UR@NLP_A_Team @ GermEval 2021: Ensemble-based Classification of Toxic, Engaging and Fact-Claiming Comments

Kwabena Odame Akomeah
University of Regensburg
kwabena-odame.akomeah@ur.de

Udo Kruschwitz
University of Regensburg
udo.kruschwitz@ur.de

Bernd Ludwig
University of Regensburg
bernd.ludwig@ur.de

Abstract

In this paper, we report on our approach to addressing the *GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments* for the German language. We submitted three runs for each subtask based on ensembles of three models each using contextual embeddings from pre-trained language models using SVM and neural-network-based classifiers. We include language-specific as well as language-agnostic language models – both with and without fine-tuning. We observe that for the runs we submitted that the SVM models overfitted the training data and this affected the aggregation method (simple majority voting) of the ensembles. The model records a lower performance on the test set than on the training set. Exploring the issue of overfitting we uncovered that due to a bug in the pipeline the runs we submitted had not been trained on the full set but only on a small training set. Therefore in this paper we also include the results we get when trained on the full training set which demonstrate the power of ensembles.

1 Introduction

The need to check and moderate conversations and text on Social Media keeps increasing proportionally to the use of Social Media over the years (Shu et al., 2018; Rizoiu et al., 2019; Waseem and Hovy, 2016). Research into the identification of hate speech or toxic comment and fake news have recently become more popular in languages other than English because the abuse of free speech online and spread of information whether false or true extends farther than we can imagine (Vosoughi et al., 2018; Zampieri et al., 2020). GermEval 2021 (Risch et al., 2021) contains three subtasks not only aimed at identifying toxic comments in German text on social media platforms like in previous years (Struß et al., 2019) but also the classification of engaging and fact-claiming comments. In a way to help the situation of diffusing toxic content and promote positive content moderators on popular social media platforms also seek to promote texts that engage other users in a healthy conversation (Welch et al., 2016). The connection between hate speech and fake news is immense as the latter can rather stir up the masses into targeted hate towards a group of people or in some instances deadly violence (Moon et al., 2020). Therefore identifying social media content that makes a-need-to-check claim is as important as identifying hate content online.

Our participation in GermEval 2021 was in all three subtasks and involved the use of the same model architectures on all three to learn, compare and analyse how models behave on subtasks. We applied Transformer-based embeddings (BERT), RNN-based embeddings (BiLSTM) with a classifier either utilising a densely connected output layer of a simple neural network or a Support Vector Machine in an ensemble constructed with majority voting of three models on all three subtasks.

The next sections discuss in detail the dataset used for our experiment and the model architectures applied. We also discuss and compare the performances of the models on the subtasks. All code used in this experiment can be accessed via GitHub.[1]

2 Dataset and Task

The dataset provided for this competition includes a trial set of 113 user comments, a training set of 3,244 user comments and a test set of 944 user comments of German text in csv format. The training set provided consists of over 3,000 Facebook anonymized user comments that were annotated by

[1] https://github.com/kaodamie/GermEval2021_Kobby_participation

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

comment_id	comment_text	Sub1_Toxic	Sub2_Engaging	Sub3_FactClaiming
1	Ziemlich traurig diese Kommentare zu lesen. Ihr könnt euch zwar belügen, dass es den vom Menschen gemachten Klimawandel nicht gibt, nur kann man die Natur nicht belügen. Wie viele Menschen müssen denn noch auf Grund des Klimawandels ihre Lebensgrundlage verlieren oder gar Sterben, bis ihr den ernst der Lage erkannt habt?	0	0	0
2	Sag ich doch, wir befeuern den Klimawandel. Raucher können ihr Lebensende meiner Meinung nach auch gerne befeuern, nur hab ich daran kein Interesse.	0	1	1
3	Schublade auf, Schublade zu. Zu mehr Denkleistung reicht es wohl bei dir nicht.	1	0	0
4	Dummerweise haben wir in der EU und in der USA einen viel höheren CO2 Fußabdruck als z.B. die Afrikaner oder Inder.	0	0	1
5	"So lange Gewinnmaximierung Vorrang hat, wird sich das nur schleppend ändern" Da gebe ich dir recht.	0	0	0
6	@USER Schon mal was von Physik gehört?	1	0	0
7	Sollte es dann doch einen Klimawandel geben, der unabhängig vom Menschen stattfindet? Lernt er nichts von periodischen Klimaveränderungen?	0	0	0

Figure 1: Small sample of the Training Data.

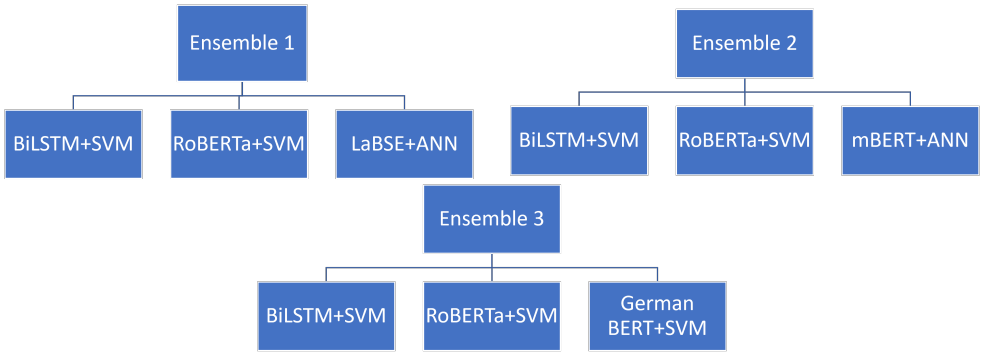

Figure 2: Ensemble models used for this experiment.

four trained annotators (Risch et al., 2021). The dataset was extracted from the home feed of the Facebook page of a political talk show of a German television broadcaster as well as the comment section discussions of posts from the same page from July 2019 till February 2021. It was shared in fully anonymized form and no user information or comment ids were revealed. Links referring to users were replaced by @USER, Links referring to the show were replaced by @MEDIUM, and links referring to the moderator of the show were replaced by @MODERATOR. The csv file contained all comments and labels for all 3 subtasks. That is to say, a user comment can be either toxic, engaging, fact-claiming or any of 2 of the labels or all 3 or neither of the labels (see Figure 1). GermEval 2021 consists of 3 subtasks (Risch et al., 2021). The first subtask is the identification of toxicity or hate speech from German text. The second and the third are the identification of engaging text and fact-claiming text, respectively. Participants were to choose any or all of the tasks they would participate in. We participated in all 3 tasks using a system of 3 different ensembles for each task (see Figure 2 for a quick overview). Submissions of the runs were submitted to Codalab.

3 Models architecture

Over the past few years, Long Short-Term Memory (LSTM) (Huang et al., 2015) and pre-trained transformer-based models (Devlin et al., 2019)

	Sub1-F1	Sub1-P	Sub1-R	Sub2-F1	Sub2-P	Sub2-R	Sub3-F1	Sub3-P	Sub3-R
Ens1	0.9750	1.0000	0.9751	0.9623	0.9273	1.0000	0.9587	0.9508	0.9667
Ens2	0.9402	1.0000	0.9024	0.9714	0.9444	1.0000	0.9594	0.9365	0.9833
Ens3	0.9750	1.0000	0.9512	0.9902	0.9808	1.0000	0.9836	0.9677	1.0000

Table 1: Results on the trial set after training on small dataset.

	Sub1-F1	Sub1-P	Sub1-R	Sub2-F1	Sub2-P	Sub2-R	Sub3-F1	Sub3-P	Sub3-R
Ens1	0.5547	0.5529	0.5565	0.6337	0.6211	0.6468	0.5970	0.5915	0.6026
Ens2	0.5545	0.5550	0.5540	0.6428	0.6406	0.6450	0.6316	0.6241	0.6392
Ens3	0.5559	0.5571	0.5547	0.6143	0.6107	0.6180	0.6150	0.6110	0.6191

Table 2: Results on the test set with models trained on *small* training set (actually submitted runs).

have proven to be effective in various NLP tasks through their ability to generate word or sentence embeddings (Qiu et al., 2020). One of such models that have widely been used in many NLP tasks is the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). It is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context to learn and produce embeddings either on sentence or word level in a transformer based architecture. The Flair embedding architecture is also an example of a model that uses a variant of bidirectional recurrent neural networks (BiLSTMs) with a conditional random field (CRF) layer to generate contextual embeddings from both directions (Akbik et al., 2018; Huang et al., 2015). In this experiment, we applied both transformer based models and Bi-directional LSTM (BiLSTM) based models to generate embeddings and further applied a Support Vector Machine (SVM) or a sigmoid activated single-layered neural network as a classifier in an ensemble of 3 models with majority voting – a simple yet effective paradigm (Kanakaraj and Guddeti, 2015; Zimmerman et al., 2018).

Each of the 3 subtasks, that is, identifying toxic, engaging and fact-claiming comments were classified with the same ensemble models. Each ensemble model however, contained three sub-models. The models were run on a standard Google Colabs runtime with a RAM size of 12 gigabyte. Below are the summaries of the sub-models.

3.1 Ensemble 1

For Ensemble 1, a sub-model with emeddings generated from the *flair* framework[2] pre-trained on the German corpus was applied. A forward and backward contextualized embeddings were generated and stacked on top of each other and then mean-pooled. An SVM classifier was fitted to the model with a linear kernel, a regularization parameter of 1, a gamma of 1 and a degree of 3. Embeddings from the XLM-RoBERTa (Conneau et al., 2019) – a multi-lingual BERT-based model designed by Facebook's AI team – was also generated for another sub-model and was also fitted with an SVM classifier with a regularization parameter of 1, a linear kernel, a gamma of 1 and a degree of 3. Finally, the last sub-model applied the language-agnostic BERT-based sentence encoder (LaBSE) with a single layered output of a fully-connected neural network with a sigmoid activation. The sub-models were not fine-tuned on the dataset due to RAM limitations.

3.2 Ensemble 2

Ensemble 2 is very similar to Ensemble 1. The only difference is that one of the sub-models does not use embeddings from a sentence encoder unlike the first Ensemble but rather embeddings were generated from fine-tuning a multilingual BERT (mBERT) and further classified with a sigmoid activated single layered output of a fully-connected neural network. SVM parameters are maintained just as with Ensemble 1.

3.3 Ensemble 3

This Ensemble model applied only SVM classifiers for its sub-models with the same parameters as stated for the other 2 Ensemble models (Hoffmann and Kruschwitz, 2020). However, unlike the other two, the third sub-model of this Ensemble applied a German based BERT model designed by Deepset AI (Chan et al., 2020). No fine-tuning was performed.

[2] https://github.com/flairNLP/flair

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

	Sub1-F1	Sub1-P	Sub1-R	Sub2-F1	Sub2-P	Sub2-R	Sub3-F1	Sub3-P	Sub3-R
Ens1	0.7024	0.7957	0.6286	0.7869	0.8536	0.7299	0.7851	0.8280	0.7466
Ens2	0.7577	0.8174	0.7060	0.8389	0.8640	0.8154	0.8148	0.8251	0.8046
Ens3	0.7886	0.8412	0.7422	0.8522	0.8864	0.8206	0.8402	0.8613	0.8201

Table 3: Results on the trial set after training on full dataset.

	Sub1-F1	Sub1-P	Sub1-R	Sub2-F1	Sub2-P	Sub2-R	Sub3-F1	Sub3-P	Sub3-R
Ens1	0.6205	0.6914	0.5629	0.6721	0.7160	0.6333	0.7211	0.7695	0.6784
Ens2	**0.6472**	0.6936	0.6067	**0.6930**	0.7197	0.6684	**0.7343**	0.7443	0.7247
Ens3	0.6241	0.6574	0.5940	0.6770	0.7023	0.6536	0.7341	0.7596	0.7103

Table 4: Results on the test set with models trained on the *full* training set.

For each of the neural networks applied in Ensemble 1 and 2, the BERT-based embedding layer was fully connected to the output layer. The output layer was activated with a sigmoid function. The neural network had a learning rate of 1-e5 , batch size of 32 and was trained with a model checkpoint on validation loss. The models were setup with 50 training epochs with early stopping on the model checkpoint at a patience of 3 epochs. The training dataset was split for train-test-validation reasons with an initial ratio of 0.8 for training. The remaining 20% was further split into 0.8 and 0.2 for validation and testing respectively. The SVM models were fitted on the whole training data.

4 Results

The results of our officially submitted runs are displayed in Table 2 (and corresponding training performance in Table 1). Note however, that the results submitted were acquired from training on a trial set of 113 comments only – an error which we only noticed after having received the results.

We subsequently re-run the three approaches – this time trained on the full training set – as illustrated in Table 4 (with corresponding training data performance in Table 3). Highest F1 performances are in bold, and we observe that Ensemble 2 consistently performs best.

The results demonstrate that, as expected, an increase in the training data has a measurable positive effect on the overall performance across all metrics.

The results recorded after training shows that the SVM models had very high metrics on the trial set whereas the ANN models had relatively low metrics peaking at 62% for F1 score, precision and recall. An ensemble approach rather seemed balanced. The Ensemble models were slightly biased towards the SVM models because in a total of three models for each ensemble, two models were SVM models for both Ensemble 1 and 2. Ensemble 3 was a model of 3 SVM models. It is fair to say that the SVM models were overfitted on the trial set. The results from the test set were lower than the results for the training data (see Table 2). Considering the fact that the training set of 113 data points is substantially smaller than the test set of 994 entries, it is also not surprising the model performed worse on the test set. The more interesting observation is that even though the training was done for a tiny dataset the results seem better than what one might expect.

Most interesting are of course the findings we derive from running our three approaches on the full training data. We observe *robust performance* of our ensemble-based approaches. We also observe that *fine-tuning* one of the models in our ensembles appears to push up performance quite substantially.

5 Conclusion

Ensemble approaches have repeatedly been shown to offer great benefits but they nevertheless rely on good underlying individual models. In our runs we combined contextual embeddings using state-of-the-art models such as BiLSTM-CRF, BERT-based models and SVM and simple neural networks as classifiers in an ensemble approach to perform binary text classification in German. We observe robust performance across different tasks, we also note a positive impact of including fine-tuned models in our ensembles.

Acknowledgements

This work was supported by the project *COURAGE: A Social Media Companion Safeguarding and Educating Students* funded by the Volkswagen Foundation, grant number 95564.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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

Julia Hoffmann and Udo Kruschwitz. 2020. UR_NLP @ haspeede 2 at EVALITA 2020: Towards robust hate speech detection with contextual embeddings. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Monisha Kanakaraj and Ram Mohana Reddy Guddeti. 2015. NLP based sentiment analysis on Twitter data using ensemble classifiers. In *2015 3rd international conference on signal processing, communication and networking (ICSCN)*, pages 1–5. IEEE.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection. *arXiv preprint arXiv:2005.12503*.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26.

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.

Marian-Andrei Rizoiu, Tianyu Wang, Gabriela Ferraro, and Hanna Suominen. 2019. Transfer learning for hate speech detection in social media. *arXiv preprint arXiv:1906.03829*.

Kai Shu, Suhang Wang, and Huan Liu. 2018. Understanding user profiles on social media for fake news detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 430–435.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, Manfred Klenner, et al. 2019. Overview of GermEval Task 2, 2019 Shared task on the identification of offensive language. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 354–365.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

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.

Vivian Welch, Jennifer Petkovic, J Pardo Pardo, Tamara Rader, and Peter Tugwell. 2016. Interactive social media interventions to promote health equity: an overview of reviews. *Health promotion and chronic disease prevention in Canada: research, policy and practice*, 36(4):63.

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). *arXiv preprint arXiv:2006.07235*.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.