German to English: Fake News Detection with Machine Translation

Jin Liu[1], Steffen Thoma[1]

[1]FZI Research Center for Information Technology, Information Process Engineering, Haid-und-Neu-Str. 10-14, 76131 Karlsruhe, Germany

Abstract

Fighting the spreading of fake news is one of the most challenging tasks on the Internet. In this paper, we experiment with various pre-trained language models (PLMs) to check the veracity of German news articles. Since there are very few PLMs for German, we translate the German benchmark dataset to English. Then, we conduct experiments with both German and translated English datasets for predicting the veracity of the news articles. In the experiments, we apply the fine-tuning and adapter methods based on corresponding PLMs. Our experiments on the FANG-COVID dataset show that the errors resulting from translating German to English can be compensated by the accuracy of available PLMs in English. With these experiments, we show that translating a dataset from a lower-resource language to English is a valid intermediate step for further processing with English PLMs.

Keywords

Fake News Detection, Pre-trained Language Model, Machine Translation

1. Introduction

Since the US election in 2016, the spread of fake news has caused more and more concern in the public [1]. There are various motives behind the spread of fake news, e.g. political and financial. In academia, there has been an increasing interest in fighting fake news with machine learning and natural language processing (NLP) based methods. There are different kinds of definitions for fake news. Following Shu et al. [2], fake news is a news article that is intentionally and verifiably false. For fighting fake news, there are many public training datasets, most of which have been crawled from the Internet. A large part of the popular datasets is in English, e.g. BuzzFeed News [3], Fake News Challenge dataset [4], FEVER [5], LIAR [6], etc. Currently, pre-trained language models (PLMs) are standard tools for NLP tasks. English datasets are also the main material for training these models. Therefore, there are obvious advantages regarding the amount of pre-trained language models and relevant datasets to detecting fake news in English.

Due to the aforementioned facts, we come up with an intuitive way for detecting fake news in German and thereby show that the translation to English is a reasonable way for

Proceedings of the Workshop on Text Mining and Generation (TMG) co-located with the 45rd German Conference on Artificial Intelligence (KI 2022), September 19, 2022, Trier (Virtual), Germany

jin.liu@fzi.de (J. Liu); thoma@fzi.de (S. Thoma)
https://www.fzi.de/team/jin-liu/ (J. Liu); https://www.fzi.de/team/steffen-thoma/ (S. Thoma)
0000-0003-2157-4161 (J. Liu); 0000-0003-1205-0630 (S. Thoma)

fake news detection. Concretely, we translate a German fake news dataset to English with a transformer-based machine translator. We then fine-tune or adapt Bidirectional Encoder Representations from Transformers (BERT) [7] and Robustly Optimized BERT Pretraining Approach (RoBERTa) [8], which are mainly trained on English datasets, to predict the veracity of news articles. For comparison, we directly fine-tune or adapt the German version of the BERT model with the German fake news dataset for further prediction. Based on the results of the experiments with indirect and direct methods, we show the feasibility of the indirect method via translation for fake news detection.

The rest of the paper is organized as follows: We briefly review related papers for fake news detection, especially on computational methods, in Section 2. In Section 3, we describe the original German open-source dataset and the translation process. In Section 4, we explain our models for the experiments in detail. In Section 5, we report the results of various experiments. Finally, we conclude and give an outlook for future work in Section 6.

2. Related Work

Fake news detection has drawn increasing attention from the research community. There is fake news in various forms of news, text, image, video, and multimedia. We restrict the literature review to text-based fake news detection. The models for fake news detection can be based on the news contents and social contexts [2]. The latter is often applied to combat fake news on social media platforms. According to Shu et al. [2], the content-based models for checking the facts in the news can be classified into three groups, namely expert-oriented, crowdsourcing-oriented, and computational-oriented. The computational-oriented models are our focus in the review.

Before the introduction of large PLMs, most computation-oriented models applied supervised learning in the form of classification with the construction of various features. Zhou et al. [1] classified the features into 5 groups, namely lexicon, syntax, semantic, discourse, and combination. The traditional machine learning models for detecting fake news include Naive Bayes [9, 10], Support Vector Machine [9, 11], Random Forest [10], and so on. These methods achieve high accuracy for verifying the news. Meanwhile, they need complex feature engineering, that demands domain expertise.

Transformer-based PLMs have achieved SOTA performance for many NLP tasks. Fine-tuning is the standard method for downstream tasks with PLMs, which updates all parameters in the PLM for the specific downstream task [7]. They have also been applied for fake news detection. Due to the extensive training corpora, transformer models can capture the linguistic features in text well. Sepúlveda-Torres et al. [12] achieved SOTA performance on the dataset of fake news challenge with PLMs. Hously et al. [13] proposed adapter modules that add only a few trainable parameters per task and freeze the parameters in the PLMs. Compared to full fine-tuning, the adapter method needs to train many fewer parameters. For fake news detection, the adapter modules are often used to introduce external knowledge into the PLMs for fact verification. Whitehouse et al. [14] evaluated fake news detection with knowledge-enhanced language models which include the K-Adapter model. Factual knowledge obtained from the text triples on Wikipedia and Wikidata, and linguistic knowledge obtained from dependency parsing, are added to the PLMs via the adapter modules [15]. The evaluation in [14] shows that given relevant and up-to-date knowledge bases, knowledge-enhanced models can significantly improve the performance of fake news detection.

Machine translation is a widely applied data augmentation method for NLP tasks [16]. Amjad et al. [17] translated English corpus to Urdu as extra training data for fake news detection. The augmented data has not improved the performance due to the quality of machine translation. De et al. [18] tackled the non-availability of annotated corpora for four low-resource languages by translating English datasets to corresponding languages.

3. Dataset

In this section, we provide an overview of the dataset used for the following experiments. We first describe the original chosen German dataset about COVID-19. Based on the German dataset, we then give an introduction to our machine translator for converting German news articles to English.

3.1. Original Dataset

There are very few German fake news datasets publicly available. We have searched and decided to choose the FANG-COVID dataset provided in [19]. The reason we choose this dataset is that the dataset is well labeled and contains a large number of training examples. The dataset contains in total 41,242 news articles about the COVID-19 pandemic, 28,056 news articles are labeled as real and 13,186 news articles are abeled as fake. The dataset has been crawled from 3 reliable news agencies, including Sueddeutsche Zeitung, Tagesspiegel, ZEIT, and 10 unreliable news agencies, e.g. AnonymousNews, Contra-Magazin, etc. The news articles from the 3 reliable news agencies are labeled as real and from the 10 unreliable news agencies as fake. In addition to labels, headers, and contents of the news articles, the dataset also contains meta-information about URL, date, source, and Twitter history. We only use the text contents of the news articles and the labels for training and prediction. The headers and meta-information are excluded from the experiments.

3.2. Dataset Translation

The original FANG-COVID dataset is in German. We translate all articles to English for further experiments. The dataset has over 40,000 articles and each article has about 48 sentences on average. Considering the progress in neural machine translation with transformer-based models, we implemented our own machine translator. The engine of our translator is the pre-trained opus-mt-de-en, based on Marian-NMT [20]. Each article is separated into sentences with spaCy sentence detector [21]. Afterwards, each sentence is put into the translator to obtain the corresponding English translation.

4. Methodology

In this section, we present our experiments with the FANG-COVID dataset. Based on pre-trained language models we explore two typical methods: fine-tuning and adapter models, for predicting the labels of the news articles in the dataset. All base models are based on the Huggingface transformers library [22].

4.1. Fine-tuning

For the experiments, we choose a German version BERT model, namely bert-base-german-cased, as the base model [23]. For the English version, we try with two base models, namely bert-base-uncased [7] and roberta-base [8]. Since PLMs are mostly trained on English corpora, we use two English base models to reflect the availability of models in English compared to other languages. We then add a binary classification head to the base models for predicting the veracity of the news articles.

We split our dataset into training, validation, and test dataset (64%, 16%, 20%). The content of each news article is tokenized with the corresponding tokenizer provided by the base model. The maximum input length of each base model is limited to 512. The tokens outside the range are dropped. We choose cross-entropy loss as our loss function with

$$\mathcal{L}(p, y) = -y * \log(p) - (1 - y) * \log(1 - p), \tag{1}$$

where y is the target value (0 or 1) and p is the predicted probability. AdamW [24] is used as the optimizer. We have tried different learning rates and choose the learning rate to be 0.00005. With some test runs, we find that within 5 epochs, the fine-tuning method already achieves very good performance regarding accuracy. So, we fine-tune each model with 5 epochs. We then choose the model with the smallest loss on the validation dataset as the final model for the predictions on the test dataset.

4.2. Adapter

As an alternative to the fine-tuning method, we also experiment with the adapter method. The adapter method adds extra layers to the original pre-trained transformer-based models. As in the fine-tuning method, we also add a classification head for predicting the veracity of the news. For the downstream task, the model only updates the added parameters by freezing the parameters of the pre-trained models [13]. With this property, the adapter model can add extra layers for each downstream task without forgetting the learned knowledge in PLMs [25]. Additionally, the adapter methods need to train many fewer parameters compared to the full fine-tuning methods. We have applied the framework of AdapterHub [26] which has been built on top of the Huggingface transformer framework. For the configuration of the adapter, we applied the Pfeiffer version configuration [26]. With the adapter method, we have the same base models, dataset split, optimizer, and loss function. We have selected the learning rate of 0.0001. The adapter models achieve very good performance regarding accuracy on training and validation datasets within 10 epochs. So, we run each adapter model with 10 epochs and choose the model with the smallest loss on the validation dataset for evaluating the test dataset.

Table 1
Performance of fine-tuning and adapter models

Model	Input language	Accuracy	Precision	Recall	F1
Fine-tuning					
bert-base-german-cased	German	0.976	0.981	0.983	0.982
bert-base-uncased	English	0.971	0.979	0.979	0.979
roberta-base	English	**0.980**	**0.983**	**0.988**	**0.985**
Adapter					
bert-base-german-cased	German	0.976	0.978	**0.988**	0.983
bert-base-uncased	English	0.969	0.974	0.980	0.977
roberta-base	English	**0.981**	**0.984**	**0.988**	**0.986**

5. Evaluation

For the evaluation, we have chosen standard accuracy, F1 score, precision, and recall as metrics. To reduce the impact of randomness, we run each model 5 times with different seeds. We report the mean of the performance of the fine-tuning and the adapter method separately in Table 1. The top performance of each group (fine-tuning and adapter) is marked in bold.

Comparing the results across two groups (fine-tuning and adapters), both methods have very similar performance. This confirms the finding in [27] that, for the dataset with a large number of training examples (FANG-COVID over 41k), adapter methods have not shown significant advantages in performance over fine-tuning methods. Within each group (fine-tuning and adapters), the RoBERTa-based models achieve the best performance among the three base models. The performance has proven the improvement of retrained RoBERTa over the original BERT model. The German BERT version has also outperformed the original BERT model in both fine-tuning and adapter methods which has shown good performance in the German domain. The results show that machine translation of the dataset from lower-resource languages to English is a valid intermediate step since more PLMs are available for the specific downstream task.

We give here a brief analysis of the prediction errors by the models. In general (averaged over 5 seeds), the articles labeled as fake have a higher probability of 4.5% being misclassified, compared to the misclassification probability of 1.5% for the articles labeled as real. This can be partly explained by the unbalanced dataset, 68% of the articles are labeled as real and 32% as fake. Fewer training examples of fake articles lead to worse performance in the test datasets. We further apply Jaccard similarity to estimate how similar the prediction errors are. Jaccard similarity is defined as $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$ [28], where $|A \cap B|$ is the number of common prediction errors of models A and B, and $|A \cup B|$ is the number of total prediction errors of models A and B. With pairwise comparison (6 models, 15 pairs) and averaged over 5 seeds, model Fine-tuning bert-base-uncased (English) and Adapter bert-base-uncased (English) have the highest similarity coefficient of 0.269. Model Fine-tuning bert-base-german-cased (German) and Adapter bert-base-german-cased (German) have a similarity coefficient of 0.255. The similarity coefficients of models with different input languages are mostly below 0.15. The models with the same input language have a higher similarity of prediction errors.

6. Conclusion and Outlook

In this paper, we have experimented with an open-source German news dataset to validate the hypothesis that machine translation to English can be a viable intermediate step for fake news detection. Besides original news articles in German, we have translated German news articles to English with a self-implemented neural machine translator. For the original German news articles, we applied a German pre-trained BERT model. For the translated news articles we experimented with the BERT and the RoBERTa model. Based on the base models, we predict the truthfulness of news articles with the fine-tuning and adapter method. The results of the experiments show that the prediction via translation is a competitive method (even slightly better than the direct method) and the errors resulting from translation can be mitigated by the accuracy of available PLMs in English.

We have tackled the fake news detection problem with coarse granularity, namely binary real and fake labels. In future work, we will focus on a taxonomy with more fine-grained labels to also generate explanations for why an article is classified as fake news. This information would be useful to gain a user's trust to use a fake news detection system.

Acknowledgments

This work was carried out with the support of the German Federal Ministry of Education and Research (BMBF) within the project "DeFaktS" (Grant 16KIS1524K).

References

[1] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Comput. Surv. 53 (2020). URL: https://doi.org/10.1145/3395046. doi:10.1145/3395046.

[2] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, SIGKDD Explor. Newsl. 19 (2017) 22–36. URL: https://doi.org/10.1145/3137597.3137600. doi:10.1145/3137597.3137600.

[3] BuzzFeedNews, 2014. URL: https://github.com/BuzzFeedNews.

[4] FakeNewsChallenge, 2017. URL: https://github.com/FakeNewsChallenge/fnc-1.

[5] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, Fever: a large-scale dataset for fact extraction and verification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819. URL: https://aclanthology.org/N18-1074. doi:10.18653/v1/N18-1074.

[6] W. Wang, "Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 422–426. URL: https://aclanthology.org/P17-2067. doi:10.18653/v1/P17-2067.

[7] J. Delvin, M. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding (2018). doi:10.48550/ARXIV.1810.04805.

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach (2019). doi:10.48550/ARXIV.1907.11692.

[9] S. Afroz, M. Brennan, R. Greenstadt, Detecting hoaxes, frauds, and deception in writing style online, in: 2012 IEEE Symposium on Security and Privacy, IEEE Symposium on Security and Privacy, San Francisco, California, USA, 2012.

[10] P. Bharadwaj, Z. Shao, Fake news detection with semantic features and text mining, International Journal on Natural Language Computing (IJNLC) (2019) 17–22.

[11] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, in: Proceeding of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 3391–3401.

[12] R. Sepúlveda-Torres, M. Vicente, E. Saquete, E. Lloret, M. Palomar, Headlinestancechecker: Exploiting summarization to detect headline disinformation, Journal of Web Semantics 71 (2021).

[13] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for nlp, in: Proceedings of the 36th International Conference on Machine Learning, California, USA, 2019, pp. 2790–2799.

[14] C. Whitehouse, T. Weyde, P. Madhyastha, N. Komninos, Evaluation of fake news detection with knowledge-enhanced language models, in: Proceedings of the International AAAI Conference on Web and Social Media, 2022, pp. 1425–1429.

[15] R. Wang, D. Tang, N. Duan, Z. Wei, X. Huang, J. Ji, G. Cao, D. Jiang, M. Zhou, K-adapter: Infusing knowledge into pre-trained models with adapters, Findings of the Association for Computational Linguistics: ACL-IJCNLP, Online, 2021, pp. 1405–1418.

[16] S. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A survey of data augmentation approaches for nlp, Findings of the Assocaition for Computational Linguistics: ACL-IJCNLP, 2021, pp. 968–988.

[17] M. Amjad, G. Sidorov, A. Zhila, Data augmentation using machine transation for fake news detection in the urdu language, in: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC), Marseille, 2020, pp. 2537–2542.

[18] A. De, D. Bandyopahyay, B. Gain, A. Ekbal, A transformer-based approach to multilingual fake news detection in low-resource languages, ACM Transactions on Asian and Low-Resource Language Information Processing (2022).

[19] J. Mattern, Y. Qiao, E. Kerz, D. Wiechmann, M. Strohmaier, Fang-covid: A new large-scale benchmark dataset for fake news detection in German, in: Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Dominican Republic, 2021, pp. 78–91. URL: https://aclanthology.org/2021.fever-1.9. doi:10.18653/v1/2021.fever-1.9.

[20] J. Tiedemann, S. Thottingal, Opus-mt - building open translation services for the world, in: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, Lisboa, Portugal, 2020, pp. 479–480. URL: https://aclanthology.org/2020.eamt-1.61.

[21] A. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spycy: Industrial-strength natural

language processing in python (2020).

[22] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstration, 2020, pp. 38–45.

[23] deepset Homepage, 2019. URL: https://www.deepset.ai/german-bert.

[24] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, Seventh International Conference on Learning Representations (ICLR), 2019.

[25] A. Lauscher, O. Majewska, L. Ribeiro, I. Gurevych, N. Rozanov, G. Glavaš, Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers, in: Proceedings of Deep Learning Inside Out, 2020, pp. 43–49.

[26] J. Pfeiffer, A. Rücklé, C. Poth, A. Kamath, I. Vulić, S. Ruder, K. Cho, I. Gurevych, Adapterhub: A framework for adapting transformers, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 46–54. URL: https://aclanthology.org/2020.emnlp-demos.7. doi:10.18653/v1/2020.emnlp-demos.7.

[27] R. He, L. Liu, H. Ye, Q. Tan, B. Ding, L. Cheng, J. Low, L. Bing, L. Si, On the effectiveness of adapter-based tuning for pretrained language model adaptation, in: Proceedings of the 59 Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2021, pp. 2208–2222.

[28] S. Niwattanakul, J. Singthongchai, E. Naenudorn, S. Wanapu, Using of jaccard coefficient for keywords similarity, in: Proceedings of the International MultiConference of Engineers and Computer Scientist, Hong Kong, 2013.