Detection and Identification of Fake News

Binary Content Classification with Pre-trained Language Models

Mina Schütz

Darmstadt University of Applied Sciences
Haardtring 100, 64295 Darmstadt, Germany and
Austrian Institute of Technology GmbH, Giefinggasse 4, 1210 Vienna, Austria
ORCID-ID: 0000-0002-1102-306X, mina.schuetz@ait.ac.at

Abstract

Fake news has emerged as a critical problem for society and professional journalism. Many individuals consume their news via online media, such as social networks and news websites. Therefore, the demand for automatic fake news detection is increasing. There is still no agreed upon definition for fake news, since it can include various concepts, such as clickbait, propaganda, satire, hoaxes, and rumors. This results in a broad landscape of machine learning approaches, which have a varying accuracy in detecting fake news. This masterthesis focused on a binary content-based classification approach, with a bidirectional Transformer (BERT), to detect fake news in online articles. BERT creates a pre-trained language model during training and is fine-tuned on a labeled dataset. The FakeNewsNet dataset is used to test two variants of the model (cased / uncased) with articles, using only the body text, the title, and a concatenation of both. Additionally, both models were tested with different preprocessing steps. The models gain in all 29 carried out experiments high accuracy results, without overfitting. Using the body text and the concatenation resulted in five models with an accuracy of 87% after testing, whereas using only titles resulted in 84%. This shows that short statements could be already enough for fake news detection using language models. Also, the preprocessing steps seem to have no major impact on the predictions. It is concluded that transformer models, such as BERT, are a promising approach to detect fake news, since it achieves notable results, even without using a large dataset.

Keywords: fake news; fake news detection; BERT; transformer; pre-trained language model; binary classification

1 Introduction

Fake news has recently gained a lot of attention in the media and research community. Through the information overload in the Internet and an increased usage of social media for news consumption the propagation of disinformation has risen (Figueira & Oliveira, 2017). Zhou and Zafarani (2018) even state that fake news is "[…] one of the greatest threats to democracy, journalism, and freedom of expression" (p. 1). This is also based on several psychological reasons, such as echo chambers (Rana et al., 2018). Each individual has a social circle, which typically has the same beliefs as them. Those views are propagated with other people, who share their ideology and therefore the facts seem more credible (Shu et al., 2017a). Individuals are more likely to believe facts the more often they have read it (validity effect) and tend to believe facts that confirm their views. A manual method to solve this problem is expert-based fact-checking, which is expensive and slow (Graves, 2018). Another approach is the automatic detection of fake news with machine learning methods and natural language processing (NLP), which could help users to identify a possible deception in a claim (Mahid et al., 2018).

2 State-of-the-art

Since there is no common definition for fake news to-date, we propose a new definition, which includes all concepts (disinformation, misinformation, hoax, propaganda, rumor, clickbait, satire) that could be categorized as fake news, regardless of the intention or factuality of the article: *Fake news is an article which propagates a distorted view of the real world regardless of the intention behind it.* The content can be examined through the extraction of linguistic features in the body text and titles with NLP methods. This can be on character, word, sentence or document level. Usually Bag-of-Words, N-Gram models, and Part-of-Speech Tagging play an important role. However, many researchers used a variant of models, such as SVM, logistic regression, decision trees or neural networks (convolutional, recurrent, LSTM) to compare the significance of features and their impact on predictions. Since 2018 several surveys were published (Rana et al., 2018; Oshikawa et al., 2018; Zhou & Zafarani, 2018; Sharma et al., 2019), which give an overview of the

various papers and methodologies. Besides content-based, there are also knowledge-based, social-context-based, and hybrid approaches, which can include user-information, network-analysis and stance detection (Shu et al., 2017a; Zhou & Zafarani 2018).

The newly invented transformers have shown to outperform many NLP state-of-the-art results and are usually pre-trained on a large corpus and then fine-tuned on a dataset for the specific task, which makes them very usable for fake news detection, since available corpora are often small. During pre-training transformers create language models, which can be used for sentence- or token-level tasks (Devlin et al., 2019).

BERT (Bidirectional Encoder Representations from Transformers) is one of the latest innovations in machine learning techniques for NLP and was developed by Google in 2019 (ibid.). It is based on the original transformer structure by Vaswani et al. (2017). This transformer "[…] is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers" (ibid., p. 1). Additionally, BERT is the first bidirectional transformer, which means that the model can read a sentence in both directions at the same time, unlike unidirectional transformers (e.g., left-to-right, right-to-left) (Devlin et al., 2019).

3 Main contribution

Text-based fake news detection is still an open research challenge, because the currently proposed solutions are not yet satisfying. These studies use various datasets with no common ground truth and are stand-alone or a mix of the different approaches. Their results are varying to a high degree and often focus on multiple fake news concepts. The purpose of this work is to set out the current state-of-the-art methods in fake news detection and use a pre-trained language model for a binary content classification to gain further knowledge about the textual representation of fake news, especially the usage of a Transformer model for long body text and short titles. Furthermore, the main research question of this work is:

To which extent is a pre-trained language model useful for content-based fake news detection and does it gain promising results in predicting the classification of body texts and titles of news articles?

Therefore, the practical part of the conducted thesis proposes an approach for fake news detection with BERT and the FakeNewsNet (Shu et al., 2017a, 2017b, 2018) dataset. This leads to the following sub-questions:

- How do the results change, if the body text or the title of a news article are used separately, or both are used together?
- Do the results change based on different preprocessing steps, as removing non-ASCII characters, transcripts or short/long articles?
- Does it make a difference, in respect to the preprocessed files, if the cased or uncased version of BERT is used?

4 Related work

Several studies focused on fake news and Transformers but with different tasks: sentence comparison/fact classification (Yang et al., 2019; Mao & Liu, 2019), propaganda detection (Gupta et al., 2019), relationship of body text/ title (Jwa et al., 2019), stance detection (Slovikovskaya, 2019), evidence retrieval (Soleimani et al., 2019). Directly comparable are the works of Liu et al. (2019) and Rodríguez and Iglesias (2019). For a multi-classification for short statements and extra metadata, Liu et al. (2019) gained a result of 40.58% accuracy. However, Rodríguez and Iglesias (2019) also did a binary classification of fake news using three neural networks. They had an accuracy of 98% and used a similar approach as this work, but tested BERT only on one experiment and combined titles and text as one input but did not differentiate between using body text or title and the cased/uncased version.

5 Methodology

In this work we used the BERT implementation by HuggingFace[1], which is based on PyTorch. The FakeNewsNet contains more than 20,000 already labeled (fake/real) articles. It is notable, that the datasets contained more real than fake articles, depending on the pre-processing steps. The content was

1 https://github.com/huggingface/transformers

pre-processed in various ways, which lead to a total of 13 different files (Table 1) with the proposed pre-processing steps and size: deleting articles based on length (removing outliers), deleting transcripts (spoken language in mostly real articles), deleting HTML-tag [edit] (mostly fake articles), deleting non-ASCII symbols and digits. Those were carried out to prevent the model from learning false features based on the classes.

Table 1: Pre-processed dataset files

File no.	Type	Length	Transcript	Edit	ASCII/Digits	Dataset size
1	Text	Yes	Yes	Yes	Yes	11,944
2	Text	Yes	Yes	Yes	No	11,944
3	Text	Yes	Yes	No	No	11,944
4	Text	Yes	No	No	No	12,172
5	Text	No	No	No	No	21,041
6	Title	No	No	No	No	21,041
7	Title	Yes	No	No	No	12,172
8	Title	Yes	No	No	Yes	12,172
9	Both	No	No	No	No	21,041
10	Both	Yes	No	No	No	15,355
11	Both	Yes	Yes	No	No	15,103
12	Both	Yes	Yes	Yes	No	15,103
13	Both	Yes	Yes	Yes	Yes	15,103

Those files were tested in a total of 29 experiments. The hyper parameters for the model were tuned with the first file (completely pre-processed, only body text). The best hyper parameters were then used for the following experiments (text, title, both) with the BERT-base-cased/uncased versions (Fig. 1).

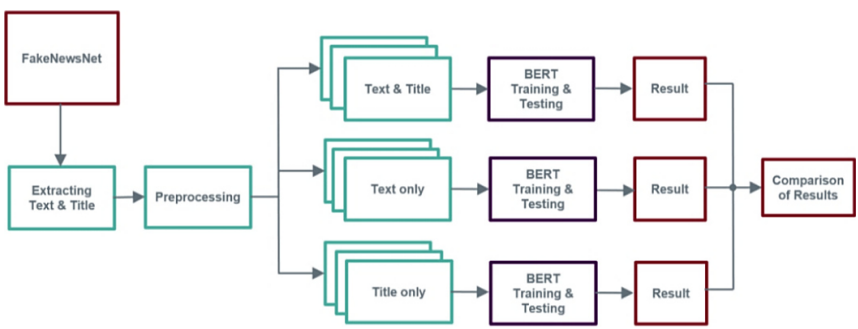

Fig. 1 Methodology overview

Afterward the most promising combinations were tested with the BERT-base-uncased model (Fig. 2). All experiments were evaluated with precision,

recall, accuracy and F1. For each experiment the loss after validation was calculated and compared to the various preprocessed files and both models.

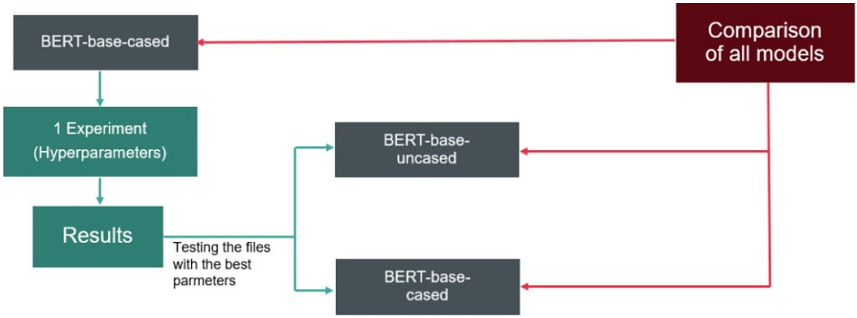

Fig. 2 Workflow for both BERT models

6 Results

The results in Table 2 show that, with the right parameters, both models gain very good detection results. Almost all experiments had results over 80% accuracy in the validation and testing set without overfitting the data. As shown, the learning rate has the highest influence on the results in comparison to other hyper parameters (see experiment 1). Furthermore, even if there is only a small difference in the evaluation metrics, the datasets with a combination of titles and text have the best results (files 9–13). The datasets with the most promising metrics were additionally carried out with the BERT-base-uncased-version (last block). Therefore, we conclude that lowercasing the text has no influence with the used data. Also, in all experiments it is shown that the pre-processing steps have no major impact on the outcome. Although for titles, a larger dataset might be helpful, since file 6 gained the best results with twice the size as other pre-processed data.

Table 2: Results

Model	File	Epochs	Batch	LR	MSL	W-U	ACE	VL	ACT	F1	P	R
B-B-C	1	5	6	5e-5	512	100	0.75	0.60	0.72	0.48	0.43	0.57
B-B-C	1	3	6	2e-5	512	100	0.86	0.30	0.86	0.79	0.81	0.80
B-B-C	1	3	6	1e-5	512	100	0.87	0.30	0.87	0.80	0.82	0.82
B-B-C	1	3	16	2e-5	256	100	0.86	0.23	0.86	0.81	0.83	0.82
B-B-C	1	5	16	5e-5	256	100	0.84	0.07	0.85	0.82	0.81	0.80
B-B-C	1	5	16	2e-5	256	100	0.85	0.09	0.85	0.80	0.83	0.81
B-B-C	1	15	16	2e-5	256	100	0.85	0.01	0.86	0.81	0.84	0.81
B-B-C	1	15	6	2e-5	512	100	0.86	0.02	0.87	0.79	0.81	0.81
B-B-C	1	10	6	2e-5	512	100	0.85	0.02	0.85	0.77	0.79	0.79
B-B-C	1	5	16	2e-5	256	0	0.86	0.09	0.85	0.81	0.83	0.81
B-B-C	1	10	16	2e-5	256	0	0.86	0.02	0.86	0.81	0.84	0.81
B-B-C	1	100	16	2e-5	256	100	0.85	0.01	0.86	0.80	0.84	0.80
B-B-C	2	10	16	2e-5	256	0	0.84	0.02	0.85	0.80	0.81	0.81
B-B-C	3	10	16	2e-5	256	0	0.85	0.02	0.86	0.80	0.82	0.81
B-B-C	4	10	16	2e-5	256	0	0.84	0.02	0.86	0.81	0.83	0.82
B-B-C	5	10	16	2e-5	256	0	0.87	0.08	0.86	0.78	0.82	0.78
B-B-C	6	10	16	2e-5	256	0	0.85	0.08	0.84	0.75	0.78	0.76
B-B-C	6	5	32	3e-5	128	0	0.84	0.11	0.84	0.76	0.79	0.76
B-B-C	7	5	32	2e-5	128	0	0.84	0.11	0.84	0.80	0.80	0.80
B-B-C	8	5	32	2e-5	128	0	0.83	0.10	0.83	0.79	0.80	0.79
B-B-C	9	10	16	2e-5	256	0	0.87	0.07	0.87	0.78	0.82	0.79
B-B-C	10	10	16	2e-5	256	0	0.87	0.08	0.86	0.78	0.82	0.78
B-B-C	11	10	16	2e-5	256	0	0.87	0.09	0.86	0.77	0.83	0.77
B-B-C	12	10	6	2e-5	512	0	0.87	0.10	0.87	0.78	0.80	0.80
B-B-C	13	10	16	2e-5	256	0	0.87	0.08	0.86	0.78	0.86	0.83
B-B-U	1	10	6	2e-5	512	0	0.86	0.02	0.86	0.82	0.84	0.83
B-B-U	9	10	16	2e-5	256	0	0.87	0.07	0.87	0.79	0.84	0.80
B-B-U	6	10	32	2e-5	128	0	0.85	0.08	0.85	0.77	0.85	0.80
B-B-U	6	30	32	2e-5	128	50	0.85	0.06	0.85	0.75	0.85	0.79

(B-B-C: BERT-base-cased, B-B-U: BERT-base-uncased, LR: Learning Rate, MSL: Maximum Sequence Length, W-U: Warm Up, ACE: Accuracy Evaluation, VL = Validation Loss, ACT = Accuracy Testing, R = Recall, P = Precision)

7 Conclusion

The results of this work show that a content-based approach can gain promising results for detecting fake news, even without setting hand-engineered features. We conclude that BERT can be used for short statements as well as complete articles. Out of 29 experiments, five gained an accuracy of 87% on the test set. For only testing the titles of the articles the best accuracy was

84%. This shows that there is only a minimal difference in accuracy using the body text or the title (or both) with pre-trained language models. In future work, especially social-context-based features should be looked upon to gain better results. Another important part is to explore methods of explainable artificial intelligence, to help understanding the difference in fake news concepts, which could help to gain further knowledge about linguistic features. Lastly, this approach should be tested with another dataset to evaluate the cross-domain prediction results.

Acknowledgements

This thesis was conducted at the Darmstadt University of Applied Sciences at the Faculty of Media within the course of studies in Information Science in collaboration with the Austrian Institute of Technology GmbH at the Center for Digital Safety and Security under the supervision of Kawa Nazemi (h_da), Melanie Siegel (h_da), and Alexander Schindler (AIT). Additionally, this work was supported by the Research in Information Science (https://sis.h-da.da/) and the Research Group on Human-Computer Interaction and Visual Analytics (https://vis.h-da.de).

Bibliography

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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* (Vol. 1, pp. 4171–4186). Stroudsburg, PA: Association for Computational Linguistics.

Figueira, A., & Oliveira, L. (2017). The Current State of Fake News: Challenges and Opportunities. *Procedia Computer Science, 121*, 817–825.

Graves, L. (2018). Understanding the promise and limits of automated fact-checking. https://reutersinstitute.politics.ox.ac.uk/our-research/understanding-promise-and-limits-automated-fact-checking

Gupta, P., Saxena, K., Yaseen, U., Runkler, T., & Schütze, H. (2019). Neural Architectures for Fine-Grained Propaganda Detection in News. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda* (pp. 92–97). Association for Computational Linguistics.

Jwa, H., Oh, D., Park, K., Kang, J. M., & Lim, H. (2019). exBAKE: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (BERT). *Applied Sciences 2019, 9*(4062).

Khan, S. A., Alkawaz, M. H., & Zangana, H. M. (2019). The use and abuse of social media for spreading fake news. *IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)* (pp. 145–148). Cham: Springer Nature.

Liu, C., Wu, X., Yu, M., Li, G., Jiang, J., Huang, W., & Lu, X. (2019). A Two-Stage Model Based on BERT for Short Fake News Detection. In Douligeris, C., Karagiannis, D., & Apostolou, D. (Eds.), *Knowledge Science, Engineering and Management* (pp. 172–183). Cham: Springer International Publishing.

Mahid, Z. I., Manickam, S., & Karuppayah, S. (2018). Fake news on social media: Brief review on detection techniques. In *2018 Fourth International Conference on Advances in Computing, Communication Automation (ICACCA)*. [Piscataway, NJ]: IEEE.

Mao, J., & Liu, W. (2019). Factuality Classification Using the Pre-Trained Language Representation Model BERT. In *IberLEF@SEPLN*, pp. 126–131.

Oshikawa, R., Qian, J., & Wang, W. Y. (2018). A Survey on Natural Language Processing for Fake News Detection. *arXiv:1811.00770 (version 1, 2018)*.

Rana, D. P., Agarwal, I., & More, A. (2018). A Review of Techniques to Combat the Peril of Fake News. *4th International Conference on Computing Communication and Automation (ICCCA)*. [Piscataway, NJ]: IEEE.

Rodríguez, À. I., & Iglesias, L. L. (2019). Fake News Detection Using Deep Learning. *arXiv:1910.03496*.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., Liu, Y. (2019). Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 37(4), 111:0–111:41.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2018). FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *CoRR, abs/1809.01286*.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017a). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter, 19*(1), 22–36.

Shu, K., Wang, S., Liu, H. (2017b). Exploiting Tri-Relationship for Fake News Detection. arXiv:1712.07709

Slovikovskaya, V. (2019). Transfer Learning from Transformers to Fake News Challenge Stance Dection (FNC-1) Task. arXiv:1910.14353

Soleimani, A., Monz, C., & Worring, M. (2019). BERT for Evidence Retrieval and Claim Verification. arXiv:1910.02655

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones., L., Gomez, A. N. … Polosukhin, A. (2017). Attention Is All You Need. arXiv:1796.03762

Yang, K.-C., Niven, T., & Kao, H.-Y. (2019). Fake News Detection as Natural Language Interference. arXiv:1907.07347

Zhou, X., & Zafarani, R. (2018). Fake News: A Survey of Research, Detection Methods and Opportunities. *ACM Computing Surveys, 1*(1), Art. No. 109.

In: T. Schmidt, C. Wolff (Eds.): Information between Data and Knowledge. Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021), Regensburg, Germany, 8th–10th March 2021. Glückstadt: Verlag Werner Hülsbusch, pp. 422–431. DOI: doi.org/10.5283/epub.44959.