AIT_FHSTP at GermEval 2021: Automatic Fact Claiming Detection with Multilingual Transformer Models

Jaqueline Böck [2], Daria Liakhovets [1], Mina Schütz [1], Armin Kirchknopf [2], Djordje Slijepčević [2], Matthias Zeppelzauer [2], Alexander Schindler [1]

[1] Austrian Institute of Technology GmbH
Giefinggasse 4, 1210 Vienna, Austria
{daria.liakhovets.fl, mina.schuetz, alexander.schindler}@ait.ac.at

[2] St. Pölten University of Applied Sciences
Institute of Creative Media Technologies
3100 St. Pölten, Austria
{jaquelineboeck1}@gmx.at
{armin.kirchknopf, djordje.slijepcevic, matthias.zeppelzauer}@fhstp.ac.at

Abstract

Spreading ones opinion on the internet is becoming more and more important. A problem is that in many discussions people often argue with supposed facts. This year's GermEval 2021 focuses on this topic by incorporating a shared task on the identification of fact-claiming comments. This paper presents the contribution of the AIT_FHSTP team at the GermEval 2021 benchmark for task 3: "identifying fact-claiming comments in social media texts". Our methodological approaches are based on transformers and incorporate 3 different models: multilingual BERT, GottBERT and XML-RoBERTa. To solve the fact claiming task, we fine-tuned these transformers with external data and the data provided by the GermEval task organizers. Our multilingual BERT model achieved a precision-score of 72.71%, a recall of 72.96% and an F1-Score of 72.84% on the GermEval test set. Our fine-tuned XML-RoBERTa model achieved a precision-score of 68.45%, a recall of 70.11% and a F1-Score of 69.27%. Our best model is GottBERT (i.e., a BERT transformer pretrained on German texts) fine-tuned on the GermEval 2021 data. This transformer achieved a precision of 74.13%, a recall of 75.11% and an F1-Score of 74.62% on the test set.

1 Introduction

Today's social media platforms allow any individual to share information and opinions easily and quickly across a wide audience with almost no restrictions. However, not only obviously offensive comments, but also comments and posts with false information are becoming a serious problem on the Internet. The sheer amount of available information and content generated every day makes it impossible to verify all information. Thus, misinformation and false information can easily spread and influence people and their decisions, which has a strong impact on our society.

As a workshop part of the KONVENS 2021 (Konferenz zur Verarbeitung natürlicher Sprache / Conference of Natural Language Processing) the GermEval 2021 focuses on the problem of fact claiming, i.e., the identification of content in social media that contains potential facts that need to be checked (Risch et al., 2021). The identification of such fact claiming content is a first step in the information verification process to separate relevant from irrelevant information for fact checking. Our team participated in the fact claiming task (task 3: identification of fact-claiming comments) of GermEval 2021 and this paper presents our methodology and the results. To solve the task we fine-tuned (supervised) the pre-trained transformer models with the original GermEval 2021 data and external data, i.e., the ClaimBuster dataset (Arslan et al., 2020). The employed datasets and our general approach are described in Section 2. A detailed description of the transformer-based mod-

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

els is provided in Section 3. In Section 4, our experimental setup is introduced. The results can be found in Section 5 followed by a brief discussion and conclusion in Section 6.

2 Methodological Approach

The GermEval 2021 provided one labeled dataset for all three tasks (task 1 and 2 not considered in our contribution). The data for task 3 contained approx. 1/3 of content that mentions claimed facts and 2/3 with no claimed facts. We applied three pre-trained transformer models (Vaswani et al., 2017) to encode and classify the content for this task, namely: German OSCAR text trained BERT (GottBERT) (Scheible et al., 2020), multilingual BERT (mBERT) (Devlin et al., 2019b) and XLM-RoBERTa (XLM-R) (Conneau et al., 2019). Transformers are usually pre-trained on a large general corpus and can be used for many natural language processing (NLP) downstream tasks, which makes them especially useful for small training corpora (Liu et al., 2019). Compared to mBERT and XLM-R, which are both pre-trained on multilingual data, GottBERT is the only one that was trained on one language (German) only. We fine-tuned these models in a supervised manner for binary classification into fact claiming comments and non fact claiming comments. Since we employ two multilingual models, we chose to fine-tune one of those (mBERT) on the GermEval 2021 data and an additional dataset. In comparison, we fine-tuned our second multilingual model (XLM-R) and the German GottBERT model using only the training data provided by the GermEval 2021 shared task.

The applied method is derived from our approach (Schütz et al., 2021a) presented in the EXIST 2021 challenge. The first shared task on sEXism Identification in Social neTworks (EXIST) at IberLEF 2021 (Rodríguez-Sánchez et al., 2021; Montes et al., 2021), covering a wide spectrum of sexist content and aims to differentiate different types of sexist content. In our EXIST 2021 contribution a comparable set of transformer models and processing steps were applied (Schütz et al., 2021a).

2.1 GermEval 2021 Data & Preprocessing

The data provided by the organizers of GermEval 2021 is an annotated dataset consisting of over 3,244 German Facebook comments on a political talk show of a German television broadcaster and user discussions from February to July 2019. The dataset was annotated and standardized. Links to users were anonymized with @USER, links to the show were replaced with @MEDIUM and links to the moderator were replaced with @MODERATOR. The original dataset was provided in CSV format. A subset of user comments from two shows were used for the train data. The comments in the test data were drawn from other shows. The dataset contained 1,103 (34%) instances which were labeled as *fact claiming* and 2,141 (66%) instances without any fact claims. The provided dataset is described in more detail in the GermEval 2021 overview paper (Risch et al., 2021).

In initial experiments, we applied different preprocessing strategies to the dataset. We tested our models on a processed version where all links in the dataset were replaced with @MEDIUM, since not every link was connected to the show that was the source of the data. Similarly, as an additional step for our multilingual models, we replaced all emojis with their English translations[1]. However, the two preprocessing steps had a slightly negative impact of 1% on average for mBERT, while they had a clearly positive impact of 3% for XLM-R. Therefore, we used the preprocessed training data only for the XLM-R model. Similarly, the replacement of links did not have a positive influence for the monolingual GottBERT model, where we also used the unpreprocessed comments as an input for training.

We did not use conventional preprocessing steps, e.g., stop-word removal, lemmatization, or stemming, because transform models do not need these due to their ability to capture more context in their word embeddings through improved pretraining capabilities and multi-head attention mechanisms (Vaswani et al., 2017; Devlin et al., 2019a; Liu et al., 2019).

2.2 External Data

As external data we use the ClaimBuster (Arslan et al., 2020) dataset, which consists of English statements from all U.S. presidential debates from 1960-2016. The original part of this dataset consists of 23,533 records. In total, 32,072 sentences were spoken in these debates. The presidential candidates spoke 26,322 sentences, debate moderators spoke 4,292 sentences and 1,319 sentences were spoken by the questioners. Sentences from

[1] https://pypi.org/project/emoji/

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

the moderators and the questioners were discarded and only the sentences spoken by the presidential candidates were considered for creating the ClaimBuster dataset. Moreover, sentences shorter then 5 words were also removed (2,789 sentences). The resulting dataset (*crowdsourced.csv*) was annotated by recruited participants (mostly university students). In addition, three experts labeled a subset of this dataset containing 1,032 sentences to create a groundtruth dataset (*groundtruth.csv*). The provided ClaimBuster dataset consists of three CSV files:

- *all_sentences.csv* (32,072 sentences): all sentences of the debates
- *crowdsourced.csv* (23,533 sentences): sentences of presidential candidates longer than 5 words, labeled by recruited participants
- *groundtruth.csv* (1,032 sentences): sentences of presidential candidates longer than 5 words, labeled by experts

For the GermEval 2021 challenge we used only the *groundtruth.csv* file to ensure high-quality data. The records in the file are annotated as follows:

- non-factual statement (NFS)
- unimportant factual statement (UFS)
- check-worthy factual statement (CFS)

Referring to the original paper (Arslan et al., 2020), the dataset is imbalanced in terms of class distribution: 23.87% belong to CFS, 10.45% to UFS and 65.68% to NFS. The instances (sentences) are annotated as numerical categories ("-1", "0", "1"). In order to match the ClaimBuster data with the original GermEval 2021 data, it was necessary to get an overview of the sentences first and afterwards match the labels to a unified format. Therefore, the comments with the labels "0" and "-1" have been mapped to "0" (not claiming). The instances labeled as "1" were not changed and thus assigned to the class of fact claiming comments. In a next step, we translated the whole dataset into German using the Google Translator API. The translation of the dataset was only used for the mBERT model, since in former work (Schütz et al., 2021a) it was shown that using additional data for this exact model can improve the predictions on a similar NLP downstream task.

3 Models

In total we used three different architectures, which are all based on the original transformer (Vaswani et al., 2017) model:

mBERT is a multilingual transformer based on the original structure of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019a). However, BERT was only trained on English data in comparison to the multilingual model which was additionally trained on Wikipedia data in 100 languages (Devlin et al., 2019b). BERT in general consists - unlike the original transformer with its encoder / decoder architecture (Vaswani et al., 2017) - only of an encoder and is pre-trained using two different strategies: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) (Devlin et al., 2019a). MLM masks words with a specific pattern in a sequence that the model has to predict using its bidirectionality and multi-headed attention (reading a sentence from left-to-right and right-to-left). NSP is the task of predicting the following sentence in the text input (Devlin et al., 2019a).

GottBERT: is a monolingual transformer model, which is based on RoBERTa (Robustly Optimized BERT Pretraining Approach) (Liu et al., 2019). The latter used the BERT architecture, but was trained with more data over a longer time period. Additionally, NSP was not used for pre-training the model and MLM was changed from static to dynamic, where they use a different mask pattern for every sequence during training instead of the same as in BERT. RoBERTa outperforms BERT in several NLP downstream tasks (Liu et al., 2019). Since the original RoBERTa model was only trained on English data, GottBERT was trained from scratch, with the same parameters as the German BERT version, on the German data of the OSCAR corpus (Scheible et al., 2020).

XLM-R: is a self-supervised cross-lingual model that was - similarly as mBERT - trained with monolingual CommonCrawl data in 100 languages (Conneau et al., 2019). The architecture is based on RoBERTa (similarly as GottBERT) in combination with the multilingual XLM transformer (Conneau and Lample, 2019). XLM uses more language modeling approaches (Conneau and Lample, 2019) than RoBERTa and is

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

only trained monolingually with MLM (Conneau et al., 2019). XLM-R outperforms mBERT on multiple tasks (Conneau et al., 2019). Evaluation results showed that XLM-R especially works well for languages with less available data in comparison to other models (Conneau et al., 2019).

The three models do not only differ in the number of languages that they were trained on: BERT and RoBERTa have different pre-training strategies, whereas the strategy of RoBERTa are used by GottBERT as well as XLM-R. As more training data is used, the vocabulary increases, resulting in longer pre-training and fine-tuning intervals. This usually has a positive influence on the performance of downstream tasks.

4 Experimental Setup

Figure 1 provides an overview of our experimental setup and the training strategies used to solve the fact claiming task. The two main approaches take two distinct parts of input data, i.e., only GermEval 2021 data or in addition ClaimBuster data as input. To evaluate the proposed methods we performed experiments by utilizing the following pre-trained transformer models provided by the HuggingFace (Wolf et al., 2020) library: mBERT (Devlin et al., 2019a), Gottbert[2], and XLM-R[3] (Conneau et al., 2019). The experimental setup for the three models is described in detail below.

4.1 mBERT

The cased multilingual BERT transformer (Devlin et al., 2019a) was fine-tuned on the original GermEval 2021 data as well as the additional English ClaimBuster data (Arslan et al., 2020) and its German translations. Note that since the model is multilingual, we expect the English ClaimBuster data to have a positive impact on model training. Both datasets were not subject to any further preprocessing. The model was trained for 4 epochs with a learning rate of 1e-5, batch size of 8 and a maximum sequence length of 284.

4.2 GottBERT

We fine-tuned the German RoBERTa model (GottBERT) (Scheible et al., 2020) using the GermEval 2021 data without any additional preprocessing. We trained the model for 4 epochs with a learning rate of 3e-5 and one more epoch with a learning rate of 1e-5, with a batch size of 8, a maximum sequence length of 128, weight decay of 0.01 and 500 warm-up steps.

[2] https://huggingface.co/uklfr/gottbert-base

[3] https://huggingface.co/xlm-roberta-base

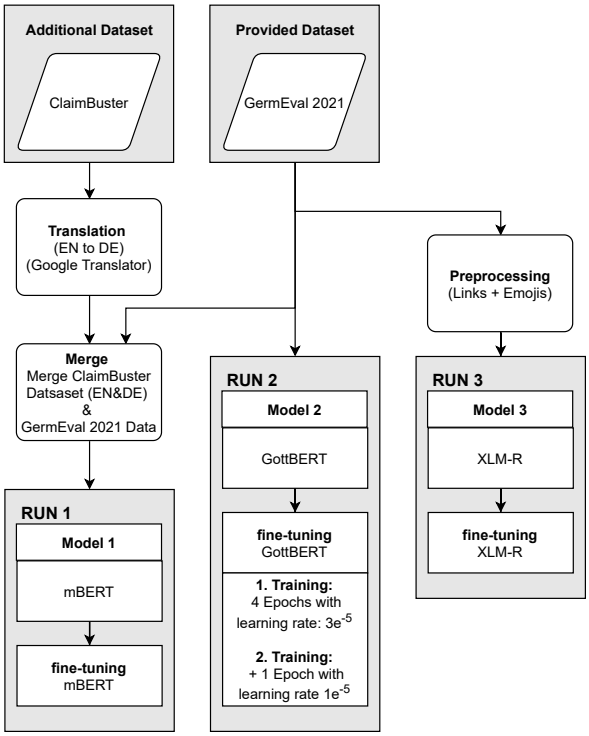

Figure 1: Overview of the setup of our submitted runs including the used models and data.

4.3 XLM-R

The XLM-R (Conneau et al., 2019) model was trained on the preprocessed (replacing links with @MEDIUM and replacing emojis with their translated text) GermEval 2021 training data. We fine-tuned the model for 10 epochs, with a batch size of 16, a maximum sequence length of 256, a learning rate of 2e-5 without warm-up steps and Adam as an optimizer.

5 Results

The validation and test results in terms of precision, recall, and macro-averaged F1-score are presented in Table 1.

Run 1: The mBERT seems to generalize well, as the F1-score on the test set of 72.84% is at a similar performance level as on the validation set (76.09%). This result on the test set is the second highest achieved in our experiments.

Model	Run	P (val)	R (val)	F1 (val)	P (test)	R (test)	F1 (test)
mBERT	1	77.52	74.72	76.09	72.71	72.96	72.84
GottBERT	2	74.26	68.80	78.90	**74.13**	**75.11**	**74.62**
XLM-R	3	76.91	78.11	76.73	68.45	70.11	69.27

Table 1: Accuracy (A), precision (P), recall (R), and macro-averaged F1-scores (F1) for the GermEval 2021. Abbreviation "val" stands for our validation set and "test" for the official benchmark test set. The performance measures are expressed in percent (%).

Run 2: The fine-tuned GottBERT on the GermEval 2021 data achieved an overall F1-score of 74.62% on the test set (78.90% on the validation set). These results speak for the generalization ability of this network because the test performance is at a similar performance level as on the validation set. This result is the highest obtained for all our models.

Run 3: The XLM-R fine-tuned on the GermEval 2021 data achieved the lowest F1-score on the test set (69.27%). This approach seems to exhibit a strong overfitting behavior, as the results on the validation set are considerably higher (F1-score of 76.73%).

In conclusion, the GottBERT model (run 2) achieves the highest results in our experiments. These results indicate that the model that is pretrained on German data allows for a better modeling of the semantics of the task than a multilingual model. All other models are also beyond the zero-rule baseline which is at 66% for the test set.

A more detailed analysis of the results shows that all three models consistently predicted the same class in 560 cases (corresponds to approx. 60% of the test set). In the following, two examples are given for both classes:

Example 1 "@USER Sie würden wahrscheinlich auch einen Kriegstreiber/in wählen, wenn es gegen Trump ginge, warten sie es ab , vielleicht geht ihr Wunsch ja in Erfüllung...."
The ground truth and predictions of all models for this example are "0" (not fact claiming).

Example 2 "@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"
The ground truth and predictions of all models for this example are "1" (fact claiming).

Furthermore, all three models consistently predicted the wrong class in 107 cases (corresponds to approx. 11% of the test set). In the following, two examples are given for both classes:

Example 1 "Hackt nicht nimmer auf den Fussball rum. Bei allem Sportarten sind wieder Zuschauer erlaubt. Hygienekonzept vorausgesetzt."
The ground truth is "1" (fact claiming) and predictions of all models are "0" (not fact claiming).

Example 2 "Biden gewinnt, Corona wird weggehen, Amerika wird reich,k alle bekommen ARbeit und die Welt wird schön. Also was sollst."
The ground truth is "0" (not fact claiming) and predictions of all models are "1" (fact claiming).

In the remaining 273 cases (corresponds to approx. 29% of the test set), one of the models did not predict the same as the others. mBERT and GottBERT predicted equally in 100 cases (70 correctly and 30 incorrectly). GottBERT and XLM-R predicted equally in 100 cases (47 correctly and 53 incorrectly). mBERT and XLM-R predicted equally in 73 cases (28 correctly and 45 incorrectly). These results show that even though both pairs mBERT and GottBERT on one side and mBERT and XLM-R on the other side predict equally in most cases (100), mBERT and GottBERT predict correctly in significantly more cases (70).

6 Conclusion & Future Work

In this paper, we described our submission to the "Fact-Claiming Comment Classification" task of the GermEval 2021. In our experiments GottBERT, a transformer-based machine learning model pretrained on German data only, achieved the best results, leading to an F1-score of 74.62% on the test set. For the multilingual transformer models, we obtained better results with mBERT (potentially because it was trained with an additional dataset) than with XLM-R, which seems to have slightly overfitted on the training data.

Future work will focus on evaluating the different models and approaches in more detail and to investigate how they specifically adapt to the underlying data. We will further investigate how the use of external data impacts the performance of all three investigated models, especially GottBERT, which seem to be the most promising option. Due to the similarity of the presented approaches in this challenge and our previous submission to the EXIST 2021 challenge, see (Schütz et al., 2021a), we plan to perform comparisons on how the applied models converge with respect to the different datasets, semantic concepts and downstream tasks addressed in the benchmarks. Furthermore, we will analyze whether the findings from this comparison can be applied to related tasks such as disinformation detection (Schütz et al., 2021b).

7 Acknowledgements

This contribution has been funded by the FFG Project "Defalsif-AI" (Austrian security research programme KIRAS of the Federal Ministry of Agriculture, Regions and Tourism(BMLRT), grant no. 879670) and the FFG Project "Big Data Analytics" (grant no. 866880).

References

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. Claimbuster: A benchmark dataset of check-worthy factual claims.

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. *CoRR*, abs/1911.02116.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Multilingual BERT (mBERT). Accessed: 2010-06-02.

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

Manuel Montes, Paolo Rosso, Julio Gonzalo, Ezra Aragón, Rodrigo Agerri, Miguel Ángel Álvarez Carmona, Elena Álvarez Mellado, Jorge Carrillo de Albornoz, Luis Chiruzzo, Larissa Freitas, Helena Gómez Adorno, Yoan Gutiérrez, Salud María Jiménez Zafra, Salvador Lima, Flor Miriam Plaza de Arco, and Mariona Taulé (eds.). 2021. Proceedings of the iberian languages evaluation forum (iberlef 2021). In *CEUR Workshop Proceedings*.

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 SharedTask on the Identification of Toxic, Engaging, and Fact-Claiming Comments co-located with KONVENS*, pages 1–12.

Francisco Rodríguez-Sánchez, Jorge Carrillo de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67(0).

Raphael Scheible, Fabian Thomczyk, Patric Tippmann, Victor Jaravine, and Martin Boeker. 2020. GottBERT: a pure german language model. *CoRR*, abs/2012.02110.

Mina Schütz, Jaqueline Boeck, Daria Liakhovets, Djordje Slijepčević, Armin Kirchknopf, Manuel Hecht, Johannes Bogensperger, Sven Schlarb, Alexander Schindler, and Matthias Zeppelzauer. 2021a. Automatic sexism detection with multilingual transformer models. *arXiv preprint arXiv:2106.04908*.

Mina Schütz, Alexander Schindler, Melanie Siegel, and Kawa Nazemi. 2021b. Automatic fake news detection with pre-trained transformer models. In *Pattern Recognition. ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Sciences*, volume 12667, Cham. Springer.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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.