Originally published in: Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 – 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg. - München [u.a.]: German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg, 2019. Pp. 352-363

Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language

Julia Maria Struß
Potsdam University of Applied Sciences
Kiepenheuerallee 5
14469 Potsdam
struss@fh-potsdam.de

Melanie Siegel
Darmstadt University of Applied Sciences
Max-Planck-Str. 2
64807 Dieburg
melanie.siegel@h-da.de

Josef Ruppenhofer
Leibniz Institute for German Language
R5, 6-13
68161 Mannheim
ruppenhofer@ids-mannheim.de

Michael Wiegand
Leibniz ScienceCampus
Heidelberg/Mannheim
wiegand@ids-mannheim.de

Manfred Klenner
University of Zurich
Andreasstrasse 15
8050 Zurich
klenner@cl.uzh.ch

Abstract

We present the second edition of the GermEval Shared Task on the Identification of Offensive Language. This shared task deals with the classification of German tweets from Twitter. Two subtasks were continued from the first edition, namely a coarse-grained binary classification task and a fine-grained multi-class classification task. As a novel subtask, we introduce the classification of offensive tweets as explicit or implicit.

The shared task had 13 participating groups submitting 28 runs for the coarse-grained task, another 28 runs for the fine-grained task, and 17 runs for the implicit-explicit task.

We evaluate the results of the systems submitted to the shared task. The shared task homepage can be found at `https://projects.fzai.h-da.de/iggsa/`

1 Introduction

The idea of social media was originally to enable an open exchange of information and opinions between people and thus to support communication. This idea of social participation is massively disturbed by current trends: Where an open exchange of views on political issues was possible, forums are increasingly inundated by offensive language. In many cases it is no longer possible to moderate forums without technical support.

The second GermEval Shared Task on the Identification of Offensive Language is intended to initiate and foster research on the identification of offensive content in German language microposts. Offensive comments are to be detected from a set of German tweets. We focus on Twitter, since tweets can be regarded as a prototypical type of micropost.

GermEval is a series of shared task evaluation campaigns that focus on natural language processing for the German language. Since 2014, there were shared tasks on named entity recognition, lexical substitution, sentiment analysis, hierarchical classification of blurbs, and identification of offensive language. These shared tasks have been run informally by self-organized groups of interested researchers and were endorsed by special interest groups within the German Society for Computational Linguistics (GSCL).

This paper will give a short overview on related work in §2. We will then describe the task in §3 and the data in §4. 13 teams participated in the shared task. We give an overview of their approaches and results in §5, and offer our conclusions in §6.

2 Related Work

For a recent overview of related work on the detection of abusive language, we refer the reader to Schmidt and Wiegand (2017) and Mishra et al. (2019). In what follows, we will briefly discuss related shared tasks as well as datasets for German.

- GermEval 2018 - To our knowledge this was the first shared task on the detection of offensive language that included German language data. (Wiegand et al., 2018b)

- SemEval 2019 - Task 5 (HatEval) concerned multilingual (English and Spanish) detection of hate speech against immigrants and women in Twitter. The two subtasks addressed binary classification (hateful or not) and the classification of the target harassed as individual or generic. (Basile et al., 2019)

- SemEval 2019 - Task 6 (OffensEval 2019) was a shared task on identification and classification of offensive language in social media. The dataset contains 14,000 English tweets. The subtasks were to identify offensive tweets, to categorize them, and to identify the targets of the offensive posts. (Zampieri et al., 2019)

- Kaggle's 2018 Toxic Comment Classification Challenge[1] was a shared task in which comments from the English Wikipedia are to be classified. There were 6 different categories of toxity to be identified (i.e. *toxic*, *severe toxic*, *obscene*, *insult*, *identity hate* and *threat*). The categories were not mutually exclusive.

- The TRAC shared task on aggression identification (Kumar et al., 2018) included both English and Hindi Facebook comments. Participants had to detect abusive comments and to distinguish between *overtly aggressive comments* and *covertly aggressive comments*.

- The shared task on Automatic Misogyny Identification (AMI) (Fersini et al., 2018) is jointly run by IberEval[2] and EVALITA[3]. It exclusively focused on the detection of misogynist tweets on Twitter. There were two subtasks. Task A addressed the identification of misogynist tweets, while Task B focused on the categorization of misogynist tweets (i.e. *Discredit*, *Derailing*, *Dominance*, *Sexual Harassment & Threats of Violence*, *Stereotype & Objectification*, *Active* and *Passive*). Both IberEval and EVALITA included a task on English tweets. IberEval also included a task on Spanish tweets while EVALITA also featured a subtask on Italian tweets.

Most existing datasets of offensive language contain English data, such as the dataset described by Waseem and Hovy (2016). With regard to publicly-available German datasets for this task, we only know of Ross et al. (2016) who present a dataset of about 500 tweets which has been annotated regarding hate speech. The authors employed a binary categorization scheme. The dataset from Ross et al. (2016) may be too small for some data-hungry learning-based approaches. Being considerably larger, the German dataset produced for the GermEval shared tasks 2018 and 2019 with about 12,000 tweets in total should be a better alternative for such approaches.

3 Task Description

Participants were allowed to participate in one, two or all three subtasks and to submit at most three runs per task.

3.1 Subtask 1: Coarse-grained Binary Classification

Subtask 1 was to decide whether a tweet includes some form of offensive language or not. The tweets had to be classified into the two classes OFFENSE and OTHER. The OFFENSE category covered abusive language, insults, as well as merely profane statements.

3.2 Subtask 2: Fine-grained 4-way Classification

The second subtask involved four categories, a non-offensive OTHER class and three sub-categories of what is OFFENSE in subtask 1. In the case of PROFANITY, profane words are used, however, the tweet does not want to insult anyone. This typically concerns the usage of swearwords (*Scheiße*, *Fuck* etc.) and cursing (*Zur Hölle! Verdammt!* etc.). This can be often found in youth language. Swearwords and cursing may, but need not, co-occur with insults or abusive speech. Profane language may in fact be used in tweets with positive sentiment to express emphasis. Whenever profane words are not directed towards a specific person or group of persons and there are no separate cues of INSULT or ABUSE, then tweets are labeled as simple cases of PROFANITY.

In the case of INSULT, unlike PROFANITY, the tweet clearly wants to offend someone. INSULT is the ascription of negatively evaluated qualities or deficiencies or the labeling of persons as unworthy

[1] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[2] https://sites.google.com/view/ibereval-2018

[3] http://www.evalita.it/2018

(in some sense) or unvalued. Insults convey disrespect and contempt. Whether an utterance is an insult usually depends on the community in which it is made, on the social context (ongoing activity etc.) in which it is made, and on the linguistic means that are used (which have to be found to be conventional means whose assessment as insulting are intersubjectively reasonably stable).

And finally, in the case of ABUSE, the tweet does not just insult a person but represents the stronger form of abusive language. By abuse we define a special type of degradation. This type of degrading consists in ascribing a social identity to a person that is judged negatively by a (perceived) majority of society. The identity in question is seen as a shameful, unworthy, morally objectionable or marginal identity. In contrast to insults, instances of abusive language require that the target of judgment is seen as a representative of a group and it is ascribed negative qualities that are taken to be universal, omnipresent and unchangeable characteristics of the group. (This part of the definition largely co-incides with what is referred to as abusive speech in other research.) Aside from the cases where people are degraded based on their membership in some group, we also classify it as abusive language when dehumanization is employed even just towards an individual (i.e. describing a person as scum or vermin etc.).

3.3 Subtask 3: Implicit vs. Explicit Classification

Implicit offensive language is a form of offensive language where the expression of hate, condemnation, inferiority etc. as directed toward an explicitly or implicitly given target has to be inferred from the ascription of (hypothesised) target properties that are insulting, degrading, offending, humiliating etc. Rather than explicitly expressing their aversion, the writers hint at something degrading, i.e. their tweets imply that the target is unworthy etc.

Offensive tweets that use figurative language such as irony or sarcasm, or a play of words also count as implicit. Implicit offensive statements sometimes are only interpretable in their context. Also, inappropriate casual language while addressing a serious topic is subsumed under implicit offensive language.

The following examples[4] illustrate our notion of implicitness:

1. *Dem Kommentar entnehme ich das auch ihre Schaukel als Kind zu nahe an der Wand gestanden hat.* (*From the commentary I can see that your swing was too close to the wall as a child.*)

2. *Flüchtlinge fliehen nach Deutschland parallel dazu lassen sie ihre Familien in der Heimat sterben sehr ehrenhaft .... .* (*Refugees flee to Germany at the same time they let their families die in their homeland very honourable ...*)

3. *Der arme ... Trauma jeden Tag , Sehnsucht nach Familiennachzug , kein eigenes Haus ... nachvollziehbar ... !* (*The poor ... Trauma every day, longing for family reunion, no own house*)

4. *Es gibt nur ein Maas das ist ein Mittelmass und heisst auch so @HeikoMaas* (*There is only one Maas that is a mediocrity and is also called so @HeikoMaas*)

5. *Also ich habe bei dem Herrn eine deutliche Alters Demenz gesehen.* (*Well I've seen that this man has an obvious age dementia*)

In example 1, it is the potential negative effect of a hypothesised situation that makes the reader understand the ascription of stupidity. Examples 2 and 3 are cases of sarcasm and irony, respectively. Neither is it honourable to leave someone in a dangerous situation (example 2) as the tweet states nor does a refugee suffer trauma just because they do not possess a house in their new host country (example 3). In example 4, a phonetic similarity between a name (*Maas*, the name of a German minister) and a negative concept (*Mittelmaß*, eng. *mediocrity*) suggests inferiority of the target (*Maas*). In example 5, the modal particle (*also*, eng. *well*) and a social distance indicating phrase (*dem Herrn*, eng. *this man*) are inappropriate in a discussion on such a topic (*Demenz*, eng. *dementia*). An honest diagnosis of a disease does not use such casual markers.

If the target is implicit, this might be an indicator of implicitness, but it is neither a necessary nor a sufficient condition. If a tweet comprises both implicit and explicit offensive language, we choose EXPLICIT as a label.

[4] These are examples from the GermEval 2018 corpus. We left misspellings untouched.

3.4 Evaluation Metrics

We evaluate the classification performance by the common evaluation measures *precision*, *recall*, and *F-score*. These measures are computed for each of the individual classes in the three subtasks. For each task, we also compute the *macro-average* precision, recall and F-score as well as the accuracy. We rank systems by their macro-average scores. We do not use accuracy for the ranking since in all three subtasks the class distribution is fairly imbalanced. Accuracy typically rewards correct classification of the majority class.

An evaluation tool computing all of the above mentioned evaluation measures on the three subtasks of the shared task was provided by the organizers prior to the release of the training data. It is publicly available and can be downloaded via the webpage of the shared task.

4 Data Set

As for last year's task, Twitter was the source for our data collection. The reasons why we chose Twitter are a) that unlike other sources, Twitter contains a much higher proportion of offensive language (Wiegand et al., 2018a) and b) that, given the Twitter terms of service, we are able to make our collection publicly available.

4.1 Data Collection

The bulk of the available training data consisted of the training and test data from the first iteration of the shared task in 2018. We newly collected and annotated this year's test data but also some further training data. To do so, we used the same approach to data collection that we had developed for the first iteration. That is, we sampled tweets from the timeline of various users rather than sampling randomly or on the basis of query term-based sampling. The latter two alternatives prove to be either too sparse in yielding offensive instances or too biased and lacking in variety.

We started by heuristically identifying users that regularly post offensive tweets. By sampling from the user's timeline, we obtained offensive tweets that exhibited a more varied vocabulary than we would have obtained by sampling by pre-defined query terms. It also enabled us to extract a substantial amount of non-offensive tweets since only very few users post offensive content exclusively.

Since the majority of last year's data came from the extreme right-wing spectrum and the dominant topic concerned migration, we explicitly added timelines of users from the extreme left-wing spectrum to the 2019 data. Additionally we identified timelines discussing antisemitism in order to increase the topic variance in the data. Most of the user timelines considered for this topic can be assigned to the right-wing spectrum, but we also included timelines of users from other political directions, however we could not identify any timelines that can be assigned to the extreme left-wing spectrum concentrating on the topic. An overview of the data distribution with respect to the political orientation is given in Table 1.

Although this extraction process prevents the data set from becoming biased towards specific topics trending at the point in time when the extraction is run (a problem one typically faces when extracting data from the Twitter stream), we still found certain topics dominating our extracted data. However, this was not as extreme in the 2019 data as it was in the 2018 data, probably due to deliberately incorporating timelines of users from different political extremes. Most of the extracted offensive tweets in 2018 concerned the situation of migrants or the German government. The tweets not considered offensive, however, often addressed different topics. In the 2019 data we still found a stronger representation of certain political parties and some of their representatives, the government and the German state as well as some minorities in the offensive categories. For example, the politician names *Stegner* and *Maas* and the abbreviation *BRD* standing for 'Federal Republic of Germany' were much more often observed in offensive tweets. Since these high-frequency words undoubtedly do not represent offensive terms, we decided to *debias* our data collection by adding further tweets from the already collected timelines, belonging to the class OTHER and containing these terms. If this was not sufficient we added timelines from different political orientations to balance the topic over the classes (see Table 1). Because it was not always possible during the debiasing process to identify user timelines focusing on relevant topics from a different political spectrum, we also sampled further arbitrary tweets containing the terms in question. We specifically sought tweets from across the entire political spectrum. We also deliberately included tweets from users that regularly post highly-critical but inoffensive tweets with respect to the above topics. Otherwise, our data col-

topic/political orientation	ABUSE	INSULT	PROFANITY	OTHER	total
extreme left-wing	64	180	61	1802	2107
extreme right-wing	794	818	177	2137	3926
antisemitism	51	86	22	548	707
non-extreme (debiasing)	1		3	281	285
total	910	1084	263	4768	7025

Table 1: Distribution of topics/political orientation of the user timelines in the 2019 data

lection would allow classifiers to do well simply by recognizing offensive content as the combination of negative polarity and particular topics (e.g. *Stegner*, *Maas* or *BRD*).

When sampling tweets from Twitter, we also imposed certain formal restrictions on the tweets to be extracted (cf. Wiegand et al. (2018b)). They had to be a) written in German, b) contain at least five ordinary alphabetic tokens, c) contain no urls and d) be no retweets. These restrictions are mainly designed to speed up the annotation process (cf. §4.2) by removing tweets that are not relevant to the gold standard.

In splitting our data collection into training and test set, we made sure that any given user's complete set of tweets was assigned to either the training set or the test set. This was done to avoid a situation where classifiers could benefit from learning user-specific writing styles or topic biases.

The data collection was also divided up in such a manner that the training and test sets have a similar class distribution. This is one of the major prerequisites for supervised learning approaches to work effectively.

The third subtask is based on the GermEval2018 data, namely those tweets from the 2018 shared task that are classified as abuse or insult (profanity was left out, because it is by definition explicit offensive language).

4.2 Annotation

4.2.1 Subtasks 1 and 2

Overall, we collected 7,025 new tweets for the 2019 Shared Task. Each of them was manually annotated by one of the organizers of the shared task. All annotators are native speakers of German.

In order to measure inter-annotation agreement, a sample of 300 tweets were annotated by all the annotators independently. We removed all tweets that were marked as HUNH or EXEMPT by at least one annotator. HUNH was used for incomprehensible utterances. We do not require that a sentence is perfectly grammatically well-formed and correctly spelled to be included in our data. However, if a sentence is so erroneous that the annotator does not understand its content, then this sentence was labeled as HUNH and removed. This label also applies if the sentence is formally correct but the annotator still does not understand what is meant by this utterance. Tweets that were marked EXEMPT were ones that only contain abuse or insults representing the view of somebody other than the tweeter; utterances which depend on non-textual information; utterances that are just a series of hashtags and/or usernames, even if they indicate abusive speech (e.g. #crimigrants or #rapefugees); or utterances that are incomplete.

On the remaining 206 tweets, an agreement of $\kappa = 0.59$ was measured between the four annotators. It can be considered moderate (Landis and Koch, 1977). All remaining tweets of the gold standard were only annotated by one of the four annotators.

Table 2 displays the class distribution among the 2019 training and the test set. It comes as no surprise that non-offensive tweets represent the majority class. The most frequent subtype of offensive language are cases of (common) insult followed by abuse. By far the smallest category are profane tweets.

4.2.2 Subtask 3

After an initial phase, where we set up the guidelines, we chose 300 offensive tweets and four annotators classified each tweet as either implicit or explicit offensive language.

Our intention in this first round was to raise a common understanding of implicitness. After harmonization, 247 of the 300 tweets were classified as explicit offensive language (82.33%) while 52 (17.33%) were deemed to be implicit. See Table 3 for pairwise Kappa values.

As we expected, the annotation of implicitness is not an easy task. Accordingly, the agreement is

		training set		test set	
categories		**freq**	**%**	**freq**	**%**
coarse-grained	OFFENSE	1287	32.2	970	32.0
	OTHER	2707	67.8	2061	68.0
fine-grained	ABUSE	510	12.8	400	13.2
	INSULT	625	15.6	459	15.1
	PROFANITY	152	3.8	111	3.7
	OTHER	2707	67.8	2061	68.0
total		3994	100.0	3031	100.0

Table 2: Class distribution on the 2019 training and test set

	B	C	D
A	0.60	0.46	0.54
B		0.48	0.52
C			0.37

Table 3: Pairwise Kappa: 4 annotators, 300 tweets

moderate (most of the time). Two annotators, A and B, almost reached a substantial agreement, while annotators C and D only have a fair agreement. We thus decided to let A and B carry out a substantial part of the annotation.

The annotation of additional 1,800 examples resulted in a Kappa value of 0.51. After harmonization, the Kappa value of A and the gold standard was 0.60, while those of B and the gold standard was 0.82. The rest of the 2,890 tweets (600) were annotated by C and D. The agreement there was 0.42.

		freq	%
training set	IMPLICIT	259	13.20
	EXPLICIT	1699	76.80
test set	IMPLICIT	134	14.37
	EXPLICIT	798	75.63

Table 4: Class distribution subtask 3

Table 4 shows the class distribution for the training and the test data. The whole corpus comprises 8,541 tweets, 2,888 are offensive (33.81%) of which 393 (13.6%) were implicitly offensive.

4.3 Data Format

Our data is distributed in the form of tab-separated value files. An example row representing one tweet for subtasks 1 and 2 is shown in Table 5. As the task is focused only on the linguistic aspect of offensive language, each tweet is represented only by its text in column 1. Meta-data contained in Twitter's json files was not used. The text column is followed by the coarse-grained label in column 2 and the fine-grained label in column 3. Note that we applied no preprocessing to the tweet text with one exception: as shown in Table 5, line breaks were replaced with the special 5-character string `|LBR|` so that each tweet could be stored on one line.

For subtask 3 the data from 2018 was used. In order to provide for an additional layer distinguishing explicit from implicit offensive language, we added an additional column. Three labels are used: IMPLICIT, EXPLICIT or OTHER, see Table 6.

4.4 Sanity checks

To make sure empirically for subtasks 1 and 2 that the combination of last year's data with this year's data was sensible and there were no crucial differences that would actually harm performance, we performed an internal pre-test using last year's winning system by TU Vienna (Padilla Montani and Schüller, 2018). We used all of last year's data as well as this year's new training data as the training set and tested on the new 2019 test set.

We performed a second sanity experiment after the task's evaluation phase because it was only then noticed that there were erroneous labels on items of the 2019 training set. Altogether about 2.9% of the labels were affected: 15 cases of the class ABUSE, 28 cases of the class INSULT and 74 cases of the class OTHER. The PROFANITY class was not affected. We repeated the sanity check on a corrected version of the dataset to evaluate if the errors might have substantially harmed results in the competition.

The results for the initial sanity check on the original, slightly erroneous data are denoted by the rows coarse$_e$ and fine$_e$ in Tables 7 and 8, while the results for the run on the corrected data are denoted

@Ralf_Stegner Oman Ralle..dich mag ja immer noch keiner. Du willst das die Hetze gegen dich aufhört? |LBR| Geh in Rente und verzichte auf die 1/2deiner Pension	OFFENSE	INSULT

Table 5: Data format for subtask 1 and 2 (2019 dataset)

Der einzige, der sich noch darüber freut, dass Merkel auf ihrem Stuhl klebt |LBR| ist der beginnende Dekubitus.	OFFENSE	INSULT	IMPLICIT

Table 6: Data format for subtask 3 (2018 dataset)

by the rows rows coarse$_c$ and fine$_c$. Overall, the results for these sanity checks are very similar to the system's results on last year's tasks, regardless of whether we the training set includes a slight number of errors or not. (The corrected version of the training set is also now publicly available from the shared task homepage.)

For subtask 3, no sanity checks were needed.

5 Submissions and Results

The full set of results for all three subtasks is available at the shared task website.

Table 9 presents descriptive statistics for the scores produced in this year's and last year's iterations of subtask 1 and 2. For subtask 1, coarse-grained classification, we can see that this year the participants' scores are more tightly clustered, yielding a lower standard deviation. For subtask 2, the fine-grained task, there was no similar development.

5.1 Coarse-grained Classification

We received 28 different runs from 12 teams for the binary classification into OFFENSE vs. OTHER. For lack of space, we only show the best 15 runs in Table 10. Compared to the previous year, this year's winning F-score is higher, but very slightly so (76.95 vs. 76.77). Of course, these number cannot be compared directly as they involved different training and test sets.

5.2 Fine-grained Classification

For the second subtask we received 28 different runs from 12 teams for the fine-grained classification. For lack of space, we only show the best 10 runs in Table 11. Compared to last year's results, the winning score is higher by about 0.9% F-score. As in the case of the coarse grained subtask, this cannot be readily interpreted without further investigation.

5.3 Implicit vs. Explicit Classification

Seven groups participated in subtask 3, which was a difficult subtask. Although the best accuracy was 86.77% with a F-score of 73.11, the numbers for the class IMPLICIT were low, the best F-score being 53.93. The subtask is difficult due to the skewed distribution of that class, just 13.9% of the offensive tweets are labeled as implicit.

5.4 General Conclusions Drawn from the Evaluation

5.4.1 System Design

Although in terms of absolute F-scores, the best performing system on all 3 subtasks was a system that employed some form of the latest transformer based language model BERT (Devlin et al., 2019) (i.e. *UPB* on subtasks 1 and 2; *hpiDEDIS* on subtask 3), at least on subtasks 1 and 2 there were systems which did not incorporate BERT (i.e. *TUWienKBS* on subtask 1 and *FoSIL* on subtask 2) but still performed very well (that is within 1% point of the top performing system). Only on subtask 3 is there a larger difference between the best performing system and the best system not employing BERT, i.e. *FoSIL*, with a gap of more than 3.5% points.

BERT seems to be generally effective. All 3 teams that participated in the shared task and incorporated some form of BERT (*UPB*, *hpiDEDIS* and *bertZH*) were among the top performing systems. The variation of BERT that consistently outperformed the other ones is a model pre-trained on 6 million German tweets (*UPB*). The other two teams just fine-tuned the existing pre-trained models.

Surprisingly, for all 3 subtasks there is no system among the top 3 teams which employs *standard* deep-learning architectures, such as LSTMs or CNN. Instead, with *FoSIL* we still find systems that are based on traditional classifiers, such as SVMs. This year's results are also mostly consis-

	OFFENSE			**OTHER**			**average**		
	P	R	F	P	R	F	P	R	F
coarse$_e$	67.23	69.18	68.19	85.29	84.13	84.71	76.26	76.65	76.46
coarse$_c$	68.20	68.76	68.48	85.24	84.91	85.08	76.72	76.84	76.78

Table 7: Results of sanity checks on error-containing and clean test data in coarse setting

		P	R	F
fine$_e$	ABUSE	47.29	41.50	44.21
	INSULT	44.71	36.82	40.38
	OTHER	83.33	88.26	85.72
	PROFANITY	42.02	45.05	43.48
	average	54.34	52.91	53.61
fine$_c$	ABUSE	47.41	41.25	44.12
	INSULT	45.76	38.78	41.98
	OTHER	83.65	88.40	85.96
	PROFANITY	43.10	45.05	44.05
	average	54.98	53.37	54.16

Table 8: Results of sanity checks on error-containing and clean test data in fine-grained setting

tent with last year's results: the best performing systems incorporated some form of word embeddings and some information on the subword level (e.g. character n-grams). Ensemble methods may be effective (*TUWienKBS*) but they seem not to be a crucial ingredient for high scores. The same holds for task-specific lexicons. Of the 3 top-performing systems on the 3 subtasks, only *FoSIL* employed that type of information.

In subtask 3, the best performing systems (*hpi-DEDIS*, *UPB*, rank 1-5 with various runs) were using BERT as a resource (fine-tuned it). Of the 7 participants, 5 used neural approaches (including BERT), i.e. RNN (*inriaFBK*), CNN (*fkie*, *HAU*) and LSTM (*fkie*). Two worked with German FastText (*RGCL*, *FoSIL*), one also considered a Random Forest approach and one also submitted a SVM based run.

5.4.2 Task and Data

With regard to subtask 1, if we compare the difference between the F-score of the best performing system to the median between this year (median: 72.95; best system: 76.95) and last year (median: 69.15; best system: 76.77), we find that the median has risen appreciably (by more than 3% points) while the best score has maintained its level of performance. From that we may conclude that the average system that took part in this year's edition of the shared task is notably stronger than last year.

In terms of the best overall scores that have been achieved in subtasks 1 and 2 in this year's edition of the shared task, there is hardly any improvement. We re-trained last year's winning system on this year's training data and compared the classification on this year's test data (cf. Tables 7 and 8) with the best performing system in this year's competition (cf. Tables 10 and 11). Surprisingly, we obtained only marginally worse results with last year's system (subtask 1: 76.46 vs. 76.95; subtask 2: 53.61 vs. 53.95). Given that this year's training set was larger, this could mean one of two things. First, the additional data might not have helped even though test and training data were otherwise similar because the system was not able to make use of relevant features. Alternatively, the increase in data this year might have been offset by the new data being more difficult so that overall the system reached only the same level of performance as last year. These questions can best be addressed by running the same system on various combinations of this and last year's data, which unfortunately is outside the scope of this overview paper.

All in all, these results underline that the problem of offensive language detection is far from solved. It also suggests that a thorough error analysis is required. Only thus can we learn which systematic errors even the best performing systems make and, hopefully, get ideas how to devise new methods which even solve these types of phenomena.

6 Conclusion

In this paper, we described the second edition of the GermEval Shared on the Identification of Offensive Language. The shared task comprised three tasks, a coarse-grained binary classification task, a fine-grained multi-class classification task and a novel classification task in which explicit tweets had to be separated from implicit ones. In total, 13 groups participated in the shared task submitting 28 runs for each of the first two subtasks and 17 runs for the last subtask.

year	subtask	# teams	# runs	min	max	median	mean	sd
2019	coarse	12	28	54.87	76.95	72.95	71.51	5.67
	fine	12	28	36.83	53.59	46.55	46.63	5.18
	implicit/explicit	7	17	55.37	73.11	68.87	67.19	5.26
2018	coarse	20	51	49.03	76.77	69.15	66.35	8.45
	fine	11	25	32.07	52.71	38.75	39.71	5.00

Table 9: Summary statistics for overall macro F1-scores in the three subtasks and as a reference the figures of last year’s edition

While for the third subtask, the data of the previous edition were augmented by the classification scheme of this new task, for the first and second subtask, completely new tweets were added to the collection. For these two subtasks, we added about 4,000 manually labeled training tweets. Similar to last year, much care was taken in order to provide a relatively unbiased dataset. Unlike the data from the previous edition, the new data also contain offensive language originating from other areas than the extreme right.

Approaches that were effective in last year’s edition, such as supervised classifiers using word embeddings, subword information and ensemble methods, also proved effective in this year’s edition. However, similar effectiveness without less task-specific design could be achieved by classifiers based on the recent BERT model.

Surprisingly, the best system of this year’s system on the coarse-grained task is on a par of last year’s winning system. This result again underlines the difficulty of this task. Further error analyses should be carried in order to determine which types of errors even the best performing systems incur. This would hopefully provide fruitful research directions for future work.

Acknowledgments

We are grateful to the large number of participants whose enthusiastic participation made GermEval 2019 a success. We would like to thank Joane Amrhein for maintaining the shared task home page and mailing lists and supporting the evaluation process. We would also like to thank Sophia Conrad (BA student), Andreas Säuberli (BA student) and Noëmi Aepli (PhD student) for their contribution to subtask 3. All annotators are native speakers of German. We also thank the Konvens 2019 conference organizers for their support. Michael Wiegand was supported by the Leibniz ScienceCampus “Empirical Linguistics and Computational Modeling”, funded by the Leibniz Association under grant no. SAS-2015-IDS-LWC and by the Ministry of Science, Research, and Art (MWK) of the state of Baden-Württemberg.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 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*, pages 54–63.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@ SEPLN*, pages 214–228.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling Online Abuse: A Survey of Automated Abuse Detection Methods. *arXiv e-prints*, August.

Joaquın Padilla Montani and Peter Schüller. 2018. TUWienKBS at GermEval 2018: German Abusive Tweet Detection. Proceedings of GermEval 2018,

	submission		Accuracy			OFFENSE			OTHER			average		
	team	runID	percent	correct	total	P	R	F	P	R	F	P	R	F
1	UPB	coarse_1	79.38	2406	3031	66.26	72.47	69.23	86.45	82.63	84.50	76.35	77.55	76.95
2	UPB	coarse_3	79.38	2406	3031	66.26	72.47	69.23	86.45	82.63	84.50	76.35	77.55	76.95
3	UPB	coarse_2	79.64	2414	3031	67.53	70.10	68.79	85.67	84.13	84.90	76.60	77.12	76.86
4	TUWienKBS	coarse_1	80.04	2426	3031	69.73	66.49	68.07	84.57	86.41	85.48	77.15	76.45	76.80
5	TUWienKBS	coarse_2	79.94	2423	3031	69.34	66.91	68.10	84.68	86.07	85.37	77.01	76.49	76.75
6	FoSIL	coarse_2	79.54	2411	3031	67.68	69.07	68.37	85.30	84.47	84.89	76.49	76.77	76.63
7	FoSIL	coarse_1	79.25	2402	3031	66.73	70.10	68.38	85.59	83.55	84.56	76.16	76.83	76.49
8	hpiDEDIS	coarse_2	78.69	2385	3031	64.86	72.89	68.64	86.45	81.42	83.86	75.66	77.15	76.40
9	hpiDEDIS	coarse_3	79.71	2416	3031	70.66	62.58	66.38	83.29	87.77	85.47	76.98	75.18	76.06
10	hpiDEDIS	coarse_1	76.97	2333	3031	61.54	74.74	67.50	86.78	78.02	82.17	74.16	76.38	75.26
11	inriaFBK	coarse_2	78.55	2381	3031	67.74	62.99	65.28	83.14	85.88	84.49	75.44	74.44	74.93
12	hshl	coarse_1	78.39	2376	3031	68.00	61.34	64.50	82.61	86.41	84.47	75.30	73.88	74.58
13	inriaFBK	coarse_1	78.55	2381	3031	69.18	59.48	63.97	82.11	87.53	84.73	75.65	73.51	74.56
14	rgcl	coarse_2	77.96	2363	3031	81.72	40.10	53.80	77.26	95.78	85.53	79.49	67.94	73.26
15	rgcl	coarse_3	77.37	2345	3031	83.49	36.49	50.79	76.37	96.60	85.30	79.93	66.55	72.63

Table 10: Top 15 runs for subask 1: coarse-grained classification

	submission		accuracy			ABUSE			INSULT			OTHER			PROFANITY			average		
	team	runID	percent	correct	total	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
1	UPB	fine_1	73.61	2231	3031	44.04	60.00	50.79	49.49	32.03	38.89	84.69	88.55	86.57	55.88	17.12	26.21	58.53	49.42	53.59
2	FoSIL	fine_2	71.36	2163	3031	42.57	58.75	49.37	45.30	45.10	45.20	85.88	82.63	84.22	46.15	16.22	24.00	54.98	50.67	52.74
3	FoSIL	fine_1	72.02	2183	3031	42.60	58.25	49.21	46.81	38.34	42.16	84.60	85.30	84.95	53.33	14.41	22.70	56.83	49.08	52.67
4	bertZH	fine_2	74.46	2257	3031	55.29	45.75	50.07	48.09	41.18	44.37	83.43	90.15	86.66	33.75	24.32	28.27	55.14	50.35	52.64
5	UPB	fine_2	71.66	2172	3031	43.33	58.50	49.79	43.51	39.43	41.37	85.67	84.13	84.90	45.10	20.72	28.40	54.40	50.70	52.48
6	UPB	fine_3	73.57	2230	3031	45.61	54.50	49.66	47.21	36.82	41.37	84.69	88.55	86.57	45.00	16.22	23.84	55.63	49.02	52.11
7	TUWienKBS	fine_2	71.86	2178	3031	45.36	41.50	43.34	44.78	38.34	41.31	82.62	87.19	84.84	40.21	35.14	37.50	53.24	50.54	51.86
8	bertZH	fine_1	73.38	2224	3031	54.55	43.50	48.40	43.42	40.96	42.15	82.63	89.33	85.85	41.18	18.92	25.93	55.44	48.18	51.55
9	TUWienKBS	fine_1	73.18	2218	3031	46.20	38.00	41.70	49.82	30.94	38.17	80.58	91.80	85.82	46.38	28.83	35.56	55.75	47.39	51.23
10	hpiDEDIS	fine_3	70.04	2123	3031	44.94	47.75	46.30	39.19	48.58	43.39	85.19	81.76	83.44	40.68	21.62	28.24	52.50	49.93	51.18

Table 11: Top 10 results for subtask 2: fine-grained classification

14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria September 21, 2018.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of the Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6–9, Bochum, Germany.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the EACL-Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 1–10, Valencia, Spain.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018a. Inducing a Lexicon of Abusive Words – A Feature-Based Approach. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, New Orleans, USA.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018b. Overview of the GermEval 2018 shared task on the identification of offensive language. Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria September 21, 2018, pages 1 – 10. Austrian Academy of Sciences, Vienna, Austria.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *SemEval@NAACL-HLT*.

7 Appendix

group id	authors	affiliation	paper title
InriaFBK	Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli and Serena Villata	Inria, France and Fondazione Bruno Kessler, Italy	InriaFBK Drawing Attention to Offensive Language at Germeval2019
hshl	Kristian Rother and Achim Rettberg	Hochschule Hamm-Lippstadt, Germany	German Hatespeech classification with Naive Bayes and Logistic Regression - hshl at GermEval 2019 - Task 2
fkie	Theresa Krumbiegel	Fraunhofer FKIE, Germany	FKIE - Offensive Language Detection on Twitter at GermEval 2019
FraunhoferSIT	Inna Vogel and Roey Regev	Fraunhofer Institute SIT, Germany	FraunhoferSIT at GermEval 2019: Can Machines Distinguish Between Offensive Language and Hate Speech? Towards a Fine-Grained Classification
bertZH	Tim Graf and Luca Salini	University of Zurich, Switzerland	bertZH at GermEval 2019: Fine-Grained Classification of German Offensive Language using Fine-Tuned BERT
FoSIL	Florian Schmid, Justine Thielemann, Anna Mantwill, Jian Xi, Dirk Labudde and Michael Spranger	University of Applied Sciences Mittweida, Germany	FoSIL - Offensive language classification of German tweets combining SVMs and deep learning techniques
h_da	Isabell Börner, Midhad Blazevic, Maximilian Komander and Margot Mieskes	Darmstadt University of Applied Sciences, Germany	2019 GermEval Shared Task on Offensive Tweet Detection h_da submission
HAU	Johannes Schäfer, Tom De Smedt and Sylvia Jaki	University of Hildesheim, Germany and University of Antwerp, Netherlands	HAU at the GermEval 2019 Shared Task on the Identification of Offensive Language in Microposts: System Description of Word List, Statistical and Hybrid Approaches
UPB	Andrei Paraschiv and Dumitru-Clementin Cercel	University Politehnica of Bucharest, Romania	UPB at GermEval-2019 Task 2: BERT-Based Offensive Language Classification of German Tweets
hpiDEDIS	Julian Risch, Anke Stoll, Marc Ziegele and Ralf Krestel	Hasso Plattner Institute, Heinrich Heine University Düsseldorf and University of Passau, Germany	hpiDEDIS at GermEval 2019: Offensive Language Identification using a German BERT model
HUIU	Melanie Andresen, Melitta Gillmann, Jowita Grala, Sarah Jablotschkin, Lea Röseler, Eleonore Schmitt, Lena Schnee, Katharina Straka, Michael Vauth, Sandra Kübler and Heike Zinsmeister	Universität Hamburg (Germany), Universitt Bamberg (Germany), Indiana University (USA)	The HUIU Contribution for the GermEval Shared Task 2
TUWienKBS19	Joaqun Padilla Montani and Peter Schüller	TU Wien, Austria	TUWienKBS19 at GermEval Task 2, 2019: Ensemble Learning for German Offensive Language Detection
RGCL	Alistair Plum, Tharindu Ranasinghe, Constantin Orasan and Ruslan Mitkov	University of Wolverhampton, UK	RGCL at GermEval 2019: Offensive Language Detection with Deep Learning

Table 12: Group IDs, Authors and Paper Titles