GermEval 2021: Using Transformer-Based Language Models to Identify Toxic, Engaging, & Fact-Claiming Comments

Kevin Fitkau
Fachhochschule Südwestfalen

***Abstract*—The research focuses on how I approached the detection of toxic, engaging, and fact-oriented comments in GermEval 2021. To do this, I utilized freely available transformer-based models from the Huggingface Model Hub. By fine-tuning these models with various hyperparameters of the training data, I evaluated their performance. The predictions of the top two models were combined and submitted. The GermEval 2021 dataset, which comprises over 3,000 comments, served as the foundation. After careful evaluation, Electra emerged as the most effective model, achieving an F1-score of 66% in the first task and 68% in the second. This approach proved particularly effective in task 3, where I achieved an impressive F1-score of 0.75.**

I. Introduction

With the advent and widespread use of social media, user-generated content has become a fundamental component of online communication. However, platforms are grappling with the increasing proliferation of offensive content. The identification of toxic language in social media has become a central concern and is expected to continue gaining importance. Research in this area has focused on developing models capable of detecting various forms of negative content, such as hate speech, cyberbullying, and abuse. GermEval2021 focuses on the identification of different types of comments in social media. The collective task of this year's competition is divided into three distinct classifications: toxic, engaging, and fact-oriented. Based on experiences from previous GermEvals, it can be assumed that this year's tasks require similar classification approaches. In particular, the identification of comments that encourage readers to participate in discussions is paramount. Therefore, the precise classification of these comments is crucial as platforms must continuously monitor and validate user-generated content. To address this task, BERT and ELECTRA were fine-tuned. This document presents the fine-tuning methods, as well as the results for GermEval21, and evaluates the performance of the transformer models based on the GermEval-2021 dataset.

II. Related work

The issue of identifying offensive language in online discussions has gained significant prominence in recent years. While much of the research has focused on English data due to the availability of annotated datasets, datasets for offensive language in other languages have also been created and studied. Researchers worldwide are grappling with the problem of offensive content in social media, whether it's for Greek [1], Arabic [2], or Italian [3]. Previous approaches to addressing this problem have ranged from traditional machine learning classifiers like logistic regression and SVMs to various deep learning models.

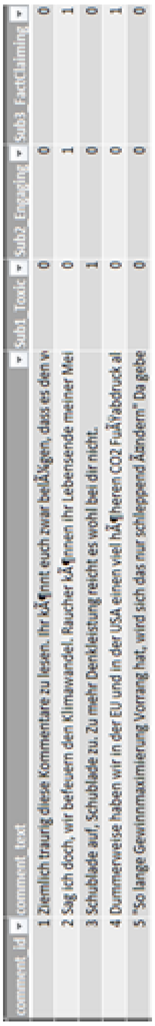
The introduction of BERT [4] has propelled the use of pre-trained transformer models for classifying offensive language. The application of pre-trained BERT models, as well as BERT-based models, has demonstrated their ability to deliver competitive performance in competitions. Additionally, language-specific and multilingual models have been developed to support NLP research in various languages, such as gBERT [5] for German or Philip May's model, one of the authors of 'german-nlpgroup/electra-base-german-uncased', which evaluated multiple models with the GermEval 2018 dataset for German, AraBERT for Arabic [6], and the multilingual XLM-R [7], which has been successfully used for identifying offensive language.

III. GermEval Dataset

The models were trained on a dataset of German-language tweets provided for classification as part of GermEval21. This dataset comprises over 3000 Facebook comments extracted from a page of a political talk show of a German television channel. The training dataset consists of approximately 3244 entries, including 1074 entries without toxic, fact-related, or engaging content. An example of the training data is depicted in Figure 1. Each entry in the training data includes a comment ID (comment_id) for identification, along with the training comment itself. Additionally, there are three columns (Toxic, Engaging, and FactClaiming), each containing either 0 or 1. A one in any of these columns indicates that the comment belongs to the corresponding category, while a zero indicates the opposite.

IV. Task description

Participants were allowed to participate in one, two, or all three subtasks and submit a maximum of three runs per task. The tasks include the Toxic Comment Classification, Engaging Comment Classification, and the final task is Fact-Claiming Comment Classification.

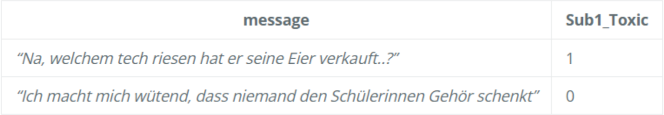


comment_id	comment_text	Sub1_Toxic	Sub2_Engaging	Sub3_FactClaiming
1	Ziemlich traurig diese Kommentare zu lesen. Ihr kÃ¶nnt euch zwar belÃ¼gen, dass es den w	0	0	0
2	Sag ich doch, wir befeuern den Klimawandel. Raucher kÃ¶nnen ihr Lebensende meiner Mei	0	1	1
3	Schublade auf, Schublade zu. Zu mehr Denkleistung reicht es wohl bei dir nicht.	1	0	0
4	Dummerweise haben wir in der EU und in der USA einen viel hÃ¶heren CO2 FuÃŸabdruck al	0	0	1
5	"So lange Gewinnmaximierung Vorrang hat, wird sich das nur schleppend Ã„ndern" Da gebe	0	0	0

Figure 1. Example table of GermEval21 data records

A. Toxic Comment Classification

The first subtask involved the identification of toxic or offensive comments in online discussions that could potentially offend or harm readers. This subtask builds upon previous GermEval tasks related to detecting offensive language and was further extended in this study.

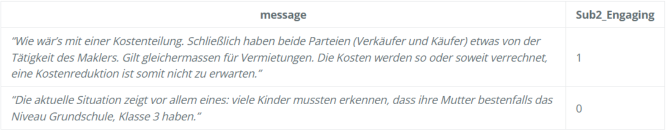


message	Sub1_Toxic
"Na, welchem tech riesen hat er seine Eier verkauft..?"	1
"Ich macht mich wütend, dass niemand den Schülerinnen Gehör schenkt"	0

Figure 2. Example of toxic comments

B. Engaging Comment Classification

In addition to detecting toxic language, community managers and moderators are increasingly interested in identifying particularly valuable user-generated content. This may include highlighting rational, respectful, and reciprocal comments to give them more visibility. Such comments encourage readers to engage in the discussion, enhance the positive perception of the discussion providers, and promote a more fruitful and less violent exchange. Therefore, the second subtask aimed to identify precisely those comments that could encourage readers to participate in the conversations.

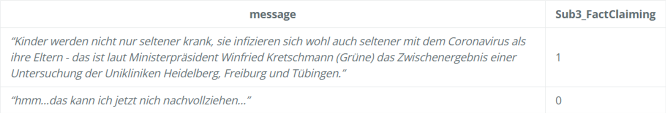


message	Sub2_Engaging
"Wie wär's mit einer Kostenteilung. Schließlich haben beide Parteien (Verkäufer und Käufer) etwas von der Tätigkeit des Maklers. Gilt gleichermassen für Vermietungen. Die Kosten werden so oder soweit verrechnet, eine Kostenreduktion ist somit nicht zu erwarten."	1
"Die aktuelle Situation zeigt vor allem eines: viele Kinder mussten erkennen, dass ihre Mutter bestenfalls das Niveau Grundschule, Klasse 3 haben."	0

Figure 3. Example of Engaging Comment comments

C. Fact-Claiming Comment Classification

In addition to ensuring non-hostile debates, platforms and moderators are challenged by the rapid spread of misinformation and fake news. They are therefore under pressure to verify and authenticate posted information to fulfill their responsibility as information providers and disseminators. Hence, the final subtask involved identifying fact-based comments, although the accuracy of the comments themselves was not to be verified.



message	Sub3_FactClaiming
"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."	1
"hmm...das kann ich jetzt nich nachvollziehen..."	0

Figure 4. Example of fact-claiming comments

D. Evaluation Metrics

Classification performance is assessed using the common assessment measures of precision, recall and F-score. These measures are calculated for each of the individual classes in the three subtasks. In addition to the class-specific metrics, the macro-average „precision", „recall" and „F-score" as well as the overall accuracy are calculated for each task. Note, however, that the precision is not used in the evaluation rank and is therefore only approximated using WANDB information. To evaluate the data, a ZIP file with the required subtasks had to be uploaded to Codalab.

V. Fine-Tuning gBERT and Electra

In this section I explain the fine tuning of both the German BERT and Electra models. I also present my settings for these models.

A. gBert

One approach was to use a German-language BERT model that is case-sensitive. This model was trained with 12 GB of raw data from the German Wikipedia dump, the OpenLegalData dump and news articles. It has the same size as the English-language BERT model with 12 layers, 768 hidden units and 12 attention heads. The model comprises a total of 110 million parameters [5].

A batch size of 8 was used for the fine-tuning of BERT. The Adam Optimiser with a learning rate of 1e-5 was used. The evaluation steps were performed every 250 steps and the maximum sequence length was 128 tokens. Training was performed for one to five epochs, and it was found that no further improvements were achieved with more epochs. The maximum number of training runs was set at 1500. It was found that at higher settings there was a deterioration in F1 score and accuracy. BERT uses tokenised parts of words instead of tokenised words. Here are two examples of tokenisation:

Example 1:

- Text: „Sag ich doch, wir befeuern den Klimawandel. Raucher können ihr Lebensende meiner Meinung nach auch gerne befeuern, nur hab ich daran kein Interesse.“
- Token: [’Sa’, ’##g’, ’ich’, ’doch’, ’,’, ’wir’, ’bef’, ’##euer’, ’##n’, ’den’, ’Klima’, ’##wandel’, ’.’, ’Rauch’, ’##er’, ’können’, ’ihr’, ’Lebens’, ’##ende’, ’meiner’, ’Meinung’, ’nach’, ’auch’, ’gerne’, ’bef’, ’##euer’, ’##n’, ’,’, ’nur’, ’hab’, ’ich’, ’daran’, ’kein’, ’Interesse’, ’.’]

It can be observed that the tokenisation symbol ## signals that the words belong together and combines them into a token.

Example 2:

- Text: „Ziemlich traurig diese Kommentare zu lesen. Ihr kÃ¶nnt euch zwar belÃ¼gen, dass es den vom Menschen gemachten Klimawandel nicht gibt, nur kann man die Natur nicht belÃ¼gen. Wie viele Menschen mÃ¼ssen denn noch auf Grund des Klimawandels ihre Lebensgrundlage verlieren oder gar Sterben, bis ihr den ernst der Lage erkannt habt?“
- Token: [’Zie’, ’##mlich’, ’traur’, ’##ig’, ’diese’, ’Kommentare’, ’zu’, ’lesen’, ’.’, ’Ihr’, ’[UNK]’, ’[UNK]’, ’n’, ’##nt’, ’euch’, ’zwar’, ’[UNK]’, ’,’, ’dass’, ’es’, ’den’, ’vom’, ’Menschen’, ’gemachten’, ’Klima’, ’##wandel’, ’nicht’, ’gibt’, ’,’, ’nur’, ’kann’, ’man’, ’die’, ’Natur’, ’nicht’, ’[UNK]’, ’.’, ’Wie’, ’viele’, ’Menschen’, ’[UNK]’, ’denn’, ’noch’, ’auf’, ’Grund’, ’des’, ’Klima’, ’##wandel’, ’##s’, ’ihre’, ’Lebens’, ’##grundlage’, ’verlieren’, ’oder’, ’gar’, ’Sterbe’, ’##n’, ’,’, ’bis’, ’ihr’, ’den’, ’ernst’, ’der’, ’Lage’, ’erkannt’, ’hab’, ’##t’, ’?’]

Please note that the second sentence has a formatting error that is treated as an unknown symbol. This may result in the classifier not recognising words correctly. In addition, the tokeniser should be able to correctly handle words written in capital letters. However, abbreviations and capital letters must be learnt by the tokenizer, such as „DE“ for German. Normally, such abbreviations are already integrated in the model manufacturer’s model, as they often occur in the pre-training data for the German BERT model.

B. Electra

The Electra model can be used as an alternative to BERT in most tasks. It is implemented as an extended BERT class and was trained with 12 GB of raw data from the German Wikipedia dump, OpenLegalData dump and news articles. The model is the same size as the English and German BERT models, with 12 layers, 768 hidden units and 12 attention heads. It also includes 110 million parameters like BERT. However, Electra aims to reduce computation time and resources while maintaining high performance. The pre-training task in Electra is based on recognising substituted tokens in the input sequence. Two transformer models are required for this task: a generator and a discriminator. The same settings were used for fine-tuning Electra as for gBERT. The batch size was 8, the Adam optimiser with a learning rate of 1e-5 was used, and the evaluation steps were performed every 250 steps. The maximum sequence length was set to 128 tokens. Training was performed for one to five epochs, and it was found that no further improvements were achieved with more epochs. The maximum number of training runs was set at 1500, and it was found that this number should be sufficient for a duration of 5 epochs. It was also found that higher settings lead to a deterioration in F1 score and accuracy.

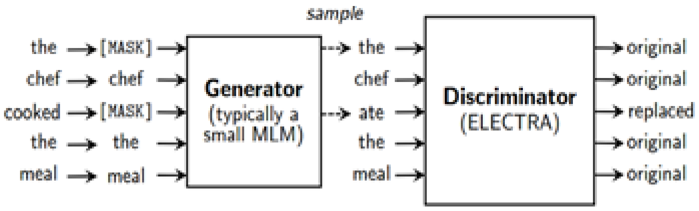


Figure 5. ELECTRA-Training process

Example 1:

- Text: „Sag ich doch, wir befeuern den Klimawandel. Raucher können ihr Lebensende meiner Meinung nach auch gerne befeuern, nur hab ich daran kein Interesse.“
- Token: [’sag’, ’ich’, ’doch’, ’,’, ’wir’, ’befe’, ’##uern’, ’den’, ’klimawandel’, ’.’, ’raucher’, ’können’, ’ihr’, ’lebens’, ’##e’, ’##nde’, ’meiner’, ’meinung’, ’nach’, ’auch’, ’gerne’, ’befe’, ’##uern’, ’,’, ’nur’, ’hab’, ’ich’, ’daran’, ’kein’, ’interesse’, ’.’]

It should be noted that the hash symbol signals that the words belong together, but indicates a different affiliation than gBERT.

Example 2:

- Text: „Ziemlich traurig diese Kommentare zu lesen. Ihr kÃ¶nnt euch zwar belÃ¼gen, dass es den vom Menschen gemachten Klimawandel nicht gibt, nur kann man die Natur nicht belÃ¼gen. Wie viele Menschen mÃ¼ssen denn noch auf Grund des Klimawandels ihre Lebensgrundlage verlieren oder gar Sterben, bis ihr den ernst der Lage erkannt habt?“
- Token: [’ziemlich’, ’traurig’, ’diese’, ’kommentare’, ’zu’, ’lesen’, ’.’, ’ihr’, ’k’, ’##ã’, ’¶’, ’nn’, ’##t’, ’euch’, ’zwar’, ’bel’, ’##ã’, ’##¼’, ’##gen’, ’,’, ’dass’, ’es’, ’den’, ’vom’, ’menschen’, ’gemachten’, ’klimawandel’, ’nicht’, ’gibt’, ’,’, ’nur’, ’kann’, ’man’, ’die’, ’natur’,

'nicht', 'bel', '##ã', '##¼', '##gen', '.', 'wie', 'viele', 'menschen', 'm', '##ã', '##¼', '##ssen', 'denn', 'noch', 'auf', 'grund', 'des', 'klimawandels', 'ihre', 'lebens', '##grundlage', 'verlieren', 'oder', 'gar', 'sterben', ',', 'bis', 'ihr', 'den', 'ernst', 'der', 'lage', 'erkannt', 'habt', '?']

VI. Results

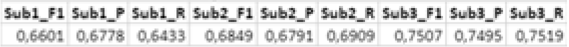
I divide the results into two categories: Test results that were uploaded and processed on WANDB and Codalab results that belong to the GermEval2021 competition and were evaluated there. WANDB stands for „Weights & Biases“ and is a platform for tracking experiments and visualising model training. It enables researchers and developers to organise their experiments, track metrics, compare models and visualise results. Codalab is a platform for scientific competitions where researchers can upload their models and test them on standardised data sets. It is often used for NLP and machine learning competitions where participants can evaluate and compare their models on a given test dataset. I used WANDB for the private results, while I used Codalab for the public results.

A. WANDB Results

Before I evaluated my results, I uploaded my runs with WANDB to test the evaluation for the final result. I evaluated the test results using both the gBERT model and the Electra model. I found that my best results were achieved with the Electra model: F1 score was 70% for Engaging, 57% for Toxic and 70% for Fact-Claiming classification. However, the fact claiming result of 70% for the Electra model was worse than for the gBERT model. The latter performed only slightly worse overall than the Electra model with the same settings. It achieved an F1 score of 65% for engagement classification, 54% for toxic classification and 71% for fact claiming classification. The accuracy of the two models was between 78% and 83%, while the precision varied between 83% and 89%. The recall was in the same range as the F1 score for the different runs.

B. Codalab Results

In the following, I present the results of the classification tasks using the Electra model for the German language. Due to the superior performance of the Electra model compared to the gBERT model, I only uploaded the results of the Electra model in the CSV file. After successfully submitting the answer file (answer.CSV) shortly before the deadline, I received the evaluation results back within a few days. I achieved my outstanding result in the final classification task, in which factual comments were classified. I achieved an F1 score, precision and recall of 75% each. This is shown in Figure 6 above.



Sub1_F1	Sub1_P	Sub1_R	Sub2_F1	Sub2_P	Sub2_R	Sub3_F1	Sub3_P	Sub3_R
0,6601	0,6778	0,6433	0,6849	0,6791	0,6909	0,7507	0,7495	0,7519

Figure 6. Results GermEval2021

VII. Fazit

In this post, I present my participation in GermEval-2021, an event that offers participants the opportunity to test computational models for identifying toxic, inciting and factual comments. In my experiment, I used different neural transformer models such as gBert and Electra. I found that Electra performed better than gBert for toxic and stimulating comments. On the other hand, gBert was superior for factual comments. Further models should be tested for future research, as other transformer models are also constantly improving. In addition, alternative approaches to data processing could further improve performance. It would also be interesting to work on similar tasks in other languages and test models that have been trained with extensive test data. It is also worth following the development of transformer models in other countries closely.

References

[1] Z. Pitenis, M. Zampieri, and T. Ranasinghe, “Offensive language identification in Greek,” English, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, *et al.*, Eds., Marseille, France: European Language Resources Association, May 2020, pp. 5113–5119, isbn: 979-10-95546-34-4. [Online]. Available: https://aclanthology.org/2020.lrec-1.629.

[2] H. Mubarak, A. Rashed, K. Darwish, Y. Samih, and A. Abdelali, “Arabic offensive language on Twitter: Analysis and experiments,” in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, N. Habash, H. Bouamor, H. Hajj, *et al.*, Eds., Kyiv, Ukraine (Virtual): Association for Computational Linguistics, Apr. 2021, pp. 126–135. [Online]. Available: https://aclanthology.org/2021.wanlp-1.13.

[3] P. Chiril, F. Benamara Zitoune, V. Moriceau, M. Coulomb-Gully, and A. Kumar, “Multilingual and multitarget hate speech detection in tweets,” in *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, E. Morin, S. Rosset, and P. Zweigenbaum, Eds., Toulouse, France: ATALA, Jul. 2019, pp. 351–360. [Online]. Available: https://aclanthology.org/2019.jeptalnrecital-court.21.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “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)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423. [Online]. Available: https://aclanthology.org/N19-1423.

[5] B. Chan, S. Schweter, and T. Möller, “German’s next language model,” in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds., Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 6788–6796. doi: 10.18653/v1/2020.coling-main.598. [Online]. Available: https://aclanthology.org/2020.coling-main.598.

[6] W. Antoun, F. Baly, and H. Hajj, “AraBERT: Transformer-based model for Arabic language understanding,” English, in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, H. Al-Khalifa, W. Magdy, K. Darwish, T. Elsayed, and H. Mubarak, Eds., Marseille, France: European Language Resource Association, May 2020, pp. 9–15, isbn: 979-10-95546-51-1. [Online]. Available: https://aclanthology.org/2020.osact-1.2.

[7] A. Conneau, K. Khandelwal, N. Goyal, *et al.*, *Unsupervised cross-lingual representation learning at scale*, 2020. arXiv: 1911.02116 `[cs.CL]`.