Originally published in: Ruppenhofer, Josef/Siegel, Melanie/Wiegand, Miachel (Eds): Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria – September 21, 2018. - Vienna, Austria: Austrian Academy of Sciences, 2018. Pp. 1-10

Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language

Michael Wiegand
Spoken Language Systems
Saarland University
michael.wiegand@lsv.uni-saarland.de

Melanie Siegel
Information Science
Darmstadt University
of Applied Sciences
melanie.siegel@h-da.de

Josef Ruppenhofer
Empirical Linguistics and
Language Modelling
Institut für deutsche Sprache
ruppenhofer@ids-mannheim.de

Abstract

We 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. It comprises two tasks, a coarse-grained binary classification task and a fine-grained multi-class classification task.

The shared task had 20 participants submitting 51 runs for the coarse-grained task and 25 runs for the fine-grained task. Since this is a pilot task, we describe the process of extracting the raw-data for the data collection and the annotation schema. We evaluate the results of the systems submitted to the shared task. The shared task homepage can be found at https://projects.cai.fbi.h-da.de/iggsa/

1 Introduction

Offensive Language is commonly defined as hurtful, derogatory or obscene comments made by one person to another person. This type of language can be increasingly found on the web. As a consequence, many operators of social media websites no longer manage to manually monitor user posts. Therefore, there is a pressing demand for methods to automatically identify suspicious posts.

The 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.

The shared task was endorsed by two of the special interest groups of the German Society for Computational Linguistics and Language Technology (GSCL): the Interest Group on German Sentiment Analysis (IGGSA) as well as the Interest Group on Social Media Analysis.

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. 20 teams participated in the shared task. We describe the participants and their approaches in §5 and give an overview of the results in §6.

2 Related Work

For a detailed summary of related work on the detection of abusive language, we refer the reader to Schmidt and Wiegand (2017). In the following, we will briefly comment on related shared tasks and datasets in German language. We will also provide some information on the GermEval evaluation campaign.

- Kaggle's Toxic Comment Classification Challenge[1] is a shared task in which comments from the English Wikipedia are to be classified. There are 6 different categories of toxity to be identified (i.e. *toxic*, *severe toxic*, *obscene*, *insult*, *identity hate* and *threat*). These categories are not mutually exclusive.
- The shared task on aggression identification[2] includes both English and Hindi Facebook comments. Participants have to detect abusive comments and to distinguish between *overtly aggressive comments* and *covertly aggressive comments*.
- The shared task on Automatic Misogyny Identification (AMI)[3] is jointly run by IberEval[4]

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

[2] https://sites.google.com/view/trac1/shared-task

[3] https://amievalita2018.wordpress.com https://amiibereval2018.wordpress.com

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

and EVALITA[5]. It exclusively focuses on the detection of misogynist tweets on Twitter. There are two subtasks. Task A addresses the identification of misogynist tweets, while Task B focuses 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 include a task on English tweets. IberEval also includes a task on Spanish tweets while EVALITA also includes a subtask on Italian tweets.

We are not aware of any shared task on the detection abusive language that includes German language data. 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. While the dataset from Ross et al. (2016) may be too small for some data-hungry learning-based approaches, we hope that the German dataset we introduce in this shared task is sufficiently large (i.e. more than 8,000 tweets) even for those approaches.

GermEval is a series of shared task evaluation campaigns that focus on Natural Language Processing for the German language. So far, there have been three iterations of GermEval, each with a different type of task: named entity recognition (Benikova et al., 2014), lexical substitution (Tristan Miller et al., 2015) and aspect-based sentiment analysis in social media customer feedback (Wojatzki et al., 2017). GermEval shared tasks have been run informally by self-organized groups of interested researchers.

3 Task Description

Participants were allowed to participate in one or both tasks and submit at most three runs per task.

3.1 Task 1: Coarse-grained Binary Classification

Task 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 Task 2: Fine-grained 4-way Classification

The second task involved four categories, a nonoffensive OTHER class and three sub-categories of what is OFFENSE in Task 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 (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.).

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

3.3 Evaluation Metrics

We evaluate the classification performance by the common evaluation measures *precision*, *recall*, and *F1-score*. These measures are computed for each of the individual classes in the two tasks. For each task, we also compute the *macro-average* precision, recall and F1-score. We also compute accuracy. We rank systems by their macro-average scores. We do not use accuracy since in both tasks the class distribution is fairly imbalanced. Accuracy typically rewards correct classification of the majority class.

An evaluation tool computing all of the above evaluation measures on the two tasks 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 a source for our data collection, we chose Twitter. Thus we are able to make our collection publicly available. Unlike existing corpora, Twitter also contains a much higher proportion of offensive language (Wiegand et al., 2018).

4.1 Data Collection

Much care was taken in sampling the tweets for our gold standard. Although a natural sample of tweets would represent the most unbiased form of data, we decided against it. A sample of a few thousand tweets would have resulted in just too few occurrences of offensive language as the proportion of offensive tweets is known to be generally low (Schmidt and Wiegand, 2017). We also decided against sampling by specific query terms (as Waseem and Hovy (2016) suggest) since our initial experiments showed that using offensive query terms, such as *Idiot* or *Schmarotzer*, greatly reduced the variety of offensive terms occurring in the retrieved tweets.[6]

Instead, we sampled tweets from the timeline of various users. In total, we considered about 100 different users. We started by heuristically identifying users that regularly post offensive tweets. By sampling from their timeline, we obtained offensive tweets that exhibited a more varied vocabulary than we would have obtained by sampling by predefined query terms. It also enabled us to extract a substantial amount of non-offensive tweets since only very few users *exclusively* post offensive content.

Although this extraction process prevents the dataset 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. Most of the extracted offensive tweets concerned the situation of migrants or the German government. The tweets not considered offensive, however, often addressed different topics. For example, the politician names *Maas* and *Merkel* and the common noun *Flüchtlinge* 'refugees' were almost exclusively observed in offensive tweets. Since these high-frequency words undoubtedly do not represent offensive terms, we decided to *debias* our data collection by sampling further arbitrary tweets containing these terms. We specifically sought tweets from across the entire political spectrum. We also deliberately included tweets from users that regularly post highly-critical tweets with respect to the above topics. Otherwise, our data collection would allow classifiers to score well that simply infer offensive content by observing a negative polarity co-occurring with particular topics (e.g. *Maas*, *Merkel* or *Flüchtlinge*).

When sampling tweets from Twitter, we also imposed certain formal restrictions on the tweets to be extracted. They are as follows:

(1) Each tweet had to be written in German.

(2) Each tweet had to contain at least five ordinary *alphabetic* tokens.

(3) No tweet was allowed to contain any URLs.

(4) No tweet was allowed to be a retweet.

All of 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. (2) was included to remove tweets that just function as an advertisement or spam. We wanted to exclude URLs (3) since our data collection should be self-contained to the degree possible.[7] We avoid retweets since they represent a form of reported content where it is often difficult to decide whether the views expressed in the reported content are shared by the user retweeting or not.

[6] Our observation was that the overwhelming proportion of retrieved tweets would contain just the query words as offensive terms.

[7] The offensive nature of tweets with an URL often only becomes visible by taking into account their linked content.

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. In this way, we wanted to avoid that classifiers could benefit from learning user-specific information. For example, if a user, who very often posts offensive tweets has a very idiosyncratic writing style and his/her tweets were distributed across training and test set, then a classifier could exploit the knowledge about the writing style in order to infer offensive language. Such a classifier would not really have learned to detect offensive language but a very specific writing style which, beyond that given dataset, would not be of any use for detecting offensive language.

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.

4.2 Annotation

Each tweet of the resulting data collection with an overall size of 8541 tweets was manually annotated by one of the three 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 the three annotators in parallel. We removed all tweets that were marked as HUNH or EXEMPT at least by 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 are EXEMPT from the subtyping annotation involve tweets which only contain abuse or insults that represent 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 240 tweets, an agreement of $\kappa = 0.66$ was measured. It can be considered substantial (Landis and Koch, 1977). All remaining tweets of the gold standard were only annotated by one of the three annotators.

Table 1 displays the class distribution among the 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 abuse followed by (common) insults. By far the smallest category are profane tweets.

4.3 Data Format

Our data is distributed in the form of tab-separated value files. An example row representing one tweet is shown in Table 2. 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 2, line breaks were replaced with the special 5-character string `|LBR|` so that each tweet could be stored on one line.

5 Participants and Approaches

Overall, we had 20 teams participating in the shared task. All teams participated in Task 1 and 11 of them took part in Task 2.

Across both tasks, the teams made use of a variety of approaches. Below, we identify some major trends and commonalities between the teams. For a detailed description of the systems, we refer readers to the dedicated system description papers.

5.1 Preprocessing

Tokenization 9 teams mention tokenization as a preprocessing step in their papers. Most used tokenizers adapted to social media: 3 teams used the TweetTokenizer in nltk (Bird et al., 2009), one team used the SoMaJo social media tokenizer (Proisl and Uhrig, 2016), one team used twokenize (Owoputi et al., 2013) and one team developed an extension of the tokenizer of Baziotis et al. (2017). Of the others, one team used the tokenizer in spaCy[8], one team split based mostly on punctuation and the last team did not give any details about its tokenizer.

POS-Tagging 6 teams used POS-Tagging. In most cases, the POS-tags were not produced by a stand-alone tagger but derived from a more complex software tool such as spaCy, the TextBlob[9]

[8] https://spacy.io/

[9] https://github.com/sloria/TextBlob

		training set		test set	
categories		**freq**	**%**	**freq**	**%**
coarse-grained	OFFENSE	1688	33.7	1202	34.0
	OTHER	3321	66.3	2330	66.0
fine-grained	ABUSE	1022	20.4	773	21.9
	INSULT	595	11.9	381	10.8
	PROFANITY	71	1.4	48	1.4
	OTHER	3321	66.3	2330	66.0
total		5009	100.0	3532	100.0

Table 1: Class distribution on training and test set.

@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 2: Data format

package or the ParZu dependency parser (Sennrich et al., 2013).

Lemmatization and stemming 5 systems used lemmatization. Three teams used spaCy, and one team each used the TreeTagger (Schmid, 1995) and ParZu. 2 teams used stemming.

Parsing Only two teams used parsing, one the ParZu parser (Sennrich et al., 2013) and the other the mate-tools parser (Björkelund et al., 2010).

5.2 Lexical Resources

While 8 teams used no task-specific lexicon, 8 other teams used one or more publicly available lexicons, and 7 teams created a new lexicon.[10] 9 teams used polarity lexicons, chief among them PolArt (Klenner et al., 2009), PolarityClues (Waltinger, 2010) and SentiWS (Remus et al., 2010), and 8 teams used dictionaries containing swearwords, slurs or offensive words. Several teams expanded available polarity of swearword lexicons. One team translated and post-edited the English dictionary of abusive terms provided by Wiegand et al. (2018).

5.3 Word Vectors

15 teams used pre-trained word embeddings in their systems. The most commonly used vectors were those provided by spinningbytes (word2vec, fasttext) on the one hand and those provided by the organizers (word2vec) on the other hand. Some teams trained on tweet collections of their own. Two teams pursuing a cross-lingual or translation approach used multi-lingual word embeddings, the aligned languages being German and English in both cases. One team used embeddings only for the purpose of lexicon expansion but not as a feature fed to their classifier.

5.4 Classifiers

The classifiers used involve a fairly broad variety of familiar non-neural types as well as (variations on) recent neural network-type classifiers. Among the non-neural types, SVMs were the most common type. 12 teams used a flavor of SVM, either as a baseline or their main system. Logistic regression was used by 7 teams, in two cases as a meta-classifier. Decision Trees were used by 2 teams and 1 team used a Naive Bayes classifier. Among the neural network classifiers common recent architectures are found: CNN (10 teams), LSTM and variants (11 teams), GRU (6 teams), as well as combinations of these.

6 Submissions and Results

The full set of results for both tasks is available at the shared task website.

A high-level summary of the results is given in Table 3, which provides summary statistics on the macro-average F1 score that was used as the official ranking criterion in the shared task. As the table shows, the scores achieved span a substantial range: more than 25% points in the case of the coarse task and more than 20% points in the case

[10] The publicly available lexicons used were often ones that the shared task organizers had pointed out on the shared task's web pages.

of the fine-grained task.

6.1 Coarse-grained Classification

We received 51 different runs from 20 teams for the binary classification into OFFENSE vs OTHER. For lack of space, we only show the best 15 runs in Table 4. As a baseline, we also included the performance of majority-class classifier always predicting the majority class OTHER.

6.2 Fine-grained Classification

We received 25 different runs from 10 teams for the fine-grained task that distinguishes three sub-types of offensive language from OTHER. We report the best 10 submissions in Table 5. As a baseline, again, we included the performance of majority-class classifier always predicting the majority class OTHER.

6.3 General Conclusions Drawn from the Evaluation

6.3.1 System Design

Given the diversity of approaches and the large number of participating groups in this shared task, it is difficult to draw general conclusions about the effectiveness about specific types of features.

With regard to the choice of classifiers, there is a competition between traditional feature-based supervised learning (typically represented by SVMs) and the more recent deep learning methods. Undoubtedly, most top performing systems in both shared tasks employed deep learning (e.g. *spMMMP, uhhLT, SaarOffDe, InriaFBK*), yet the top performing system in Task 1 and the second-best performing system in Task 2 (i.e. *TUWienKBS*) exclusively employed traditional supervised learning. This team even explicitly states in its participation paper that the usage of deep learning did not improve their results. This makes us wonder whether the frequent occurrence of such methods in top performing systems is just a result of the current popularity of deep learning algorithms and whether traditional engineering is not similarly effective (at least for the classification task in GermEval 2018). We also note that there was quite a bit of variation among the specific deep learning approaches used. It was not necessarily the most complex approach that produced the best results. For example, *SaarOffDe* with its straightforward approach of using RNNs and CNNs produced top scores. The scores of systems employing complex transfer-learning (e.g. *spMMMP*, *InriaFBK* or *uhhLT*) are not necessarily better.

Although overall it may not always be a crucial aspect of system design, the usage of ensemble classification seems to very often improve classification approaches (e.g. *Potsdam*, *RuG*, *SaarOffDe*, *TUWienKBS*, *UdSW*).

With regard to traditional feature engineering, the features found effective very much reflect the insights of recent research on English data, particulary the extensive study presented in Nobata et al. (2016). Several submissions include a combination of word embeddings, character n-grams and some form of (task-specific) lexicon. Both HaUA and *UdSW* report that high performance scores can already be achieved with a classifier solely relying on a lexicon. Yet both groups show that such classifiers can be outperformed by classifiers using additional (typically more generic) features, e.g. character n-grams.

The usage of datasets from other languages (typically English) to augment the training data provided by GermEval may be a very popular idea (e.g. *InriaFBK*, *hpiTM*, *UdSW*, *spMMMP*), however, the results of this shared task do not support systematic effectiveness.[11] There are two issues that may stand in the way. Firstly, the definition of abusive language varies throughout the different datasets. Secondly, the predominant type of abuse may be different: Not every English dataset on abusive language detection similarly has so many abusive comments towards migrants as the GermEval dataset.

6.3.2 Task and Data

Overall, we can conclude that the task of identifying offensive language on German tweets is doable. However, with the highest F-scores up to 76% F1-score on Task 1 and 52% on Task 2, the task is clearly far from solved. If we consider the large span of different F1-scores within the same task (i.e. 27% points on Task 1 and 20% points on Task 2), we also have to acknowledge that building classifiers that achieve top scores is <u>not</u> a trivial undertaking.

The overall performance scores achieved on Task 2 are considerably lower than on Task 1. This does not come as a surprise as Task 2 is considerably more difficult, having 4 instead of 2 classes. More-

[11] UdSW reports that no matter how crosslingual information is added to a classifier, the performance compared to a monolingual classifier drops.

task	# teams	# runs	min	max	median	mean	sd
coarse	20	51	49.03	76.77	69.15	66.35	8.45
fine	11	25	32.07	52.71	38.76	39.71	5.00

Table 3: Summary statistics for overall macro-F1 scores in the two tasks

over, for some classes, particularly PROFANITY there are simply too few instances in the dataset (Table 1).

On several comparable English datasets, much higher classification scores have been reported (Agrawal and Awekar, 2018; Badjatiya et al., 2017). Again, there may be several reasons for that. German is undoubtedly more difficult than English. Due to its higher degree of inflection and compounding, the issue of data sparseness is more prominent. Additionally, we took great efforts in removing biases from the dataset allowing classifiers to overfit (§4.1). For example, we found that if we were to eliminate the constraint that tweets in training and test data have to originate from different users, performance of supervised classifiers would increase by approximately 7% points in F1-score.

Although a proper error analysis is beyond the scope of this overview paper, we inspected the output of the best performing systems and found that while offensive utterances that contain predictive keywords, also referred to as *explicit offense* (Waseem et al., 2017), are mostly detected, offensive utterances that lack such keywords, also referred to as *implicit offense* (Waseem et al., 2017), are mostly missed. Examples (5)-(9) display some of the latter tweets. Clearly, many of these cases require world knowledge and thus remain out of reach for systems that solely employ text classification.

(5) Ich verstehe immer weniger, warum die Polen, Tschechen und Ungarn unsere vorbildliche Migrationspolitik nicht mitmachen wollen. Ist es denen nicht langweilig mit Weihnachtsmärkten so ganz ohne Barrieren, Polizisten und Nagelbomben?

(6) Sei mal ehrlich, wie sollen man Frauen noch ernst nehmen?

(7) Zion wird sein Nürnberg jetzt erleben!

(8) Wenn wir Glück haben, wird China die Welt beherrschen. Wenn wir Pech haben, der Islam.

(9) Da zeigt sich leider mal wieder dass uns der Fall der Mauer nicht nur viel Gutes gebracht hat sonder eben auch @RenateKuenast. #falldermauer

A final aspect of the task design and evaluation that leads to significantly lower scores on the fine-grained task is the combination of macro-F1-based scoring and the inclusion of a very low-frequency class among the labels, namely PROFANITY. Performance on that class was low even for the overall best teams (cf. Table 5), dragging down the macro-F1 score. By comparison, the accuracy for the fine-grained task is only about 6% lower than for the coarse-grained task.

7 Conclusion

In this paper, we described the pilot edition of the GermEval Shared on the Identification of Offensive Language. The shared task comprises two tasks, a coarse-grained binary classification task and a fine-grained multi-class classification task. 20 groups submitted to the former task while 10 groups submitted to the latter task.

Our results show that both tasks are doable but difficult and far from solved. In terms of features and classifiers, there is no clear winner. While many deep-learning approaches produce good scores, traditional supervised classifiers may produce similar scores. Word embeddings, character n-grams and lexicons of offensive words are popular features, but a robust system does not necessarily have to include all three components. Ensemble methods mostly help. The effectiveness of crosslingual methods is debatable. Implicitly offensive language seems particularly difficult.

Though much care was taken in creating the annotated data of the shared task, it is not clear in how far the top performing systems in our shared task overfit to the dataset we created. Therefore, an obvious extension to this task that could shed more light onto the question of generalization would consist of including data from additional domains.

We introduced a new dataset of 8,000 annotated tweets as part of this shared task. All this data has been made publicly available to the research community via the shared task website.

	Submission		Accuracy			Offense			Other			Average		
	Team	RunID	Percent	Correct	Total	P	R	F	P	R	F	P	R	F
1	TUWienKBS	coarse_1	79.53	2809	3532	71.87	65.47	68.52	82.97	86.78	84.83	77.42	76.13	76.77
2	spMMMP	coarse_2	78.85	2785	3532	74.65	57.32	64.85	80.34	89.96	84.88	77.49	73.64	75.52
3	spMMMP	coarse_1	78.60	2776	3532	73.98	57.24	64.54	80.25	89.61	84.67	77.11	73.43	75.22
4	uhhLT	coarse_3	77.49	2737	3532	66.29	68.89	67.56	83.62	81.93	82.77	74.96	75.41	75.18
5	SaarOffDe	coarse_3	77.27	2729	3532	66.72	66.22	66.47	82.64	82.96	82.80	74.68	74.59	74.64
6	SaarOffDe	coarse_1	77.32	2731	3532	67.12	65.39	66.25	82.38	83.48	82.92	74.75	74.43	74.59
7	InriaFBK	coarse_1	76.90	2716	3532	66.11	65.89	66.00	82.43	82.58	82.50	74.27	74.23	74.25
8	InriaFBK	coarse_2	78.20	2762	3532	73.18	56.74	63.92	80.00	89.27	84.38	76.59	73.00	74.25
9	SaarOffDe	coarse_2	76.50	2702	3532	65.68	64.81	65.24	81.97	82.53	82.25	73.83	73.67	73.75
10	InriaFBK	coarse_3	77.24	2728	3532	70.43	57.07	63.05	79.83	87.64	83.55	75.13	72.36	73.72
11	HaUA	coarse_1	76.70	2709	3532	72.91	50.17	59.44	77.86	90.39	83.65	75.38	70.28	72.74
12	UdSW	coarse_3	75.62	2671	3532	66.47	57.24	61.51	79.42	85.11	82.16	72.94	71.17	72.05
13	DFKILT	coarse_2	76.02	2685	3532	77.95	41.18	53.89	75.60	93.99	83.80	76.77	67.59	71.89
14	Potsdam	coarse_3	75.91	2681	3532	72.41	47.17	57.13	76.90	90.73	83.24	74.66	68.95	71.69
15	uhhLT	coarse_1	75.42	2664	3532	71.52	46.17	56.12	76.52	90.52	82.93	74.02	68.34	71.07
	majority-class classifier		65.97	2330	3532	-N/A-	-N/A-	-N/A-	65.97	100.00	79.50	32.98	50.00	39.75

Table 4: Top 15 runs for Task 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	uhhLT	fine_3	73.67	2602	3532	54.71	51.88	53.25	55.19	30.71	39.46	81.13	88.93	84.85	36.36	25.00	29.63	56.85	49.13	52.71
2	TUWienKBS	fine_1	74.52	2632	3532	63.70	44.50	52.40	50.87	38.32	43.71	80.83	91.42	85.80	17.14	25.00	20.34	53.14	49.81	51.42
3	uhhLT	fine_2	72.79	2571	3532	56.64	47.99	51.96	46.39	35.43	40.18	80.52	88.37	84.26	20.69	12.50	15.58	51.06	46.07	48.44
4	uhhLT	fine_1	70.44	2488	3532	49.92	42.82	46.10	43.80	13.91	21.12	76.61	90.26	82.88	33.33	2.08	3.92	50.92	37.27	43.04
5	InriaFBK	fine_2	70.50	2490	3532	58.99	31.82	41.34	37.76	29.13	32.89	76.46	91.46	83.29	5.88	4.17	4.88	44.77	39.15	41.77
6	InriaFBK	fine_3	68.66	2425	3532	54.24	37.26	44.17	29.75	28.35	29.03	77.57	86.95	81.99	11.54	6.25	8.11	43.27	39.70	41.41
7	spMMMP	fine_3	67.89	2398	3532	48.86	38.68	43.18	32.43	18.90	23.88	75.57	86.82	80.81	19.05	8.33	11.59	43.98	38.18	40.88
8	InriaFBK	fine_1	67.89	2398	3532	51.64	30.53	38.37	30.24	33.33	31.71	77.04	87.25	81.83	12.50	4.17	6.25	42.85	38.82	40.74
9	fkieITF	fine_1	68.74	2428	3532	66.36	18.89	29.41	34.31	18.37	23.93	71.21	94.89	81.36	33.33	2.08	3.92	51.30	33.56	40.58
10	RuG	fine_1	69.42	2452	3532	53.29	31.44	39.54	43.17	15.75	23.08	73.34	92.19	81.69	12.50	2.08	3.57	45.57	35.35	39.82
	majority-class classifier		65.97	2330	3532	-N/A-	-N/A-	-N/A-	-N/A-	-N/A-	-N/A-	65.97	100.00	79.50	-N/A-	-N/A-	-N/A-	16.49	25.00	19.87

Table 5: Top 10 results for Task 2: fine-grained classification

Acknowledgments

We would like to thank Torsten Zesch for providing very constructive feedback in the early stages of producing the gold standard data of the shared task.

We would like to thank Markus Meyer for maintaining home page and mailing lists and supporting the evaluation process.

Michael Wiegand was partially supported by the German Research Foundation (DFG) under grant WI 4204/2-1.

References

Sweeta Agrawal and Amit Awekar. 2018. Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms. In *Proceedings of the European Conference in Information Retrieval (ECIR)*, pages 141–153, Grenoble, France.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 759–760, Perth, Australia.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August. Association for Computational Linguistics.

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Padó. 2014. GermEval 2014 Named Entity Recognition Shared Task: Companion Paper. In *Workshop Proceedings of the KONVENS Conference*, pages 104–112, Hildesheim, Germany.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, COLING '10, pages 33–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Manfred Klenner, Angela Fahrni, and Stefanos Petrakis. 2009. Polart: A robust tool for sentiment analysis.

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

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 145–153, Republic and Canton of Geneva, Switzerland.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 380–390.

Thomas Proisl and Peter Uhrig. 2016. Somajo: State-of-the-art tokenization for german web and social media texts. In *WAC@ACL*.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. Sentiws-a publicly available german-language resource for sentiment analysis. In *LREC*.

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.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop. Dublin*.

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.

Rico Sennrich, Martin Volk, and Gerold Schneider. 2013. Exploiting synergies between open resources for german dependency parsing, pos-tagging, and morphological analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 601–609.

David J. Tristan Miller, Darina Benikova, and Sallam Abualhaija. 2015. GermEval 2015: LexSub A Shared Task for German-language Lexical Substitution. In *Proceedings of GermEval 2015: LexSub*, pages 1–10, Essen, Germany.

Ulli Waltinger. 2010. Germanpolarityclues: A lexical resource for german sentiment analysis. In *LREC*.

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.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the ACL-Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. 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 Wojatzki, Eugen Ruppert, Sarah Holschneider, Torsten Zesch, and Chris Biemann. 2017. GermEval 2017: Shared Task on Aspect-based Sentiment in Social Media Customer Feedback. In *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 1–12, Berlin, Germany.