UPAppliedCL at GermEval 2021: Identifying Fact-Claiming and Engaging Facebook Comments Using Transformers

Robin Schaefer
Applied Computational Linguistics
University of Potsdam
Potsdam, Germany
robin.schaefer@uni-potsdam.de

Manfred Stede
Applied Computational Linguistics
University of Potsdam
Potsdam, Germany
stede@uni-potsdam.de

Abstract

In this paper we present UPAppliedCL's contribution to the GermEval 2021 Shared Task. In particular, we participated in Subtasks 2 (Engaging Comment Classification) and 3 (Fact-Claiming Comment Classification). While acceptable results can be obtained by using unigrams or linguistic features in combination with traditional machine learning models, we show that for both tasks transformer models trained on fine-tuned BERT embeddings yield best results.

1 Introduction

In the last decade social media platforms, like Facebook[1], have gained a notable momentum, which is reflected by the increasing number of users of social media.[2] While facilitating communication across the globe, from the perspective of NLP however, systems need to be specifically adapted to social media for the following reasons.

First, social media data is unedited and contains certain conventions which can pose challenges for systems trained on more well-formed texts (Šnajder, 2016). Second, social media platforms are used for different kinds of communication ranging from everyday conversations to sophisticated evidence-based argumentation on political issues. While the latter have the potential to contribute to public political discourse in general, social media has been found to contain not only respectful and engaging discussions but also hateful speech, which threatens the respectful exchange and possibly also the mental well-being of its participants. The GermEval 2021 Shared Task (Risch et al., 2021) aims to stimulate research on this issue, while also going beyond the single task of toxic comment classification.

In this paper we present UPAppliedCL's contribution[3] to the GermEval 2021 Shared Task which consists of three subtasks revolving around the mentioned characteristics of social media discussions: 1. Toxic Comment Classification; 2. Engaging Comment Classification; 3. Fact-Claiming Comment Classification. Here we especially focus on Subtask 3 (fact-claiming comments), which is also relevant for tasks in the field of argument mining (AM) (Dusmanu et al., 2017; Schaefer and Stede, 2021). In addition, we also participate in Subtask 2 (engaging comments), which we consider as a first albeit facultative step in an AM system in order to identify potential argumentative comments. As we consider Subtask 1 (toxic comments) as a task which is more independent of AM, we will not attend to it in this work.

The paper is structured as follows: in Section 2 we give a short overview of relevant previous work. We present the dataset provided by the organizers in Section 3. In Section 4 we describe our approach including the developed baselines, and in Section 5 we continue with the obtained results, which are discussed in Section 6. We conclude the paper in Section 7.

2 Related Work

Given that we do not participate in Subtask 1 (toxic comments) we will not go further into details here. For surveys on tackling this issue using NLP techniques we refer the reader to Schmidt and Wiegand (2017) and Mishra et al. (2019).

Subtask 2 (engaging comments) may be seen as a complement task to toxic comment classification as it focuses more on the identification of respectful

[1] https://www.facebook.com/

[2] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[3] Code Repository: https://github.com/RobinSchaefer/GermEval2021

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

conversation. Approaches include work by Risch and Krestel (2020) who propose a system that relies on upvotes and replies in order to identify news comments that potentially attract user engagement. A neural network model obtained classification accuracies ranging from 0.68 to 0.72.

Subtask 3 (fact-claiming comments) can be approached from the perspective of AM, i.e., identifying fact-claiming content can be an important first step for further proving its actual correctness. Related work was published by Dusmanu et al. (2017) who investigated the classification of factual and opinionated tweets, which is defined as a pre-task for later checks of correctness, e.g., via source identification. A logistic regression model trained on a set of lexical, twitter-specific, syntactic/semantic and sentiment features yielded an F1 score of 0.80. Note, however, that no information is given whether micro or macro F1 scores are reported.

Factual information can also be used as evidence for claims. In that sense, fact-claiming comment classification can be interpreted as a pre-task for evidence detection, which has previously been investigated for different text sources including social media. For instance, in our previous work, we investigated different AM tasks including evidence detection on an expert and crowd annotated German tweet dataset. To this end we used classification and sequence labeling techniques. For evidence detection on the expert annotated dataset we obtained macro F1 scores of 0.60-0.75 for classification (XGBoost) and 0.61-0.72 for sequence labeling (CRF) (Iskender et al., 2021).

3 Data

The provided training set consists of 3244 German comments, which were collected from the Facebook page of a German political talk show. The comments were posted from February to July 2019 on two shows. All comments were anonymized. This includes replacement of user links with @USER, show links with @MEDIUM and moderator links with @MODERATOR. The comments were annotated by four trained expert annotators. For measuring inter annotator agreement (IAA) the Krippendorff's α metric was used. In total three binary annotation layers were created, each for one of the three subtasks of GermEval 2021.

Toxic Comments: Toxic comments include different types of uncivil behavior like insults, sarcastic language, discrimination, and threats of violence. It also comprises attacks on democratic principles (IAA: $0.73 < \alpha < 0.90$).

Engaging Comments: Engaging comments comprise language centering around rationality, mutual respect, empathy for others and their standpoints, and mediation (IAA: $0.71 < \alpha < 1.0$).

Fact-Claiming Comments: Fact-claiming comments focus on the assertion of facts, or evidence provided by external sources (IAA: $0.73 < \alpha < 0.84$).

For the development of our system we conducted a stratified split on the provided training set in order to obtain training, development and test sets. Both development and test set consisted of about 12.5% of the former training set. We used the development set to experiment with different feature sets and hyperparameters, while the test set was only used to calculate the preliminary test results presented in this paper.

For final system evaluation, 944 additional unlabeled comments were provided. These were drawn from discussions on a different show to avoid a topical bias.

4 System Description

In this paper we follow a machine learning (ML) approach based both on traditional ML methods and more recent deep learning (DL) techniques. We define three baselines against which we compare our submitted systems. All systems are evaluated using macro F1, precision and recall scores.

4.1 Baselines

As the first baseline (**majority**) we consider a simplistic model that outputs the most frequent class for all comments. Proportions of the most frequent class are 0.73 for Subtask 2 (non engaging) and 0.66 for Subtask 3 (non fact-claiming),[4] which indicates some imbalance in both datasets.

We define two more baselines which we had first considered for submission. However, given

[4]Importantly, these values equal the micro F1 score obtained by the first baseline model. Given that the subtasks are evaluated using macro scores, we calculate these for the baselines as well. This leads to results that diverge from the proportions but are directly comparable to the system run evaluations.

Linguistic Feature	Definition
Citation Ratio	ratio of citations
Comma Ratio	ratio of commas
First Person Ratio	ratio of 1st person pronouns
Initial Capital Ratio	ratio of tokens starting with capital
Medium Ratio	ratio of medium links
Modal Ratio	ratio of modal verbs
Moderator Ratio	ratio of moderator links
Question Ratio	ratio of question marks
Sentiment	the comment's sentiment
Text Length	the comment length
Token Length	the average token length
User Ratio	ratio of user links

Table 1: Definitions of Linguistic Features

that they cannot compete against the more sophisticated DL approaches we decided on using them for mere comparison. For baseline 2 (**unigram**) we derive unigrams from the data. We experimented with different variations of n-grams but simple unigrams perform best. During preprocessing we set all tokens to lower case and removed stopwords. Final vocabulary size is 19085. Baseline 3 (**linguistic features**) is based on a set of linguistic and text-related features which was compiled manually (see Table 1). Features for baseline 3 are partly inspired by Krüger et al. (2017). However, features *medium ratio*, *moderator ratio* and *user ratio* are based on the anonymization of the comments conducted by the organizers.

In addition to different feature sets we experimented with different classification algorithms: AdaBoost, Decision Trees (DT), eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016), Gaussian Naive Bayes, Logistic Regression (LR), Support Vector Machines (SVM) and Random Forest (RF). Except for XGBoost[5] all algorithms are implemented using Scikit-Learn (Pedregosa et al., 2011). We only present results of the best systems.

4.2 Submitted Approaches

Our submitted approaches are more heavily based on DL techniques (see Table 2). All three approaches make use of pretrained German BERT document embeddings[6], which were published by *deepset.ai*. Note that the embeddings were pretrained on a set of Wikipedia texts, legal texts and news articles and not on social media data.

For submissions I and II we trained transformer models (Vaswani et al., 2017) using Flair (Akbik et al., 2019), an NLP framework which provides simple interfaces for different tasks including the creation of text embeddings and training of classification models. In addition, BERT embeddings used for submission I are fine-tuned during training, whereas for submission II the pretrained BERT embeddings are directly used for feature extraction.

Recall that the final evaluation set diverges from the training set with respect to the discussed show, i.e., the topic. To account for the possibility that during fine-tuning the BERT embeddings overfit to the training data, we decided against fine-tuning in submissions II and III.

For submission III we employed the same pretrained BERT embeddings. Instead of training a transformer model, however, we trained the same set of ML models on the encoded comments that we used for baselines 2 and 3. Our experiments revealed that XGBoost models perform best for this feature type, which is why, in the following, we will exclusively focus on this classifier. This approach is comparable to other previous work of ours, which focused on argument detection in tweets (Iskender et al., 2021; Schaefer and Stede, 2020).

We hypothesize the following ranking of submitted approaches for both subtasks:

1. Fine-tuned BERT Embeddings + Transformer
2. BERT Embeddings + Transformer
3. BERT Embeddings + XGBoost

Despite the possibility of overfitting we assume that the classifier will actually benefit from fine-tuning as the embeddings were not originally pretrained on social media data. We further hypothesize that transformers will obtain better results than traditional ML models given their success in recent years.

5 Results

Our results are based on two different datasets: 1. The test set that we obtained from our own splitting

[5] https://xgboost.readthedocs.io/en/latest/index.html

[6] https://huggingface.co/bert-base-german-cased

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

Submission	Features	Classifier
I	BERT Emb (FT)	Transformer
II	BERT Emb	Transformer
III	BERT Emb	XGBoost

Table 2: Submitted Approaches (Emb=Embeddings; FT=fine-tuned)

of the provided training set (henceforth **Test Set**); 2. The evaluation set we were provided with for creation of the submitted runs that were evaluated by the organizers (henceforth **Evaluation Set**). We present results obtained by both baseline and submitted models. Recall that we only participated in subtasks 2 and 3 and that all presented results are macro scores.

5.1 Test Set Results

Table 3 shows results obtained from baseline and submitted models that were applied to the test set. Due to the macro analysis the simple majority model only obtains weak results. Both the unigram baseline and the linguistic feature baseline yield substantially higher scores. Importantly, the unigram baseline performs better for both tasks than the linguistic feature baseline (Subtask 2: F1 0.728 vs 0.694; Subtask 3: F1 0.705 vs 0.704), although the better score for Subtask 3 is likely due to chance. Interestingly, precision is higher than recall.

F1 scores reveal that transformer models trained on fine-tuned BERT embeddings yield best results for both subtasks (Subtask 2: 0.775; Subtask 3: 0.790). It is noteworthy, however, that highest precision scores are obtained by the transformer models that were trained without embedding fine-tuning (Subtask 2: 0.845; Subtask 3: 0.817), while fine-tuning led to higher recall. Interestingly, an XGBoost model performs more successfully on Subtask 2 than a transformer if both are trained without fine-tuning (0.751 vs 0.737). For Subtask 3, however, the outcome was vice versa (0.754 vs 0.761). In general, scores for Subtask 3 tend to be higher than scores for Subtask 2 with the exception of precision.

5.2 Evaluation Set Results

Results obtained from finally evaluating the submitted runs are shown in Table 4. For comparison we also evaluated the unigram and linguistic feature baselines. This was possible as the organizers provided us with the labels of the evaluation set, once the deadline for the submission runs had passed. We ignore the majority baseline, as class distributions in the evaluation set are comparable to the training set.

Both baseline models show reduced F1 scores on both subtasks compared to the model outcomes from the test set. Notably, the reduction for the unigram model is larger than for the linguistic feature model. The unigram model further shows a higher recall, while the linguistic feature model benefits from a higher precision.

The first submitted system, i.e., fine-tuned BERT embeddings with transformer, yield best results (Subtask 2: 0.689; Subtask 3: 0.736), although F1 scores are again somewhat reduced compared to the test set results. Scores are higher for Subtask 3 than for Subtask 2 including precision, which contrasts with results obtained from the test set.

This pattern repeats for Submissions II (BERT embeddings (not fine-tuned) with Transformer) and III (BERT embeddings (not fine-tuned) with XGBoost classifier). Notably the XGBoost approach yields equal results in Subtask II as the transformer approach (F1: 0.669).

6 Discussion

In this section we discuss some of the results obtained by the submitted models.

As shown in Section 5 transformers trained on fine-tuned BERT embeddings yield best F1 scores, which indicates that fine-tuning does not lead to overfitting. This is the case for testing with the in-domain testing set, evaluating with the final evaluation set and for both subtasks. Further, this is in line with our ranking hypothesis.

Interestingly, however, an XGBoost model performs better on the test set of Subtask 2 than a transformer if both are trained on non-fine-tuned BERT embeddings, which contradicts our ranking hypothesis. In contrast, a transformer is more successful than an XGBoost model on Subtask 3. Model differences on the evaluation set, however, are less substantial. Evaluation F1 scores on Subtask 2 are equal. It is difficult to argue why these patterns arise. However, from these results we can carefully conclude that DL models like transformers do not necessarily outperform traditional ML models.

Furthermore, precision appears to be reduced if embeddings are fine-tuned while recall benefits

Approach	Subtask (ST) 2			Subtask (ST) 3		
	F1	Precision	Recall	F1	Precision	Recall
Majority	0.423	0.367	0.500	0.398	0.330	0.500
Unigram SVM (ST 2)/LR (ST 3)	**0.728**	**0.817**	**0.700**	**0.705**	**0.778**	0.691
Linguistic Features XGBoost (ST 2)/RF (ST 3)	0.694	0.729	0.678	0.704	0.728	**0.694**
BERT Emb (FT) Transformer	**0.775**	0.817	**0.752**	**0.790**	0.807	**0.780**
BERT Emb Transformer	0.737	**0.845**	0.706	0.761	**0.817**	0.742
BERT Emb XGBoost	0.751	0.818	0.724	0.754	0.796	0.738

Table 3: Test Set Results (Emb=Embeddings; FT=fine-tuned)

Submission	Approach	Subtask (ST) 2			Subtask (ST) 3		
		F1	Precision	Recall	F1	Precision	Recall
-	Unigram SVM (ST 2)/LR (ST 3)	**0.671**	0.665	**0.688**	0.654	0.667	**0.688**
-	Linguistic Features XGBoost (ST 2)/RF (ST 3)	0.670	**0.681**	0.664	**0.693**	**0.710**	0.685
I	BERT Emb (FT) Transformer	**0.689**	**0.708**	**0.672**	**0.736**	0.740	**0.732**
II	BERT Emb Transformer	0.669	0.701	0.640	0.722	**0.758**	0.690
III	BERT Emb XGBoost	0.669	0.685	0.654	0.717	0.736	0.698

Table 4: Evaluation Set Results (Emb=Embeddings; FT=fine-tuned)

from it. This may have interesting implications with respect to the application's focus. The results suggest that a model needing a high recall can benefit from embedding fine-tuning, while ML practitioners requiring a higher precision may refrain from fine-tuning. This finding, of course, requires more investigation before making generalisations, especially as it is less pronounced in the evaluation results.

Scores yielded for Subtask 3 tend to be higher than for Subtask 2. We argue that this might be related to the class distribution, which is more balanced in Subtask 3.

Scores obtained by evaluation are lower than by testing. This, however, is expected due to the different topics covered in training and evaluation data. Recall that the test data is topically closer related to the training set than the evaluation set. Given that we still achieved good results, especially for Subtask 3, we argue that our models are capable of solving both tasks to a promising degree.

7 Conclusion

In this paper we presented approaches to fact-claiming and engaging comment classification. We applied different combinations of features (unigrams, linguistic features, BERT embeddings) and classification algorithms including more traditional ML techniques like SVM, RF or XGBoost and more recent DL techniques like transformer models. Our experiments show that best results can be achieved by using fine-tuned BERT embeddings in combination with a transformer. We also found that fine-tuning leads to a higher recall while precision benefits from refraining from fine-tuning. As this pattern is less obvious in the evaluation set we do

not argue that this finding necessarily generalizes to other datasets. However, it may be fruitful to shed more light on this in future work.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. Association for Computing Machinery.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark. Association for Computational Linguistics.

Neslihan Iskender, Robin Schaefer, Tim Polzehl, and Sebastian Möller. 2021. Argument Mining in Tweets: Comparing Crowd and Expert Annotations for Automated Claim and Evidence Detection. In H. Horacek E. Métais, F. Meziane and E. Kapetanios, editors, *Natural Language Processing and Information Systems (NLDB)*, Lecture Notes in Computer Science. Springer, Cham.

Katarina Krüger, Anna Lukowiak, Jonathan Sonntag, and Manfred Stede. 2017. Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Enginnering*, 23(5):687–707.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling online abuse: A survey of automated abuse detection methods. *CoRR*, abs/1908.06024.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Julian Risch and Ralf Krestel. 2020. Top comment or flop comment? predicting and explaining user engagement in online news discussions. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):579–589.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments co-located with KONVENS*, pages 1–12.

Robin Schaefer and Manfred Stede. 2020. Annotation and detection of arguments in tweets. In *Proceedings of the 7th Workshop on Argument Mining*, pages 53–58, Online. Association for Computational Linguistics.

Robin Schaefer and Manfred Stede. 2021. Argument Mining on Twitter: A survey. *it - Information Technology*, 63(1):45–58.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Jan Šnajder. 2016. Social media argumentation mining: The quest for deliberateness in raucousness. ArXiv:1701.00168.