them as 'OFFENSE' but they are not. In examples Ch1.e3 and Ch1.e4, the samples are exhibiting an abusive character but are annotated as 'OTHER'. Third, there are samples that do not contain abusive terms and sound innocuous, which our BERT combination models classify as 'OFFENSE' (see examples Ch1.e5 and Ch1.e6). We could not identify a reason for this behavior.

1. Die gesellschaftlichen Bemühungen um die Integration von Migranten wurzeln in den im Grundgesetz normierten Grundwerten.
2. Fakt Über 80% der Menschen, die aus dem Südsudan fliehen, sind Frauen und Kinder.
3. @USER Aber mit Spinnern kann man umgehen, solange sie noch eine Minderheit sind. Was aber, wenn die Spinner an die Macht gelangen?
4. Zeigen wir diesen Sklaven, wo sie stehen! Unter unseren Füßen!
5. @USER @USER .sie haben mir aus der Seele gesprochen!!!
6. @USER @USER Ick kann Dir nicht folgen.

For the false negatives, we found that these samples do contain abusive terms but these terms were hardly or not at all present in the training data. In example Ch1.e1a, the term 'Bolschewistenhure' is used but was never seen during training. The term 'hassen' (see example Ch1.e2a) occurred only six times in the training data. We assume that this is the reason that our models mispredict such samples.

1. Die Bolschewistenhure wird endlich da enden, wo viele geendet haben!
2. Ich weiß, ich habe das schon mal geschrieben, aber Ich hasse die Grünen. Mit tiefster Inbrunst.

5.2.2 GermEval2021 Test Set

As for the previous test sets, we reach the same conclusion regarding the false positives. However, our findings on the false negatives for this test set differ. While inspecting these samples and following the annotation guidelines, we find the labels provided by the annotators rather controversial. We present four samples from the test set in examples Ch1.e1b–Ch1.e4a that are annotated as 'TOXIC' by the shared task organizers. From our point of view, those samples are not abusive. Since the annotation guidelines state explicitly that the context has to be taken into account while annotating, we assume that those samples are abusive when evaluated in context. But the context is not provided for training, hence, our model lacks the proper knowledge to correctly predict such samples. Furthermore, we assume this missing context is the reason why so many false negatives have been predicted on this test set.

1. Hackt nicht nimmer auf den Fussball rum. Bei allem Sportarten sind wieder Zuschauer erlaubt. Hygienekonzept vorausgesetzt.
2. Ich bin alt(66), mir hat Corona viel mehr verbleibende Zeit gestohlen. Es macht keinen Unterschied wem diese Zeit gestohlen wird, sie ist verlohren.

3. @USER ein Jahr im Leben eines jungen Menschen ist genauso viel wert wie ein Jahr im Leben eines älteren Menschen.
4. @USER das ist leider auch wahr

6 Conclusion and Future Work

We focus on the generalizability of models trained on homogeneous German datasets and investigate whether generalizability is dependent on the method used to obtain a model. This work is mainly analytical and does not attempt to achieve state-of-the-art model performances.

For generalizability, we conclude that generalization from the GermEval2018 model on the GermEval2019 test set works significantly better than on the GermEval2021 test set.

The combination model GermEval2018+2019 can be successfully used for training more effective models, as the evaluation on both the GermEval2018 and the GermEval2019 test sets has shown. The differences in political opinions (left-wing vs. right-wing) do not seem to interfere with the model performances. Again, the evaluation of the GermEval2021 test set shows not much improvement.

The evaluation of the combination model GermEval2018 +2019+2021 shows that the addition of the GermEval2021 data set does not further increase but decreases the generalizability.

We assume that three factors are responsible for the lower results on the GermEval2021 test set. First, the source of data collection differs: For GermEval2018+2019, it was Twitter and for GermEval2021, it was Facebook. Second, the abusive categories in GermEval2021 contain more types of abusive subcategories than the other datasets. Third, the context was explicitly considered in the annotation of GermEval2021, but not in GermEval2018 and GermEval2019.

Our manual sample inspection revealed three main reasons as the cause of the models' high false positive rate: (i) non-abusive samples contain terms often associated with abusive content in different contexts, (ii) samples that should be labeled as abusive according to the annotation guidelines are labeled as not abusive, and (iii) some samples containing no abusive terms are mispredicted as abusive by our models. For the false negatives on the GermEval2018 and GermEval2019 test sets, we found that these samples contain abusive terms that are rarely or not at all present in the training data. Contrarily, the false negatives on the GermEval2021 test set are mostly due to the samples not being abusive from our point of view. As mentioned before, the context of the samples had to be considered during the annotation process and we assume that the missing context provides the abusive character. We suppose that this is also the reason why the addition of GermEval2021 data to any combination model hinders generalizability. The seemingly innocuous samples apparently confuse the model if no context is provided.

The results for the HASOC test set are surprising. The model trained on HASOC itself performs worse than all other models, even though the other models were trained on data from the GermEval series. To rule out class imbalance as the cause of this effect, we created a balanced training set for HASOC. But even then, the evaluation showed no increase in performance.

Regarding the question of whether different methods have an influence on generalizability, we can conclude that although BERT shows the best performance across all datasets, the relative generalizability depends solely on the (combinations of) training sets and is consistent for all models obtained by different methods.

This work – as well as previous work – focused on binary classification tasks. In future work, we plan to investigate the fine-grained classification of different types of abusive language. Furthermore, we know that in other research contexts, data augmentation leads to better generalization [32]. Therefore, we plan to also explore approaches to data augmentation, for example, by randomly replacing words in the training dataset with words from GermaNet [12] synsets, or by using a transformer model such as BERT to randomly replace masked words. The use of back translation to create additional training data could also be investigated [2].

Funding Partial financial support was received from dtec.bw – Digitalization and Technology Research Center of the Bundeswehr within MuQuaNet. dtec.bw is funded by the European Union – NextGenerationEU.

Conflict of Interest
The authors have no conflicts of interest to declare that are relevant to the content of this article.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Basile V, Bosco C, Fersini E et al (2019) SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and

women in Twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp 54–63 https://doi.org/10.18653/v1/S19-2007

2. Beddiar DR, Jahan MS, Oussalah M (2021) Data expansion using back translation and paraphrasing for hate speech detection. Online Soc Netw Media 24:100–153. https://doi.org/10.1016/j.osnem.2021.100153
3. Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. Bmc Genomics. https://doi.org/10.1186/s12864-019-6413-7
4. Davidson T, Warmsley D, Macy M et al (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
5. Devlin J, Chang MW, Lee K et al (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Long and short papers. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol 1. ACL, Minneapolis, pp 4171–4186 https://doi.org/10.18653/v1/N19-1423
6. Fersini E, Rosso P, Anzovino M (2018) Overview of the task on automatic misogyny identification at IberEval 2018. In: Rosso P, Gonzalo J, Martínez R et al (eds) Proceedings of the third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), vol 2150. Ceur Workshop Proceedings, Sevilla, pp 214–228
7. Fortuna P, Soler-Company J, Wanner L (2021) How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? Inf Process Manag 58(3):102524. https://doi.org/10.1016/j.ipm.2021.102524
8. Founta A, Djouvas C, Chatzakou D et al (2018) Large scale crowdsourcing and characterization of Twitter abusive behavior. ICWSM 12(1):491–500
9. Gao L, Huang R (2017) Detecting Online hate speech using context aware models. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. INCOMA Ltd., Varna, Bulgaria, pp 260–266 https://doi.org/10.26615/978-954-452-049-6_036
10. de Gibert O, Perez N, García-Pablos A et al (2018) Hate speech dataset from a white supremacy forum. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). Association for Computational Linguistics, Brussels, Belgium, pp 11–20 https://doi.org/10.18653/v1/W18-5102
11. Golbeck J, Ashktorab Z, Banjo RO et al (2017) A large labeled corpus for Online harassment research. In: Proceedings of the 2017 ACM on Web Science Conference. Association for Computing Machinery, New York, NY, USA, WebSci '17, pp 229–233 https://doi.org/10.1145/3091478.3091509
12. Hamp B, Feldweg H (1997) GermaNet - a lexical-semantic net for German. In: Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications
13. Jigsaw/Conversation AI (2018) Toxic comment classification challenge. https://tinyurl.com/y7qmd8lm. Accessed 09.05.2022
14. Karan M, Šnajder J (2018) Cross-domain detection of abusive language Online. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). Association for Computational Linguistics, Brussels, Belgium, pp 132–137 https://doi.org/10.18653/v1/W18-5117
15. Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: Bengio Y, LeCun Y (eds) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings
16. Kolhatkar V, Wu H, Cavasso L et al (2018) The SFU opinion and comments corpus: a corpus for the analysis of Online news comments. Corpus Pragmat 4:155–190. https://doi.org/10.1007/s41701-019-00065-w
17. Kumar R, Reganti AN, Bhatia A et al (2018) Aggression-annotated corpus of Hindi-English code-mixed data. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan
18. Ljubešić N, Fišer D, Erjavec T (2019) The FRENK Datasets of socially unacceptable discourse in Slovene and English. In: TSD. Lecture notes in computer science, vol 11697. Springer, Cham, pp 103–114
19. Ljubešić N, Markov I, Fišer D et al (2020) The LiLaH emotion lexicon of Croatian, Dutch and Slovene. In: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media. Association for Computational Linguistics, Barcelona, Spain, pp 153–157
20. Markov I, Ljubešić N, Fišer D et al (2021) Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In: Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Association for Computational Linguistics, pp 149–159
21. Modha S, Mandl T, Majumder P et al (2019) Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages. In: Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation. CEUR-WS, FIRE '19, pp 167–190
22. Pamungkas EW, Patti V (2019) Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Association for Computational Linguistics, Florence, Italy, pp 363–370 https://doi.org/10.18653/v1/P19-2051
23. Prakash A (2019) Fine-tuning BERT model using PyTorch. https://medium.com/@prakashakshay90/f34148d58a37. Accessed 14.10.2022
24. Risch J, Stoll A, Wilms L et al (2021) Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In: Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments. Association for Computational Linguistics, pp 1–12
25. Struß JM, Siegel M, Ruppenhofer J et al (2019) Overview of germEval task 2, 2019 shared task on the identification of offensive language. In: German Society for Computational Linguistics (ed) Proceedings of the 15th Conference on Natural Language Processing (KONVENS), pp 354–365 https://doi.org/10.5167/uzh-178687
26. Swamy SD, Jamatia A, Gambäck B (2019) Studying Generalisability across abusive language detection datasets. In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). Association for Computational Linguistics, Hong Kong, China, pp 940–950 https://doi.org/10.18653/v1/K19-1088
27. Van Hee C, Lefever E, Verhoeven B et al (2015) Detection and fine-grained classification of cyberbullying events. In: Proceedings of the International Conference Recent Advances in Natural Language Processing. INCOMA Ltd. Shoumen, Bulgaria, Hissar, Bulgaria, pp 672–680
28. Waseem Z (2016) Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In: Proceedings of the First Workshop on NLP and Computational Social Science. Association for Computational Linguistics, Austin, Texas, pp 138–142 https://doi.org/10.18653/v1/W16-5618
29. Waseem Z, Hovy D (2016) Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In: Proceedings of the NAACL Student Research Workshop. Association for Com-

putational Linguistics, San Diego, California, pp 88–93 https://doi.org/10.18653/v1/N16-2013

30. Wiegand M, Siegel M, Ruppenhofer J (2018) Overview of the GermEval 2018 shared task on the identification of offensive language. In: Ruppenhofer J, Siegel M, Wiegand M (eds) Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), pp 1–10
31. Wulczyn E, Thain N, Dixon L (2017) Ex machina: Personal attacks seen at scale. In: Proceedings of the 26th International Conference on World Wide Web, pp 1391–1399 https://doi.org/10.1145/3038912.3052591
32. Wullach T, Adler A, Minkov E (2021) Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In: Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, Punta Cana, Dominican Republic, pp 4699–4705 https://doi.org/10.18653/v1/2021.findings-emnlp.402
33. Zampieri M, Malmasi S, Nakov P et al (2019) Predicting the type and target of offensive posts in social media. In: Long and short papers. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol 1. Association for Computational Linguistics, Minneapolis, pp 1415–1420 https://doi.org/10.18653/v1/N19-1144