MeVer team tackling Corona virus and 5G conspiracy using ensemble classification based on BERT

Olga Papadopoulou
Information Technologies Institute, CERTH, Thessaloniki, Greece
olgapapa@iti.gr

Giorgos Kordopatis-Zilos
Information Technologies Institute, CERTH, Thessaloniki, Greece
georgekordopatis@iti.gr

Symeon Papadopoulos
Information Technologies Institute, CERTH, Thessaloniki, Greece
papadop@iti.gr

ABSTRACT

This paper presents the approach developed by the Media Verification (MeVer) team to tackle the task of FakeNews: Coronavirus and 5G conspiracy at the MediaEval 2020 Challenge. We build a two-stage classification approach based on ensemble learning of multiple classification networks. Due to the imbalanced and relatively small dataset, our ensemble method leads to improved performance compared to a single classification model. We fine-tune pre-trained Bidirectional Encoder Representations from Transformers (BERT), one of the most popular transformer models, on the problem of Coronavirus and 5G conspiracy detection. Our approach achieved a score of 0.413 in terms of the Matthews Correlation Coefficient (MCC), which is the official evaluation metric of the task.

1 INTRODUCTION

COVID-19 emerged as a health crisis (pandemic) and soon evolved into an infodemic ('infodemic' refers to an overabundance of information). There are already harmful impacts of COVID-19 Conspiracy theories and specifically around 5G disinformation on society. The incident of the British 5G towers fires because of coronavirus conspiracy theories [14] is a representative example of how important is to detect and prevent the dissemination of such theories. The FakeNews: Coronavirus and 5G conspiracy task is a challenge of MediaEval 2020 that focuses on the analysis of tweets around Coronavirus and 5G conspiracy theories in order to detect misinformation spreaders. For further details on the subtasks and the respective dataset, the reader is referred to [9].

Our approach focuses on ensemble classification in order to overcome the relatively small training dataset and predict more accurately the Coronavirus and 5G conspiracy tweets. In short, a first-level classification is applied using majority voting over nine classifiers to detect conspiracy and non-conspiracy tweets. A second-level classification is then applied to detect the conspiracy tweets related to 5G over the other conspiracy ones. For the training process, we leverage on the pre-trained BERT [1] model and the implementation provided by the HuggingFace library [15][1].

2 RELATED WORK

In case of a pandemic such as that of the Coronavirus, the intentional or unintentional dissemination of manipulated content, conspiracy theories, and propaganda are critical [12]. Several works have been recently published dealing with the detection and verification of COVID-19-related misinformation [2, 3, 10, 11]. Misinformation can be spread in the form of text, images, and videos. Natural language processing (NLP) is a means of dealing with many types of content. For example, the authors of [8] collected a database of debunked and verified user-generated videos and developed a method to detect them using the contextual information surrounding them rather than the video content. The emergence of BERT (Bidirectional Encoder Representations from Transformers) has led many researchers to use it for text classification and thus in the detection of fake news [5, 7]. A key limitation of emerging topics and the need to build models dedicated to a specific topic is the lack of sufficient training samples. To this end, researchers are leaning towards solutions based on ensemble methods, unsupervised learning, and data augmentation.

3 PROPOSED APPROACH

Figure 1 illustrates the pipeline of the proposed approach. We follow a two-step classification approach:

- First step consists of an initial classification based on ensemble learning in order to provide a first-level classification of *Conspiracy* and *Non-conspiracy* tweets.
- The second step consists of the final prediction that classifies the detected *Conspiracy* tweets as *5G-conspiracy* or *Other-conspiracy*.

The provided dataset consists of 1,135 samples of the *5G-conspiracy* class, 712 of the *Other-conspiracy* class and 4,198 samples of *Non-conspiracy* class. As described in [4], imbalanced datasets for training machine learning algorithms or deep learning approaches pose risks of bias towards the majority class. To this end, we sub-sample training tweets of the majority classes in order to balance the training sets and build the proposed classifiers. Specifically, Table 1 presents the number of training samples considered per classifier. In CL_i, the training samples of *5G-conspiracy* and *Other-conspiracy* are concatenated into an overall *Conspiracy* class (1,847 tweets) and an equal number of tweets is randomly sampled from the *Non-conspiracy* tweets.

In the first step of our approach, we train N classifiers CL_i, which are used to predict *Conspiracy* and *Non-conspiracy* tweets. N is empirically selected to be nine. An odd number of classifiers makes it possible to apply majority voting. Each classifier CL_i predicts a label of 1 for *Conspiracy* or 0 for *Non-conspiracy* tweets. Majority voting is applied and a final prediction per tweet is given by $\sum_{i=1}^{N} CL_i > N/2$ where, $N = 9$, and if *true* prediction = *Conspiracy* else prediction = *Non-conspiracy*. For each model, different sample of *Non-conspiracy* tweets is selected.

[1] https://huggingface.co/transformers/model_doc/bert.html

MediaEval'20, December 14-15 2020, Online

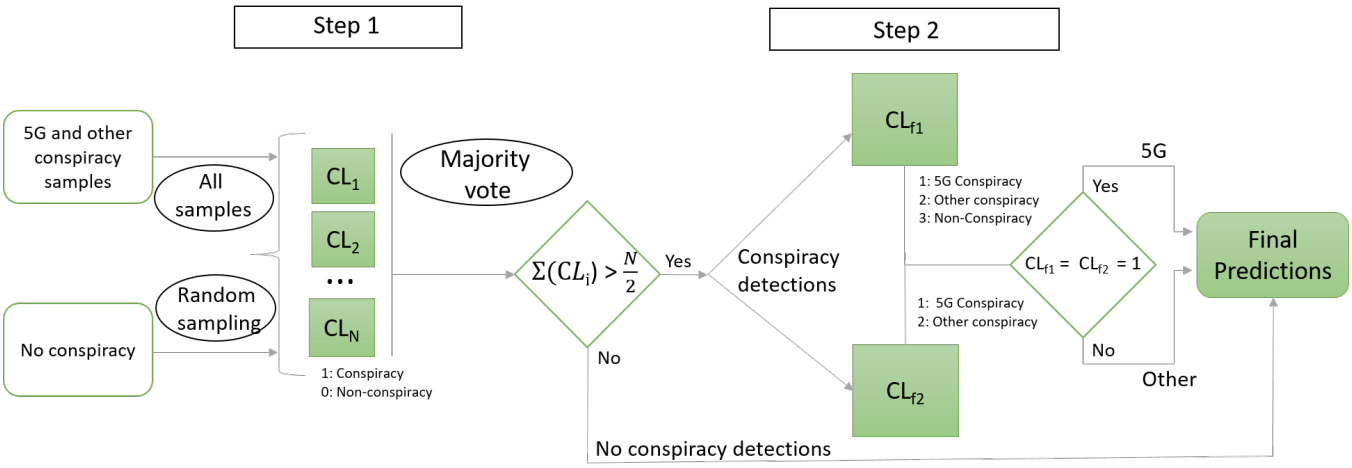

Figure 1: Our proposed pipeline for tackling the challenge of Corona virus and 5G conspiracy

Table 1: Summary of the training samples used to build the respective models

Label	CL_i	CL_{multi}	CL_{consp}
5G conspiracy	1847	712	712
Other conspiracy		712	712
Non-conspiracy	1847	712	-

Table 2: Evaluation results in terms of MCC, the official metric proposed for the task.

Method	MCC
three-class BERT	0.42
Proposed approach	**0.81**

In the second step, the predictions of *Non-conspiracy* are considered as final predictions without further processing while the *Conspiracy* tweets are further processed to distinguish *5G-conspiracy* from *Other-conspiracy*. In this step, two additional models are trained focusing on the detection of *5G-conspiracy* tweets. The first, CL_{f1}, is a three-class model (1: *5G-conspiracy*, 2: *Other-conspiracy* and 3: *Non-conspiracy*) trained using random samples from the majority classes and the total number of minority class samples (*Other-conspiracy*). The other model, CL_{f2}, is a binary classifier trained on the two *Conspiracy* classes. The final decision is taken if $CL_{f1} = CL_{f2} = 1 =$ *5G-conspiracy*. In any other case, the tweet is labeled as *Other-conspiracy*.

3.1 Implementation details

For tokenization, we employ `bert-base-uncased` of BertTokenizer applied to the text of the tweets. The text is limited to 160 tokens as input to the network. Considering that the maximum tweet length is 280 characters, it is most likely that the entire text is processed to calculate the prediction. As a backbone network, we employ the `bert-base-uncased` version of BERT [13], which is a compact transformer model, trained on lower-cased English text. The network architecture consists of 12 layers (i.e., Transformer blocks), with 768 hidden units, and 12 heads for multi-head attention layers, resulting in a total of 109M parameters.

We fine-tune our networks using Adam optimizer [6] with learning rate $2 * 10^{-5}$. The models are trained for 10 epochs with batch size 32 and categorical cross-entropy as the loss function. During training, we use dropout after the backbone network with 0.3 drop rate to prevent overfitting. Our models are evaluated against a validation set, and we select the versions that achieve the best performance in terms of accuracy as our final models.

4 RESULTS AND ANALYSIS

Initially, we trained a three-class model using the implementation details presented in subsection 3.1. From the annotated dataset, we randomly selected 100 samples per class as testing set and discarded them from the training phase in all runs. The performance of the model is 0.42 in terms of MCC. In order to improve the performance, we implemented the presented two-step classification approach resulting in increase of the MCC metric to 0.81 as presented in Table 2.

Our proposed approach achieved a score of 0.413 in terms of MCC on the provided testing set of unseen tweets.

5 DISCUSSION AND OUTLOOK

The proposed method achieves fairly accurate results in the task of FakeNews: Coronavirus and 5G conspiracy. More deep learning models, variants of BERT or other models, will be used in future experiments trying to achieve better performance. To tackle the limitation of insufficient training samples, we also intend to experiment with data augmentation approaches in order to create more samples of the minority classes and build more robust classifiers.

ACKNOWLEDGMENTS

This work is supported by the WeVerify project, which is funded by the European Commission under contract number 825297.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[2] Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. 2020. Detecting Misleading Information on COVID-19. *IEEE Access* 8 (2020), 165201–165215.

[3] Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sameer Singh, and Sean Young. 2020. Detecting covid-19 misinformation on social media. (2020).

[4] Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data* 6, 1 (2019), 27.

[5] Heejung Jwa, Dongsuk Oh, Kinam Park, Jang Mook Kang, and Heuiseok Lim. 2019. exBAKE: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (BERT). *Applied Sciences* 9, 19 (2019), 4062.

[6] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[7] Chao Liu, Xinghua Wu, Min Yu, Gang Li, Jianguo Jiang, Weiqing Huang, and Xiang Lu. 2019. A Two-Stage Model Based on BERT for Short Fake News Detection. In *International Conference on Knowledge Science, Engineering and Management*. Springer, 172–183.

[8] Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2019. A corpus of debunked and verified user-generated videos. *Online information review* (2019).

[9] Konstantin Pogorelov, Daniel Thilo Schroeder, Luk Burchard, Johannes Moe, Stefan Brenner, Petra Filkukova, and Johannes Langguth. 2020. FakeNews: Corona Virus and 5G Conspiracy Task at MediaEval 2020. In *MediaEval 2020 Workshop*.

[10] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

[11] Karishma Sharma, Sungyong Seo, Chuizheng Meng, Sirisha Rambhatla, Aastha Dua, and Yan Liu. 2020. Coronavirus on social media: Analyzing misinformation in Twitter conversations. *arXiv preprint arXiv:2003.12309* (2020).

[12] Samia Tasnim, Md Mahbub Hossain, and Hoimonty Mazumder. 2020. Impact of rumors or misinformation on coronavirus disease (COVID-19) in social media. (2020).

[13] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962* (2019).

[14] Tom Warren. 2020. British 5G towers are being set on fire because of coronavirus conspiracy theories. (Apr 2020). https://www.theverge.com/2020/4/4/21207927/5g-towers-burning-uk-coronavirus-conspiracy-theory-link

[15] 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*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6