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

Christian Gawron
Fachhochschule Südwestfalen
Frauenstuhlweg 31
58644 Iserlohn
gawron.christian@fh-swf.de

Sebastian Schmidt
Fachhochschule Südwestfalen
Frauenstuhlweg 31
58644 Iserlohn
schmidt.sebastian2@fh-swf.de

Abstract

In this paper we describe the methods we used for our submissions to the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. For all three subtasks we fine-tuned freely available transformer-based models from the Huggingface model hub. We evaluated the performance of various pre-trained models after fine-tuning on 80% of the training data with different hyperparameters and submitted predictions of the two best performing resulting models. We found that this approach worked best for subtask 3, for which we achieved an F1-score of 0.736.

1 Introduction

Compared to the detection of offensive language in GermEval 2018 (Wiegand et al., 2019) and 2019 (Struß et al., 2019), this year's task adds two important additional categories found in social media comments, namely *engaging* and *fact-claiming* comments (Risch et al., 2021). With federal elections being held in 2021, identifying fact-claiming statements (subtask 3) in German social media posts has gained additional relevance as "fake news" might have had an influence on other important elections, e. g. the 2016 US presidential elections (Allcott and Gentzkow, 2017; Bovet and Makse, 2019). A system identifying fact-claiming comments could help to identify potential attempts to spread false factual statements.

The identification of *engaging* comments (subtask 2) is potentially interesting for the ranking algorithms used by social network providers. Increasing the visibility of these comments might help improving the attractiveness of a social network by encouraging the users to employ a more respectful and rational style of discussion.

With the classification of toxic comments (subtask 1), the GermEval Shared Tasks on the identification of offensive language mentioned above are continued. This category is also useful for the ranking algorithms of social media providers and could be used to decrease the visibility of such comments. However, we have made the experience that this year's *toxic* category is harder to identify than the former offensive categories – at least by our approach.

The best performing systems in GermEval 2019 were based on BERT (Devlin et al., 2019). Leveraging the transformer architecture (Vaswani et al., 2017) with its attention mechanism, BERT is able to model relations between words and to create semantic embeddings of sentences (Feng et al., 2020). In the last two years, various modifications of BERT like RoBERTa (Liu et al., 2019) or ELECTRA (Clark et al., 2020) have been proposed and shown to achieve state-of-the-art results on various NLP tasks. Other transformer-based models, especially GPT-2 (Radford et al., 2019) and its successor GPT-3, even made it into the press (Drösser, 2020) due to their ability to create high-quality artificial text or to create source code for various programming languages (Metz, 2020).

Probably the most important feature of these models is that they allow transfer learning: After an unsupervised *pre-training*, the resulting models can be *fine-tuned* for various NLP tasks like token classification (e. g. NER) and sequence classification. Pre-training a language model for German imposes two challenges: It requires a large corpus of text and is computationally expensive. According to Brown et al. (2020), GPT-3 was trained on a corpus of 400 billion byte-pair-encoded tokens or roughly 570 GB of text. Compared to this, the "Huge German Corpus"[1] with 204 million tokens is rather small. BERT-large was trained on 64 TPU chips for four days at an estimated cost of $7,000

[1] See https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/hgc

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

(Schwartz et al., 2020), the training of GPT-3 took 3.640 petaflop-days (Brown et al., 2020). Due to the high computational effort and costs to train a model from scratch, we decided to evaluate freely available pre-trained models for our system.

For English, pre-trained models of high quality are freely available for most of the model architectures mentioned above (with the notable exception of GPT-3). Unfortunately, the groups which developed and trained these models and the companies behind them do not deem German important enough to provide pre-trained models for German. Although there is currently no active academic community in Germany training and publishing these language models, there is a growing number of companies and individuals publishing such pre-trained models. For example, Deepset.ai has published a German ELECTRA model achieving an F1-score (macro average) of 80.70% on GermEval 2018 Coarse and an F1-score (micro average) of 88.95% on GermEval 2014 (Chan et al., 2020). Philipp Reissel and Philip May have published both a German ELECTRA model (Reissel and May, 2020) and a "German colossal, cleaned Common Crawl corpus" (GC4) (Reissel and May, 2021) with about 540 GB of German text from the web It would be helpful for the development of language models for German if an extensive and high-quality corpus of German language text would be available through infrastructure projects like CLARIN-D (Hinrichs and Trippel, 2017).

2 Setup

Our experiments were performed using Jupiter Notebooks (Kluyver et al., 2016). This had the advantage that we could use local computing resources and cloud platforms like Google Colaboratory (Bisong, 2019) without modifications to the code. The code used to generate our submissions is available on GitHub[2].

We used the web application *Weights & Biases* (Biewald, 2020) to record and compare the results of experiments with different language models and hyperparameters (learning rate, number of training epochs), which was of great help especially when using cloud-based computing resources without a persistent storage medium.

[2]The repository `https://github.com/fhswf/GermEval2021` will be made public after the submission of this paper.

3 Model Library

A large repository of pre-trained transformer based language models along with an open-source library of implementations of them is operated by Huggingface (Wolf et al., 2020). As of July 2021, the *model hub* contains about 2,900 pre-trained models for English and more than 200 pre-trained models for German provided by a fast-growing number of contributors, including the groups mentioned above. Due to the large number of available pre-trained models for German, we decided to use the Huggingface transfer library for our submission and to choose among the models available on the model hub.

The transformer library makes it very easy to use and to fine-tune the models provided on the hub. Besides the model implementations, it also contains recent optimization algorithms like AdamW (Loshchilov and Hutter, 2017) and Adafactor (Shazeer and Stern, 2018), provides integration with the experiment-tracking software *Weights & Biases* (Biewald, 2020), code for loading and handling training data, and commonly used metrics.

4 Data Preprocessing

The transformer-based language models we used for our experiments use either SentencePiece (Kudo and Richardson, 2018) or byte pair encoding (Gage, 1994) for tokenization and can handle rare words and emojis. So we did actually not preprocess the texts in any way.

One of the models we used in our experiments, `german-nlp-group/electra-base-german-uncased`, is an uncased model that converts all characters to lower case during tokenization. Unlike other 'uncased' models published on the model hub, this model does not remove accents.

5 Model Selection

With more than 200 pre-trained models for German available on the model hub, we needed to do some preselection for our experiments. Philip May, one of the authors of `german-nlp-group/electra-base-german-uncased`, has evaluated several models on the GermEval 2018 dataset (see figure 1).

We chose the best three models of this evaluation as our candidates. Due to the success of GPT-2 on various NLP tasks (Radford et al., 2019), we also included `benjamin/gerpt2-large`, a German

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

	Sub1_Toxic			Sub2_Engaging			Sub3_FactClaiming		
Submission	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.
deepset/gelectra-large	**0.707**	0.743	0.675	**0.697**	0.694	0.700	0.734	0.728	0.740
benjamin/gerpt2-large	0.658	0.678	0.640	0.690	0.684	0.696	**0.736**	0.736	0.735

Table 1: Results of our submissions based on the models deepset/gelectra-large and benjamin/gerpt2-large.

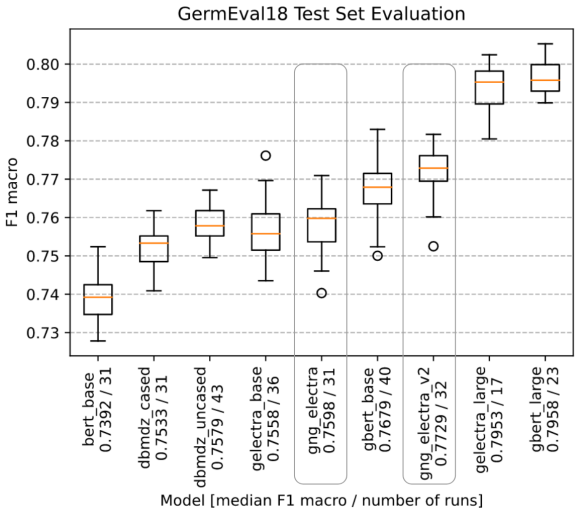

Figure 1: Results of some German language models on the GermEval 2018 dataset. Figure by Philip May, taken from the `german-nlp-group/electra-base-german-uncased` model card.

GPT-2 model recently published by Minixhofer (2020), an AI student from Johannes Kepler Universität Linz.

The following list contains some information on these models. Since we are not sure how to calculate the number of model parameters from the specification in the model configuration file, we specify the size of the binary file containing the model parameters as a measure of model complexity.

`gbert-large` has been published by Chan et al. (2020). It is a large BERT model with a binary size of 1.3 GB.

`gelectra-large` by the same group is a German ELECTRA model. The binary size is also 1.3 GB.

`electra-base-german-uncased` by Reissel and May (2020) is a smaller ELECTRA model with a binary size of 424 MB.

`gerpt2-large` published by Minixhofer (2020) is a GPT-2 model using an embedding dimension of 1280, 1024 position encodings and 20 attention heads. Although GPT-2 is mainly used for text generation, it also produces sentence embeddings which can be used for text classification. The transformer library provides the class `GPT2ForSequenceClassification` for this purpose. With a size of 3.2 GB it is the largest model we used.

6 Computing Resources

Most calculations were done on a local server using a Tesla V100S GPU card. We used fp16 precision for the training runs on the V100S for better performance as some tests with double precision did not show better results. In addition, we used cloud-based computing resources provided by GraphCore and Google Colaboratory.

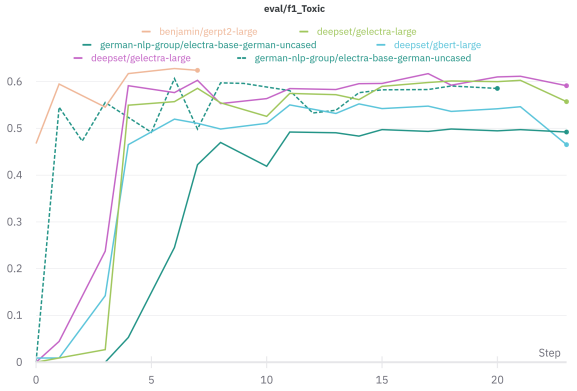

Figure 2: F1 scores of different experiments for subtask 1 with a train-test split of 0.8.

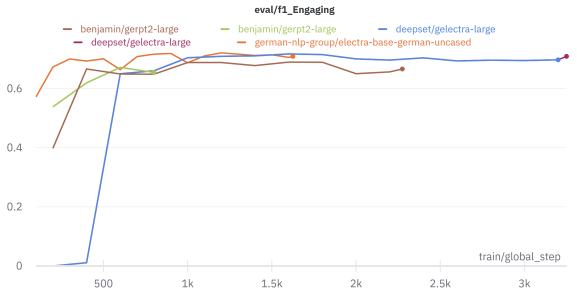

Figure 3: F1 scores of different experiments for subtask 2 with a train-test split of 0.8.

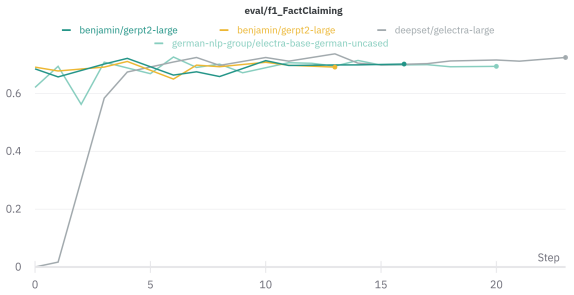

Figure 4: F1 scores of different experiments for subtask 3 with a train-test split of 0.8.

7 Results

Using the four models (see section 5) we performed several training runs with a train-test split of 80%. We did not have the time and computing resources to do a systematic hyperparameter optimization but rather tried different learning rates and number of training epochs. Figures 2 – 4 show the resulting F1-scores of several runs ans models for the three subtasks. Unfortunately, the fluctuations of the F1-scores measured on the 20% test split during the training were about as large as the differences between the different models. At this point, we would have needed more time and resources to perform a larger number of training runs and a statistical analysis similar to the one shown in figure 1. In some runs, declining F1-scores at the end of the training runs indicated overfitting – additional training data would probably have improved the results.

Overall, we achieved the best results by fine-tuning `deepset/gelectra-large` and `bjamin/gerpt2-large`. For the final system submissions, we fine-tuned these two models using the complete training dataset for all three subtasks. Table 1 shows the scores of the two submissions on the test data of the Shared Task.

8 Using Additional Training Data

Assuming that offensive language is also considered toxic, we tried to add data from GermEval 2018 and 2019 to our training dataset for subtask 1. However, compared to experiments without this additional training data, accuracy and F1-score on our validation dataset (i. e. 20% of this year's training data) were worse for these experiments. At least for an AI, toxic comments on facebook seem to be quite different from offensive language used on twitter.

9 Error Analysis

Before the gold labels were released, we compared our model predictions with our personal predictions for the first test comments. When we looked at the gold labels, we were surprised by some of the labels, especially with respect to examples having more than one label.

For example, our system flagged a fact claim in comment 3246

> @USER , ich glaube,Sie verkrnnen gründlich die Situation. Deutschland mischt sich nicht ein, weil die letzte Einmischung in der Ukraine noch nicht bereinigt ist. Es geht nicht ums Militär

which we considered correct. We did not expect that this comment is also considered engaging.

In the case of comment 3248

> Als jemand, der im real existierenden Sozialismus aufgewachsen ist, kann ich über George Weineberg nur sagen, dass er ein Voll...t ist. Finde es schon gut, dass der eingeladen wurde. Hat gezeigt, dass er viel Meinung hat, aber offensichtlich wenig Ahnung. Er hat sich eben so gut wie er kann, für alle sichtbar, zum Trottel gemacht.

we agreed with our system that the second sentence ("I think it's good that he was invited") could be considered engaging, but according to the gold labels, this comment is only toxic. On the other hand, comment 3269

> Sry aber Preetz hat nicht viel beizutragen. Er MUSS der Politik in den Hintern kriechen damit sein Verein Zuschauer ins Stadion bekommt. Er ist abhängig von der Politik.

is both toxic and engaging according to the gold labels, while we agreed with our system that this is only toxic.

These three examples demonstrate that this year's task is really hard – even for humans. It would be interesting to measure the score of human annotators getting just the category names and the training examples.

10 Conclusion

When we first looked at the development data, our impression was that fact-claiming statements would be the hardest category to recognize for an NLP system due to the wide range of different facts in the statements. The rather low range of annotator agreement of $0.73 < \alpha < 0.84$ for subtask 3 also suggests that this should be the "hard" category. We were quite surprised that our system actually achieved the best F1-score (0.736 in the case of `benjamin/gerpt2-large`) for this category.

Regarding the toxic category, the F1-score of 0.707 on subtask 1 is surprisingly low considering the F1-score of `deepset/gelectra-large` of about 0.80 reported by Chan et al. (2020) on GermEval 2018 (coarse). This year's 'toxic' category seems to be quite different from the offensive language category of the GermEval tasks in 2018 and 2019 and – at least for an AI – more difficult to recognize.

The approach we used to create our submissions is a rather simple one that did not require preprocessing of the training data or much programming. Free libraries containing implementations of a wide range of language models and the availability of an increasing number of pre-trained model instances make it quite easy to apply state-of-the-art language models for NLP tasks like text classification. It still, however, requires some coding to train and select models and to create predictions for the test dataset. Integrated tools like the recently announced AutoNLP[3] will probably enable non-experts (and non-coders) to train such models in the next few years.

Acknowledgments

This research was supported by grants from NVIDIA and utilized NVIDIA CUDA on Tesla & Ampere GPUs. This research also used free computing resources provided by the GraphCore Academic Program and Google Colab.

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Company website. Application available on wandb.com, Last accessed on 2021-07-12.

Ekaba Bisong. 2019. Google colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, pages 59–64. Apress, Berkeley, CA.

Alexandre Bovet and Hernán A. Makse. 2019. Influence of fake news in twitter during the 2016 us presidential election. *Nature Communications*, 10(1):7.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Christoph Drösser. 2020. Sie klingt wie wir. Eine Software vermittelt die Illusion eines Zwiegesprächs. *Die Zeit*, 54/2020.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Erhard Hinrichs and Thorsten Trippel. 2017. CLARIN-D: eine Forschungsinfrastruktur für die sprachbasierte Forschung in den Geistes- und Sozialwissenschaften. *Bibliothek Forschung und Praxis*, 41(1):45–54.

[3] See `https://huggingface.co/autonlp`

Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. 2016. Jupyter notebooks – a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87–90. IOS Press.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Cade Metz. 2020. Meet GPT-3. It has learned to code (and blog and argue). *New York Times*.

Benjamin Minixhofer. 2020. GerPT2-large – A large German GPT2. Huggingface model hub. `https://huggingface.co/benjamin/gerpt2-large`, Last accessed on 2021-07-12.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Philipp Reissel and Philip May. 2020. German Electra Uncased. Huggingface model hub. `https://huggingface.co/german-nlp-group/electra-base-german-uncased`, Last accessed on 2021-07-12.

Philipp Reissel and Philip May. 2021. GC4 Corpus. GitHub pages. `https://german-nlp-group.github.io/projects/gc4-corpus.html`, Last accessed on 2021-07-12.

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.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green ai. *Commun. ACM*, 63(12):54–63.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. *CoRR*, abs/1804.04235.

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 *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg*, pages 352–363, München [u.a.]. German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2019. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria September 21, 2018*, pages 1–10. Austrian Academy of Sciences, Vienna, Austria.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rmi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.