HunterSpeechLab at GermEval 2021: Does Your Comment Claim A Fact? Contextualized Embeddings for German Fact-Claiming Comment Classification

Subhadarshi Panda
Hunter College
City University of New York
spanda@gc.cuny.edu

Sarah Ita Levitan
Hunter College
City University of New York
sarah.levitan@hunter.cuny.edu

Abstract

In this paper we investigate the efficacy of using contextual embeddings from multilingual BERT and German BERT in identifying fact-claiming comments in German on social media. Additionally, we examine the impact of formulating the classification problem as a multi-task learning problem, where the model identifies toxicity and engagement of the comment in addition to identifying whether it is fact-claiming. We provide a thorough comparison of the two BERT based models compared with a logistic regression baseline and show that German BERT features trained using a multi-task objective achieves the best F1 score on the test set. This work was done as part of a submission to GermEval 2021 shared task on the identification of fact-claiming comments.[1]

1 Introduction

Deception and misinformation have become increasingly common on social media (Ciampaglia, 2018). With the endless cycle of production and consumption of information, social media users may be influenced to make choices which are unsafe for them and for society. As a result, misinformation media has attracted the interest of the NLP community to develop methods to combat it. Since information on social media is often not filtered based on veracity, it is necessary to apply fact-checking in order to verify the claims made in social media posts and comments (Thorne and Vlachos, 2018). In this paper, we address the problem of automatic fact-claiming comment detection in German, which is a crucial aspect of the fact checking process.

Fact check-worthiness detection has been studied in prior work by using Naive Bayes and SVM classifiers to identify fact check worthy statements in political debates (Hassan et al., 2015). Gencheva et al. (2017); Patwari et al. (2017) find that it is useful to use the context of a statement to detect fact worthiness. Vasileva et al. (2019) frame the problem as a multi-task prediction problem where the model predicts the labels for multiple fact worthiness annotation sources.

In this work we use contextualized embeddings from multilingual BERT and German BERT for detecting fact-claiming comments. We also train our model on two auxiliary tasks of toxicity and engagement prediction in a multi-task learning setup. Upon comparing the methods, we find that the multitask model based on German BERT achieves the best macro-average F-1 score on the test set.

We outline the remaining part of the paper by first describing the dataset used for our experiments in Section 2. Then we discuss the approaches we use in Section 3. Sections 4 and 5 contain details of the experiment setup and the results obtained. Section 6 concludes the paper.

2 Data

We use the data provided for the GermEval 2021 shared task (Risch et al., 2021). The sizes of the training and test splits along with the number of samples in each class are shown in Table 1. Since development data was not provided, we use 5-fold cross validation by splitting the training data into 5 folds using stratified splitting. We then train our models 5 times by picking 4 folds for training and the remaining fold for development. To report evaluation results, we use the mean and standard deviation of the scores on the 5 folds.

3 Methodology

The task is to classify a given comment as fact-claiming or not. It is a binary classification task.

[1] Code is available at https://github.com/subhadarship/GermEval2021.

Data split	Fact-claiming	Non fact-claiming	Total
Train	1103	2141	3244
Test	314	630	944

Table 1: Data statistics.

3.1 Baseline

We first build a linear logistic regression model as our baseline. The comments are converted to TF-IDF feature vectors with unigram word features.

3.2 BERT-based models

BERT (Devlin et al., 2019) is a large language model trained in a self-supervised task of masked language modeling. It was trained using a huge amount of text data from BookCorpus (Zhu et al., 2015) and Wikipedia. It has been shown that fine-tuning pre-trained BERT or using it as a feature extractor leads to improvements in performance in a wide range of downstream NLP tasks (Devlin et al., 2019). In this work which involves data in German, we use multilingual BERT (Devlin et al., 2019) and German BERT (Chan et al., 2020).[2] For both the models, we add a binary classification head on top for classification of comments to whether they are fact-claiming or not. The input to the classification head is the [CLS] representation. We minimize the cross-entropy loss during training.

Multilingual BERT Multilingual BERT (mBERT) is a single big language model trained using unsupervised corpora in 104 languages including German. Although not all of the 104 languages are represented with equal quality in mBERT (Pires et al., 2019), mBERT has been shown to perform well on a number of non-English languages (Devlin et al., 2019).

German BERT German BERT is a BERT model trained using a huge amount of German data which includes Wikipedia dump (6 GB), OpenLegalData (2.4 GB) and news articles (3.6 GB). It has been shown to achieve impressive results on German sentiment analysis, document classification and named entity recognition.

3.3 Multi-task training

Multi-task training has been successful in a wide range of NLP tasks where the model predicts the labels for multiple tasks for a given input (Chauhan et al., 2020; Barnes et al., 2021; Goo et al., 2018). We also employ multitask learning for our problem where instead of only predicting whether a comment is fact-claiming or not, the model also predicts whether the comment is toxic and engaging. The idea is that the model learns a representation that is useful for all three of these tasks, which may lead to better overall performance at the primary task which is fact-claiming classification. To do this we add three binary classification heads, one each for fact-claiming, toxicity and engagement prediction, on top of the base BERT model. The training is done with a multi-task objective where the total loss is computed as the sum of the cross entropy losses for the three subtasks.

4 Experiments

In this section we outline the baseline experiments using logistic regression model and experiments using BERT-based models.

4.1 Logistic regression

We used the TF-IDF feature extractor and the logistic regression model implemented in scikit-learn. L2 regularization was applied to the model parameters. The coefficient which specifies the inverse of the regularization strength (C parameter) was tuned across five values $\{1, 2, 3, 4, 5\}$. For each hyperparameter setting, 5 systems were trained corresponding to the 5 folds, where in each run 4 folds were used for training and the held out fold was used for evaluation.

4.2 BERT-based models

We used the bert-base-multilingual-cased and bert-base-german-cased identifiers in Huggingface transformers library (Wolf et al., 2020) for loading the models and tokenizers of pretrained mBERT and German BERT respectively. Both the BERT based models were trained by adding a linear classification layer on top, the hidden size of which was tuned across the values 128, 256, and 512. Optimizer used was Adam with a learning rate value tuned in $\{0.05, 0.005, 0.0005\}$. Gradients greater than 1 were clipped during training. The parameters of the

[2] https://www.deepset.ai/german-bert

Hyperparameter (C)	Train F-1	Dev F-1
1.0	0.800 ± 0.00	0.688 ± 0.00
2.0	0.876 ± 0.00	0.701 ± 0.01
3.0	0.919 ± 0.00	0.709 ± 0.01
4.0	0.946 ± 0.00	0.707 ± 0.01
5.0	**0.964** ± 0.00	**0.711** ± 0.01

Table 2: Baseline results using logistic regression model with TF-IDF features.

Model	F-1	Precision	Recall
Monotask mBERT	0.737 ± 0.01	0.745 ± 0.01	0.735 ± 0.01
Monotask German BERT	**0.762** ± 0.01	**0.774** ± 0.02	**0.754** ± 0.01
Multitask mBERT	0.737 ± 0.01	0.742 ± 0.01	0.734 ± 0.01
Multitask German BERT	0.759 ± 0.01	**0.774** ± 0.02	0.751 ± 0.01

Table 3: BERT based model results on dev data.

BERT layers were either frozen or were fine-tuned during the training process. We found that freezing the parameters of the BERT layers resulted in better scores consistently. Training was stopped when the development macro average F-1 score did not improve for 10 consecutive epochs. Similar to the logistic regression model training, 5 systems were trained for each hyperparameter setting using the 5 fold cross-validation data.

We found that for both mBERT and German BERT, less than 0.5% of the running tokens were out-of-vocabulary. However the best system for mBERT required 34 epochs to converge whereas the best system for German BERT required only 21 epochs to converge, where the best systems were decided based on the average cross-validation macro-average F-1 score. There was almost no difference in training time in epochs for monotask vs multitask best systems of German BERT.

5 Results

The predictions were evaluated using the macro-average F1 score. For the dev set results, the macro-average F-1 score is computed using the implementation provided in `scikit-learn` whereas for test set results the macro-average F-1 score is computed using the evaluation script provided by the organizers of the shared task.[3] The difference between the two is that while the former computes the arithmetic mean of the F-1 scores for each class, the latter computes the arithmetic mean of the precisions and the arithmetic mean of the recalls for each class which are then used to compute F-1 score using $F1 = \frac{2 \times P \times R}{P+R}$.[4]

5.1 Results on dev data

The mean and standard deviation of the F-1 scores for each hyperparameter setting of the logistic regression baseline model are shown in Table 2. The best training and development F-1 scores are obtained using $C = 5.0$.

The BERT-based results using monotask training and multitask training are shown in Table 3. All the results shown are for the case where the BERT layers were frozen. The best set of hyperparameters are hidden size 512 and learning rate 0.0005.

5.2 Results on test data

For a given model, we only used one of the 5 cross-validation trained systems to predict the labels of the test set. The test set results are shown in Table 4. Notably the logistic regression's F-1 and precision scores are higher than monotask mBERT scores. The German BERT scores are higher than both the logistic regression baseline and monotask mBERT. The overall best score is obtained using the multitask German BERT which achives 71.5% F-1 score, 72.72% precision and 70.32% recall.

5.3 Analysis of results

We analyze the test results using the confusion matrices of the three submitted BERT-based systems (see Figure 1). All the three systems have very close true negatives (149, 148 and 143). However, the multitask German BERT has the lowest false positives (87) as compared to monotask mBERT (107) and monotask German BERT (95).

[3] Link here.

[4] The two approaches result in almost identical scores.

System	F-1	Precision	Recall
Logistic regression	0.6873	0.7013	0.6738
Monotask mBERT	0.6851	0.6924	0.6778
Monotask German BERT	0.6991	0.7097	0.6889
Multitask German BERT	**0.7150**	**0.7272**	**0.7032**

Table 4: Results on the test set. Only the 3 BERT based systems were submitted for shared task evaluation.

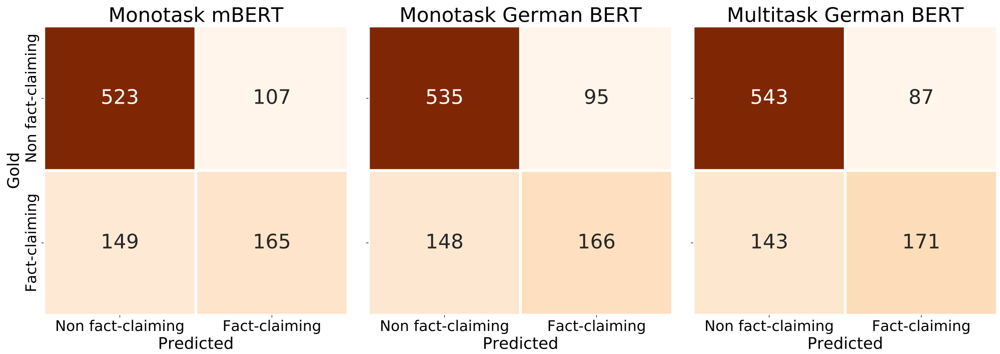

Figure 1: Confusion matrix on the test set using different BERT based systems.

6 Conclusion

In this paper we presented a baseline model based on logistic regression and two BERT-based models for identifying whether German comments in social media are fact-claiming or not. We thoroughly compared the two models: multilingual BERT which is pre-trained on 104 languages, and German BERT which is pre-trained only on German data. We also formulated the learning problem as a multitask problem by addition of two auxiliary tasks of toxicity and engagement classification. The multitask German BERT achieved the best results on the test set. This work contributes models and insights for detecting fact-claiming comments on social media, which is an important step towards hopefully combating misinformation that is pervasive on social media.

References

Jeremy Barnes, Erik Velldal, and Lilja Øvrelid. 2021. Improving sentiment analysis with multi-task learning of negation. *Natural Language Engineering*, 27(2):249–269.

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.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360, Online. Association for Computational Linguistics.

Giovanni Luca Ciampaglia. 2018. *The Digital Misinformation Pipeline*, pages 413–421. Springer Fachmedien Wiesbaden, Wiesbaden.

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.

Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 267–276, Varna, Bulgaria. INCOMA Ltd.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757, New Orleans, Louisiana. Association for Computational Linguistics.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1835–1838, New York, NY, USA. Association for Computing Machinery.

Ayush Patwari, Dan Goldwasser, and Saurabh Bagchi. 2017. *TATHYA: A Multi-Classifier System for Detecting Check-Worthy Statements in Political Debates*, page 2259–2262. Association for Computing Machinery, New York, NY, USA.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Slavena Vasileva, Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. It takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1229–1239, Varna, Bulgaria. INCOMA Ltd.

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.

Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.