Overview of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments

Julian Risch[1], Anke Stoll[2], Lena Wilms[2], and Michael Wiegand[3]

[1]deepset
[1]julian.risch@deepset.ai
[2]Department of Social Sciences, Heinrich Heine University Düsseldorf
[2]anke.stoll@hhu.de, lena.wilms@hhu.de
[3]Digital Age Research Center, Alpen-Adria-Universität Klagenfurt
[3]michael.wiegand@aau.at

Abstract

We present the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. This shared task comprises three binary classification subtasks with the goal to identify: toxic comments, engaging comments, and comments that include indications of a need for fact-checking, here referred to as fact-claiming comments. Building on the two previous GermEval shared tasks on the identification of offensive language in 2018 and 2019, we extend this year's task definition to meet the demand of moderators and community managers to also highlight comments that foster respectful communication, encourage in-depth discussions, and check facts that lines of arguments rely on. The dataset comprises 4,188 posts extracted from the Facebook page of a German political talk show of a national public television broadcaster. A theoretical framework and additional reliability tests during the data annotation process ensure particularly high data quality. The shared task had 15 participating teams submitting 31 runs for the subtask on toxic comments, 25 runs for the subtask on engaging comments, and 31 for the subtask on fact-claiming comments. The shared task website can be found at https://germeval2021toxic.github.io/SharedTask/.

1 Introduction

User-generated content on the web, particularly on social media, has become a regular part of our everyday life. Given the heavy increase of such content within the last decade, the demand for approaches to classify online content automatically is more pressing than ever. Two previous GermEval shared tasks (Wiegand et al., 2018; Struß et al., 2019) mark important references for research teams from both academia and industry that develop and evaluate approaches to detect offensive language in German-language online discussions. With this year's edition of GermEval, we want participants to go beyond the identification of offensive comments. To this end, we extend the focus to two other classes of comments that are highly relevant to moderators and community managers on online discussion platforms: engaging comments, which should be considered to be highlighted and fact-claiming comments, which should be considered as a priority for fact-checking. This shift aims to bridge the gap between the theoretical view on comment classification and the practical needs of discussion moderators.

GermEval is a series of shared task evaluation campaigns that focus on natural language processing for the German language and has been held since 2014. The topics of the individual shared tasks range from named entity recognition, over lexical substitution, sentiment analysis, and hierarchical classification of blurbs to the identification of offensive language. Teams from both academia and industry are invited to develop and evaluate their approaches on datasets provided by the organizers. The shared tasks are run informally by self-organized groups of interested researchers and are endorsed by special interest groups within the German Society for Computational Linguistics (GSCL).

The remainder of this paper is structured as follows. We describe the task in Section 2 and give an overview of related work addressing the subtasks in Section 3. The dataset is described in detail in Section 4. In Section 5, we briefly comment on the evaluation we conducted, while in Section 6, we discuss the results. Section 7 concludes the paper.

2 Task Description

In this section, we detail the different subtasks of the shared task. Teams could participate either

Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments
co-located with KONVENS

in all three subtasks or just in one or two of the following subtasks. Every team was allowed to submit at most three runs per subtask.

Subtask 1: Toxic Comment Classification. Toxic, offensive, or hateful language in social media and online discussion platforms remains a widespread and particularly pressing problem. Research in the field of communication science has shown that the occurrence of hate speech in online discussions decreases quality perceptions of participants and observers and may trigger stereotypical thinking, hateful commenting behavior or even withdrawal from the debate (Hsueh et al., 2015; Prochazka et al., 2018; Ziegele et al., 2018). While the automatic detection of toxic content is considered to be a promising approach in tackling this problem, it remains challenging and new approaches are constantly being developed. With this subtask we continue the series of previous GermEval Shared Tasks on Offensive Language Identification (Wiegand et al., 2018; Struß et al., 2019).

Subtask 2: Engaging Comment Classification. Normative approaches such as Online Deliberation Theory (Friess and Eilders, 2015) assume that rational, respectful, and reciprocal comments contribute to fostering constructive and non-violent exchange among discussants (Stroud et al., 2015). Such comments can even increase the perceived quality of the related news articles (Ziegele et al., 2018). Therefore, community managers and moderators increasingly express interest in identifying such valuable user comments, for example, to highlight them and to give them more visibility (Risch and Krestel, 2020). We refer to these comments as engaging comments. Engaging comments have been previously defined as comments that make readers join a discussion, e.g. by posting a reply or reacting with a thumbs up/thumbs down (Risch and Krestel, 2020). In this shared task, we expand the definition in favor of comments that meet communication standards of deliberative quality (Ziegele et al., 2018), namely rationality, reciprocity, and mutual respect (Gutmann and Thompson, 1998).

Subtask 3: Fact-Claiming Comment Classification. Beyond the challenge to ensure non-hostile debates, platforms and moderators are under pressure to act due to the rapid spread of misinformation and disinformation. Platforms need to review and verify information that has been posted to meet their responsibility as information providers and distributors. As a result, there is an increasing demand for systems that automatically identify comments that should be fact-checked manually. Note that this subtask is neither about the fact-checking itself nor about the identification of fake news. Instead, the identification of fact-claiming comments should be regarded as an important preprocessing step for manual fact-checking.

3 Related Work

Detection of Toxic Comments. The detection of toxicity, which may also be referred to as *offensive language* (Razavi et al., 2010), *abusive language* (Nobata et al., 2016), *hate speech* (Warner and Hirschberg, 2012), or *incivility* (Stoll et al., 2020) is currently one of the most active fields in natural language processing. For a recent overview of different approaches, we refer the reader to Schmidt and Wiegand (2017) or Fortuna and Nunes (2018), and to Vidgen and Derczynski (2020); Risch et al. (2021) for a comprehensive overview of existing datasets. There has also been a high number of different shared tasks on this topic. For English, several of these shared tasks have been organized as part of the SemEval shared task series (Zampieri et al., 2019; Basile et al., 2019; Zampieri et al., 2020; Pavlopoulos et al., 2021). For German, there have also been two editions of GermEval focusing on this task (Wiegand et al., 2018; Struß et al., 2019). The major difference between those two editions and this year's subtask on toxic comments is the data source. While the data by Wiegand et al. (2018) and Struß et al. (2019) exclusively comprise tweets, this shared task deals with Facebook posts.

Detection of Engaging Comments. The task of detecting engaging comments is motivated by the idea to highlight comments that encourage and foster reasoned and civil discussions (Ziegele et al., 2018). Napoles et al. (2017b) laid groundwork by creating an annotated dataset of engaging, respectful, and informative conversations. They identified characteristics of these conversations, such as being on-topic of the discussed news article and persuasive but not sarcastic or mean. The authors used these characteristics in their follow-up work to automatically identify these conversations (Napoles et al., 2017a). Kolhatkar and Taboada (2017) introduce another publicly available dataset and use editor picks of comments posted on the website of the New York Times as examples of constructive comments. Examples of non-constructive com-

Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments
co-located with KONVENS

ments comprise a subset of comments from non-constructive threads in the dataset by Napoles et al. (2017b). While Risch and Krestel (2020) applied deep learning methods to identify engaging comments automatically, there has been no related work on transformer-based models for this task.

Detection of Fact-Claiming Comments. Detecting check-worthy factual claims recently gained increasing attention – not least because of false claims spread in the context of presidential elections or COVID-19. Hassan et al. (2017) present a semi-automated approach for fact-checking, including automated querying of a knowledge base. Only if querying the knowledge base fails and if several other criteria are met, a claim is considered check-worthy according to their approach. As a follow-up work, they released the *ClaimBuster* dataset, which can be used as a training dataset for identifying check-worthy claims (Arslan et al., 2020). Another publicly available dataset comprises claims made in political debates (Patwari et al., 2017). There is a series of shared tasks on automatic identification and verification of claims in social media, called *CLEF - CheckThat! Lab* (Nakov et al., 2018; Elsayed et al., 2019; Barrón-Cedeno et al., 2020; Nakov et al., 2021). Note that fact-checking of news articles, often referred to as fake news detection, is different from fact-checking of user comments reacting to an article. These two tasks require different approaches, such as taking into account a much longer text or the reputation of the source.

4 Data & Resources

We manually annotated a dataset of more than 4,000 Facebook user comments, which is drawn from the Facebook page of a German political talk show of a national public television broadcaster. The user comments usually revolve around the political topic discussed in a particular edition of the show and contain feedback to political standpoints, the performance of talk show guests and the TV format as a whole. The training dataset contains more than 3,000 comments that were posted in the time span from January to July 2019. To constitute a realistic use case, the test dataset includes comments on editions of the show that were aired after the period of the training dataset. It includes about 1,000 comments that were posted in the time span from September to December 2020. We deliberately decided against producing our training and test data via random sampling to avoid similar word distributions in both data sets. Further, since different people post comments to different editions of the talk show, it is unlikely that our dataset is dominated by the same person posting comments of a particular category (e.g. toxic comments) to any topic: our training data contain user comments of 157 especially active users debating in 141 discussion threads. Therefore, we consider a topic bias and person bias (Wiegand et al., 2019) unlikely. The dataset is released in anonymized form, which means that all user information and comment IDs have been removed.

For annotating our dataset, we made use of a theory-based annotation scheme, which is designed to identify fine-grained forms of toxic and engaging commentary behavior as well as fact-claiming in online discussions (Wilms et al., 2021). An overview of the resulting fine-grained subcategories used in the annotation can be found in Table 1. For the shared task, these subcategories have been subsumed to the three main categories of the subtasks (i.e. toxic, engaging and fact-claiming comments) in a second step. The publicly released dataset only contains the annotation for these three coarse-grained categories.

The dataset we release contains 4,188 Facebook comments (training data = 3,244, test data = 944), which were labeled by trained annotators. High annotation quality was ensured by intensive annotator training as well as intercoder reliability testing using Krippendorff's alpha.[1] Apart from the discussion topic and the user id of a comment, the annotators had no access to further context information. However, it must be noted, that during their annotation, the annotators gained a certain insight into the course of the discussion, which allowed them to interpret the correct meaning of ambiguous statements. Table 1 provides an extensive summary on annotation instructions, frequency distribution and intercoder reliability for both, the main categories as well as the fine-grained subcategories.

In the following, we provide a list of the fine-grained communication features that constitute each of the three main categories, i.e., toxic, engaging and fact-claiming comments. Annotators assigned a particular main category if they identified at least one underlying communication feature.

[1] Krippendorff's alpha corrects for random agreement between coders by relating the observed mean deviation to the assumed mean deviation of a random agreement (Krippendorff, 2018).

Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments
co-located with KONVENS

	Training Data			Test Data		
	Frequency		Intercoder Reliability	Frequency		Intercoder Reliability
	n	%	K-Alpha	n	%	K-Alpha
Subtask 1: Toxic comments	1122	34.5		504	46.2	
Screaming Implying volume by using all-caps at least twice	163	5.0	0.88	101	9.2	0.88
Vulgar language Use of obscene, foul or boorish language	190	5.8	0.73	37	3.4	0.86
Insults Swear words and derogatory statements	205	6.3	0.83	79	7.2	0.83
Sarcasm Ruthless, biting mockery	419	12.9	0.89	295	27.0	0.73
Discrimination Disparaging remarks about entire groups with sweeping condemnation	104	3.2	0.83	145	13.3	0.76
Discrediting Attempt to undermine the credibility of persons, groups or ideas, or deny their trustworthiness	360	11.0	0.83	26	2.4	-*
Accusation of lying Insinuation that ideas, plans, actions or policies are dishonest, subterfuge and misleading	136	4.1	0.84	75	6.9	0.76
Subtask 2: Engaging Comments	865	26.6		293	26.8	
Argument Statements to substantiate or refute theses	506	15.5	0.72	197	18.0	0.80
Additional information Additional information are cited as references for personal opinions	184	5.6	0.84	37	3.4	0.85
Personal experience Personal experiences or values are cited as references for personal opinions	125	3.8	0.86	25	2.3	0.69
Solution proposal Constructive solution proposals are democratic, realistic and rational in the broadest sense	89	2.7	0.88	58	5.3	0.77
Empathy Serious attempt to understand and acknowledge a perspective or emotion	31	0.9	0.86	10	0.9	0.79
Mutual Respect Giving credit or praising personality traits or accomplishments	59	1.7	0.86	24	2.2	0.85
Polite salutation Use of polite language indicated by e.g. polite salutation	30	0.9	1	11	1.0	0.90
Subtask 3: Fact-Claiming Comments	1103	34.0		353	32.3	
Assertion of facts Statements with a truth claim, which is accessible for proof	1013	31.2	0.73	343	31.4	0.82
Provision of evidence Additional information are cited as references for personal opinions	184	5.6	0.84	37	3.4	0.85
	N = 3244		n = 105 4 annotators	N = 1092		n = 123 6 annotators

Table 1: Overview of frequency distribution and reliability (Krippendorff's Alpha) of fine-grained class labels on training and test dataset. Annotation scheme was adapted from Wilms et al. (2021). Note that the test set used in the shared task is a subset of the test set listed in this table where we filtered out 148 of the samples. Thereby, we ensure a similar class distribution in the training and test set of the shared task. The size and class distribution of the downsampled test set are displayed in Table 2. *The category *Discrediting* was re-labeled in the test dataset by one person.

Please note, that a comment can be assigned to more than one main category at the same time. Figure 1 shows examples for all three classes.

Toxic Comments. Toxic comments comprise uncivil forms of communication that can violate the rules of polite behavior, such as insulting participants of a discussion, using vulgar or sarcastic language or implied volume via capital letters. Additionally, incivility can be characterized as a violation of democratic discourse values, e.g. by verbally attacking basic democratic principles or making it difficult for others to participate (Papacharissi, 2004). It includes discrimination or discreditation of participants as well as threats of violence or the accusation of lying.

Engaging Comments. Engaging comments include behavior that is in line with deliberative principles, namely rationality, reciprocity, and mutual respect (Gutmann and Thompson, 1998). The first category covers communication features, such as justification, solution proposals, or the sharing of personal experiences. The second category covers empathy with regard to other users' standpoints. The third category is present when the comment is in line with rules of polite interaction or includes the expression of mutual respect.

Fact-Claiming Comments. All comments that contain any assertion of facts are considered as fact-claiming comments. In addition, the provision of evidence by external sources that have been cited fall into the class of fact-claiming comment. Figure 1 shows example comments of each class.

Sampling for the Final Dataset For the shared task, we resampled the original test dataset as presented in Table 1 so that for all subtasks, there is a similar class distribution between the training and test dataset. This was achieved by downsampling the test set. We decided in favor of this modification to allow supervised machine-learning approaches to be effective. Table 2 shows the size and class distribution of the training and test dataset as used in this year's edition of GermEval and as publicly available via the shared task website.

5 Evaluation

Following in the footsteps of the GermEval 2019 Shared on Hierarchical Classification of Blurbs (Remus et al., 2019) and the GermEval 2020 Shared Task on the Classification and Regression of

"Na, welchem tech riesen hat er seine Eier verkauft..?" *TOXIC*

"Ich macht mich wütend, dass niemand den Schülerinnen Gehör schenkt" *NOT TOXIC*

(a) Subtask 1: identification of toxic comments.

"Wie wär's mit einer Kostenteilung. Schließlich haben beide Parteien (Verkäufer und Käufer) etwas von der Tätigkeit des Maklers. Gilt gleichermassen für Vermietungen. Die Kosten werden so oder so weiterverrechnet, eine Kostenreduktion ist somit nicht zu erwarten." *ENGAGING*

"Die aktuelle Situation zeigt vor allem eines: viele Kinder mussten erkennen, dass ihre Mütter bestenfalls das Niveau Grundschule, Klasse 3 haben." *NOT ENGAGING*

(b) Subtask 2: identification of engaging comments.

"Kinder werden nicht nur seltener krank, sie infizieren sich wohl auch seltener mit dem Coronavirus als ihre Eltern - das ist laut Ministerpräsident Winfried Kretschmann (Grüne) das Zwischenergebnis einer Untersuchung der Unikliniken Heidelberg, Freiburg und Tübingen." *FACT-CLAIMING*

"hmm...das kann ich jetzt nich nachvollziehen..." *NOT FACT-CLAIMING*

(c) Subtask 3: identification of fact-claiming comments.

Figure 1: Example comments and their class labels.

Cognitive and Motivational Style (Johannßen et al., 2020), we use the platform codalab for evaluation.[2]

The evaluation uses precision, recall, and macro-average F1-score as metrics. Macro-average F1-scores give equal importance to each class, which is suited because classes in our dataset are not uniformly distributed but are equally important to identify. It is calculated as the harmonic mean of the arithmetic means of class-wise precision and recall:

$$F_1 = 2\frac{\bar{P}\bar{R}}{\bar{P}+\bar{R}} = 2\frac{(\frac{1}{n}\sum_i P_i)(\frac{1}{n}\sum_i R_i)}{\frac{1}{n}\sum_i P_i + \frac{1}{n}\sum_i R_i}$$

with P_i and R_i referring to precision and recall of class i out of n classes. We rank systems by

[2]The competition page is https://competitions.codalab.org/competitions/32854.

Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments
co-located with KONVENS

Subtask	Class Label	Training Data Freq	Training Data %	Test Data Freq	Test Data %
(1) toxic comments	toxic	1122	34.6	350	37.1
	not toxic	2122	65.4	594	62.9
(2) engaging comments	engaging	865	26.7	253	26.8
	not engaging	2379	73.3	691	73.2
(3) fact-claiming comments	fact-claiming	1103	34.0	314	33.3
	not fact-claiming	2141	66.0	630	66.7
total		3244	100.0	944	100.0

Table 2: Class distribution of the training and test dataset as used in the shared task.

their macro-average F1-score and do not consider accuracy in this shared task, since there is an imbalanced class distribution in each subtask. Accuracy typically rewards correct classification of the majority class. An evaluation tool computing all of the above mentioned evaluation measures is available on the website of the shared task.

6 Results

A high-level summary of the results by the participants in the different subtasks is given in Table 3. It provides summary statistics on the macro-average F1-score, which is the metric that was used as the official ranking criterion in the shared task. In comparison to subtask 1, the results of subtasks 2 and 3 are more tightly clustered suggesting that the methods pursued by the different participants are similarly effective. Overall, the best F1-scores reached in the different subtasks range from 69.98 (subtask 2) to 76.26 (subtask 3). These absolute numbers suggest that all three tasks are difficult and that there is still room for improvement.

Toxic Comments. We received 31 different runs from twelve teams for subtask 1, i.e. the detection of toxicity. The results are shown in Table 4. As a baseline, we also included the performance of a majority-class classifier always predicting the majority class, which is the absence of toxicity.

Engaging Comments. We received 25 different runs from nine teams for subtask 2, i.e. the detection of engaging comments. The results are shown in Table 5. As a baseline, we also included the performance of a majority-class classifier always predicting the majority class, which is the absence of engaging comments.

Fact-Claiming Comments. We received 31 different runs from eleven teams for subtask 3, i.e. the detection of fact-claiming comments. The results are shown in Table 6. As a baseline, we also included the performance of a majority-class classifier always predicting the majority class, which is the absence of fact-claiming comments.

General Conclusions Drawn from the Evaluation. Given that the overwhelming majority of participants followed generic classification approaches for the different subtasks, we discuss the results in this section jointly. All teams that participated in this year's shared task tested some form of deep learning. All teams except one considered contextual embeddings, most predominantly some type of transformer (i.e. BERT (Devlin et al., 2019)). Since the participants made use of various publicly available pre-trained models and given that the models of the best performing systems are different, it is difficult to determine any publicly available model that is particularly effective. Other types of classifiers, be it traditional supervised classifiers (e.g. Support Vector Machines, Logistic Regression, Forests) or other deep learning algorithms (e.g. CNN, GRU, or LSTM) were only used by a handful of teams each. Only one participant also tested a rule-based classifier.

An additional method that has already proved effective in previous editions of GermEval (Wiegand et al., 2018; Struß et al., 2019) are ensemble methods. Slightly more than half of the participants employed some form of ensemble, including virtually all top-performing systems. However, we do not see a clear pattern what type of classifiers should be combined into an ensemble, be it simply different initializations of the same classifier (i.e.

Subtask	# Teams	# Runs	Min	Max	Median	Mean	SD
(1) toxic comments	12	31	35.97	71.75	66.85	63.63	8.49
(2) engaging comments	9	25	61.43	69.98	68.72	67.70	2.14
(3) fact-claiming comments	11	31	59.70	76.26	72.55	71.84	3.94

Table 3: Summary statistics for overall macro F1-scores in the three subtasks.

Team ID	Codalab Run ID	F1	P	R
FHAC	921610	71.75	73.10	70.44
FHAC	921609	71.61	70.87	72.37
FHAC	920735	71.27	70.55	72.00
FH-SWF SG	918686	70.73	74.28	67.51
WLV-RIT	921323	69.14	73.54	65.24
WLV-RIT	921321	69.14	72.56	66.03
ur-iw-hnt	921615	68.98	71.83	66.35
DFKI SLT	921619	68.59	68.99	68.18
TUW-Inf	921590	68.42	70.44	66.52
ur-iw-hnt	921616	68.33	71.68	65.29
ur-iw-hnt	921614	68.10	70.47	65.88
WLV-RIT	921318	67.96	71.74	64.56
TUW-Inf	921582	67.71	70.06	65.51
TUW-Inf	921594	67.46	69.22	65.79
Precog-LTRC-IIITH	920506	66.87	67.42	66.33
DFKI SLT	920147	66.85	66.35	67.35
Data Science Kitchen	921663	66.85	66.98	66.73
Precog-LTRC-IIITH	920089	66.54	67.17	65.92
FH-SWF SG	921306	65.81	67.77	63.95
DFKI SLT	921621	65.73	65.90	65.56
Data Science Kitchen	921319	64.79	65.95	63.67
Data Science Kitchen	921587	63.78	64.89	62.71
Universität Regensburg MaxS	921252	61.53	62.30	60.79
DeTox	921281	58.95	63.06	55.35
IRCologne	921157	57.63	58.24	57.03
IRCologne	921667	57.40	58.03	56.77
UR@NLP_A_Team	921640	55.59	55.71	55.47
UR@NLP_A_Team	919179	55.47	55.29	55.65
UR@NLP_A_Team	921263	55.45	55.50	55.40
DeTox	921278	38.12	38.54	37.71
DeTox	921282	35.97	36.22	35.72
majority-class classifier (baseline)		38.62	31.46	50.00

Table 4: Results of subtask 1: identification of toxic comments.

transformer), different pre-trained models or the combination of a transformer with a traditional supervised classifier. While the participants applied different methods to combine all predictions of the ensembled models into a single prediction, the most frequent method was simple (soft) majority voting.

Only three teams considered using the data from previous related GermEval editions (Wiegand et al., 2018; Struß et al., 2019) as additional training data. This low number does not come as a surprise since those previous editions addressed text from a different source, i.e. Twitter rather than Facebook. Being

Team ID	Codalab Run ID	F1	P	R
Data Science Kitchen	921663	69.98	71.71	68.34
FHAC	921609	69.91	68.39	71.51
FW-SWF SG	918686	69.69	69.41	69.97
WLV-RIT	921321	69.47	68.95	69.99
WLV-RIT	921323	69.34	69.44	69.24
ur-iw-hnt	921614	69.29	72.28	66.53
WLV-RIT	921318	69.26	68.27	70.27
FW-SWF SG	921306	69.02	68.42	69.63
FHAC	920735	69.01	67.52	70.56
Precog-LTRC-IIITH	920506	68.93	68.37	69.50
UPAppliedCL	921269	68.92	70.77	67.16
ur-iw-hnt	921615	68.75	71.24	66.42
Data Science Kitchen	921319	68.72	69.70	67.78
Precog-LTRC-IIITH	920089	68.60	68.21	69.00
Data Science Kitchen	921587	68.33	69.26	67.43
ur-iw-hnt	921616	67.64	70.03	65.42
UPAppliedCL	921271	66.91	68.49	65.39
UPAppliedCL	921270	66.88	70.07	63.97
TUW-Inf	921590	66.34	78.02	57.70
TUW-Inf	921582	66.34	78.02	57.70
TUW-Inf	921594	66.34	78.02	57.70
FHAC	921610	65.80	66.68	64.95
UR@NLP_A_Team	921263	64.28	64.06	64.50
UR@NLP_A_Team	919179	63.37	62.11	64.68
UR@NLP_A_Team	921640	61.43	61.07	61.80
majority-class classifier (baseline)		42.26	36.60	50.00

Table 5: Results of subtask 2: identification of engaging comments.

out-of-domain data, the data from those previous GermEval shared tasks are unlikely to produce a notable improvement for this year's shared task.

Only two teams considered exploiting the plethora of available English training datasets for this task by following some multilingual approach. This low number, too, is in line with recent findings. Even for subtask 1, i.e. toxicity detection, for which many English datasets exist (Vidgen and Derczynski, 2020; Risch et al., 2021), Nozza (2021) recently identified reasons why multilingual approaches are highly problematic. One team also explored harnessing synthetically generated training data. However, that approach did not produce the expected outcome. Despite the similarity of many approaches pursued by the different participants of this year's edition of GermEval, the difference in performance for subtask 1 is still fairly large (Table 3). We assume that due to the complexity of those state-of-the-art learning methods and frameworks, there is still a very high number of degrees of freedom (e.g. settings of hyperparameters) that apparently plays a significant role in the overall performance of classifiers. As a basis for our analysis of the results, we asked all participants to complete a survey in which we asked about details of their submission. A summary of the survey responses is available on the shared task website.

7 Conclusion

In this paper, we described the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. For each of the three classes of comments, there was an individual subtask that defined a binary classification problems. As part of this shared task, we introduced a hand-annotated dataset of 4,188 Facebook-posts. The results for all three subtasks show that state-of-the-art classification approaches perform well and achieve macro-average F1-scores between 70%

Team ID	Codalab Run ID	F1	P	R
FHAC	921609	76.26	74.97	77.59
ur-iw-hnt	921615	76.02	77.56	74.54
ur-iw-hnt	921616	75.79	77.25	74.38
ur-iw-hnt	921614	75.43	77.91	73.10
FHAC	920735	74.82	73.52	76.16
WLV-RIT	921318	74.72	74.50	74.95
WLV-RIT	921321	74.68	75.30	74.07
AITFHSTP	921165	74.62	74.13	75.11
Precog-LTRC-IIITH	920506	73.91	73.44	74.39
WLV-RIT	921323	73.69	73.54	73.83
Precog-LTRC-IIITH	920089	73.69	73.14	74.24
UPAppliedCL	921269	73.60	74.01	73.19
FH-SWF SG	921306	73.57	73.63	73.51
FH-SWF SG	918686	73.37	72.76	74.00
AITFHSTP	921162	72.84	72.71	72.96
Data Science Kitchen	921663	72.55	73.03	72.08
Data Science Kitchen	921587	72.44	73.39	71.52
Data Science Kitchen	921319	72.34	73.25	71.44
FHAC	921610	72.28	73.75	70.88
UPAppliedCL	921270	72.21	75.78	68.96
TUW-Inf	921590	72.07	71.18	72.97
TUW-Inf	921582	72.07	71.18	72.97
UPAppliedCL	921271	71.69	73.63	69.84
HunterSpeechLab	921571	71.50	72.72	70.32
HunterSpeechLab	921569	69.91	70.97	68.89
AITFHSTP	921168	69.27	68.45	70.11
TUW-Inf	921594	68.80	82.35	59.08
HunterSpeechLab	921565	68.51	69.24	67.78
UR@NLP_A_Team	919179	63.16	62.41	63.92
UR@NLP_A_Team	921640	61.50	61.10	61.91
UR@NLP_A_Team	921263	59.70	59.15	60.26
majority-class classifier (baseline)		40.03	33.37	50.00

Table 6: Results of subtask 3: identification of fact-claiming comments.

and 76%. However, all of them should be considered far from solved. In terms of methods, we cannot determine a clear winner. All participants employed some form of transformer-based neural network. Due to the complexity of that method, there is a large number of degrees of freedom, such as hyperparameters, which need to be carefully set. They still seem to have a significant impact upon the resulting overall classification performance.

Acknowledgments

We are grateful to the large number of participants whose enthusiastic participation made GermEval 2021 a great success. We would like to thank Marc Ziegele and Dominique Heinbach for the provision of annotated data as well as their valuable insight and support during data annotation. We would also like to thank our student assistants Sebastian Joppien, Saskia Jende, Alena Palkowski, Charlotte Pape and Noah Schmitt, who performed parts of the data annotation. Parts of this research were conducted in the project "AI-supported collective-social moderation of online discussions" (KOSMO) supported by the Federal Ministry of Education and Research Grant 01IS19040C to Marc Ziegele.

References

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A benchmark dataset of check-worthy factual claims. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 821–829. AAAI Press.

Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. 2020. Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 215–236. Springer.

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 International Workshop on Semantic Evaluation (SemEval@NAACL)*, pages 54–63. ACL.

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 North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186. ACL.

Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeno, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. 2019. Overview of the clef-2019 checkthat! lab: Automatic identification and verification of claims. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 301–321. Springer.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):85:1–85:30.

Dennis Friess and Christiane Eilders. 2015. A systematic review of online deliberation research. *Policy & Internet*, 7(3):319–339.

Amy Gutmann and Dennis F Thompson. 1998. *Democracy and disagreement*. Harvard University Press.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1803–1812. ACM.

Mark Hsueh, Kumar Yogeeswaran, and Sanna Malinen. 2015. "Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors? *Human Communication Research*, 41(4):557–576.

Dirk Johannßen, Chris Biemann, Steffen Remus, Timo Baumann, and David Scheffer. 2020. Germeval 2020 task 1 on the classification and regression of cognitive and motivational style from text: Companion paper. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 1–10. CEUR-WS.org.

Varada Kolhatkar and Maite Taboada. 2017. Using new york times picks to identify constructive comments. In *Proceedings of the Natural Language Processing meets Journalism Workshop (NLPmJ@EMNLP)*, pages 100–105. ACL.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 372–387. Springer.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021. The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, pages 639–649. Springer.

Courtney Napoles, Aasish Pappu, and Joel R Tetreault. 2017a. Automatically identifying good conversations online (yes, they do exist!). In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 628–631. AAAI Press.

Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. 2017b. Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of the Linguistic Annotation Workshop (LAW@EACL)*, pages 13–23.

Chikashi Nobata, Joel R. 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. ACM.

Debora Nozza. 2021. Exposing the limits of Zero-shot Cross-lingual Hate Speech Detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL.

Zizi Papacharissi. 2004. Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New media & society*, 6(2):259–283.

Ayush Patwari, Dan Goldwasser, and Saurabh Bagchi. 2017. Tathya: A multi-classifier system for detecting check-worthy statements in political debates. In *Proceedings of the Conference on Information and Knowledge Management (CIKM)*, pages 2259–2262. ACM.

John Pavlopoulos, Jeffrey Sorensen, Leo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 Task 5: Toxic Spans Detection. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval@ACL-IJCNLP)*, pages 59–69. ACL.

Fabian Prochazka, Patrick Weber, and Wolfgang Schweiger. 2018. Effects of civility and reasoning in user comments on perceived journalistic quality. *Journalism studies*, 19(1):62–78.

Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Proceedings of the Canadian Conference on Advances in Artificial Intelligence (Canadian AI)*, pages 16–27. Springer.

Steffen Remus, Rami Aly, and Chris Biemann. 2019. Germeval 2019 task 1: Hierarchical classification of blurbs. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 280–292. German Society for Computational Linguistics and Language Technology (GSCL).

Julian Risch and Ralf Krestel. 2020. Top comment or flop comment? predicting and explaining user engagement in online news discussions. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 579–589. AAAI Press.

Julian Risch, Philipp Schmidt, and Ralf Krestel. 2021. Data integration for toxic comment classification: Making more than 40 datasets easily accessible in one unified format. In *Proceedings of the Workshop on Online Abuse and Harms (WOAH@ACL)*, pages 157–163. ACL.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the International Workshop on Natural Language Processing for Social Media (SocialNLP@EACL)*, pages 1–10. ACL.

Anke Stoll, Marc Ziegele, and Oliver Quiring. 2020. Detecting impoliteness and incivility in online discussions: Classification approaches for german user comments. *Computational Communication Research*, 2(1):109–134.

Natalie Jomini Stroud, Joshua M Scacco, Ashley Muddiman, and Alexander L Curry. 2015. Changing deliberative norms on news organizations' facebook sites. *Journal of Computer-Mediated Communication*, 20(2):188–203.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 352–363. German Society for Computational Linguistics and Language Technology (GSCL).

Bertie Vidgen and Leon Derczynski. 2020. Directions in Abusive Language Training Data. *PLoS One*, 15(12).

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Workshop on Language in Social Media (LSM@ACL)*, pages 19–26. ACL.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 602–608. ACL.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 1–10. Austrian Academy of Sciences.

Lena Wilms, Dominique Heinbach, and Marc Ziegele. 2021. Annotation guidelines for GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. Excerpt of an unpublished codebook of the DEDIS research group at Heinrich-Heine-University Düsseldorf (full version available on request).

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 *Proceedings of the International Workshop on Semantic Evaluation (SemEval@NAACL)*, pages 75–86. ACL.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of the International Workshop on Semantic Evaluation (SemEval@COLING)*, pages 1425–1447. ACL.

Marc Ziegele, Mathias Weber, Oliver Quiring, and Timo Breiner. 2018. The dynamics of online news discussions: effects of news articles and reader comments on users' involvement, willingness to participate, and the civility of their contributions. *Information, Communication & Society*, 21(10):1419–1435.

8 Appendix

Team ID	Affiliation	Paper Title
AITFHSTP	Austrian Institute of Technology GmbH/St. Pölten University of Applied Sciences	AITFHSTP at GermEval 2021: Automatic Fact Claiming Detection with Multilingual Transformer Models
Data Science Kitchen	Data Science Kitchen	Data Science Kitchen at GermEval 2021: A Fine Selection of Hand-Picked Features, Delivered Fresh from the Oven
DeTox	Darmstadt University of Applied Sciences/Fraunhofer Institute for Secure Information Technology	DeTox at GermEval 2021: Toxic Comment Classification
FHAC	FH Aachen University of Applied Sciences	FHAC at GermEval 2021: Identifying German toxic, engaging, and fact-claiming comments with ensemble learning
FH-SWF SG	Fachhochschule Südwestfalen	FH-SWF SG at GermEval 2021: Using Transformer-Based Language Models to Identify Toxic, Engaging, & Fact-Claiming Comments
HunterSpeechLab	City University of New York	HunterSpeechLab at GermEval 2021: Does Your Comment Claim A Fact? Contextualized Embeddings for German Fact-Claiming Comment Classification
IRCologne	TH Köln	IRCologne at GermEval 2021: Toxicity Classification
Precog-LRTC-IIITH	International Institute of Information Technology, Hyderabad, India	Precog-LTRC-IIITH at GermEval 2021: Ensembling Pre-Trained Language Models with Feature Engineering
DFKI SLT	DFKI GmbH	DFKI SLT at GermEval 2021: Multilingual Pretraining and Data Augmentation for the Classification of Toxicity in Social Media Comments
Universität Regensburg MaxS	Universität Regensburg	Universität Regensburg MaxS at GermEval 2021 Task 1: Toxic Comment Classification
UPAppliedCL	University of Potsdam	UPAppliedCL at GermEval 2021: Identifying Fact-Claiming and Engaging Facebook Comments Using Transformers
ur-iw-hnt	University of Regensburg	ur-iw-hnt at GermEval 2021: An Ensembling Strategy with Multiple BERT Models
UR@NLP_A_Team	University of Regensburg	UR@NLP_A_Team @ GermEval 2021: Ensemble-based Classification of Toxic, Engaging and Fact-Claiming Comments
TUW-Inf	TU Wien	TUW-Inf at GermEval2021: Rule-based and Hybrid Methods for Detecting Toxic, Engaging, and Fact-Claiming Comments
WLV-RIT	University of Wolverhampton/Rochester Institute of Technology	WLV-RIT at GermEval: Multitask Learning with Transformers to Detect Toxic, Engaging, and Fact-Claiming Comments

Table 7: Team ID, affiliation and paper title.