MeVer team tackling Corona virus and Conspiracies using Ensemble Classification

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

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

ABSTRACT

This paper presents the approach developed by the Media Verification (MeVer) team to tackle the task of Corona Virus and Conspiracies Multimedia Analysis Task at the MediaEval 2021 Challenge. We utilized ensemble learning and propose a two-stage classification approach that aims to overcome the challenge of the imbalanced and relatively small training dataset. We deal with the problem as binary classification in the first stage and in the second stage we predict the multi-labels. We experimented with fine-tuning pre-trained Bidirectional Encoder Representations from Transformers (BERT) and achieved a score of 0.294 in terms of the Matthews Correlation Coefficient (MCC), which is the official evaluation metric of the task. Additionally, leveraging on the proposed two-stage classification approach, we extracted a set of feature representations (BoW, TfIDF, embeddings) and classify them using traditional machine learning algorithms (Support Vector Machines, Logistic Regression) achieving in the best run a score of 0.292 of MCC.

1 INTRODUCTION

The challenge of COVID 19-related misinformation has emerged with the COVID 19 pandemic and continues to concern the community about the amount of misinformation being disseminated and its implications for many areas, such as health and society [6, 15]. The need to develop methods to combat the dissemination of COVID-related conspiracies triggered the organization of the last year's task of FakeNews: Coronavirus and 5G conspiracy in the MediaEval 2020 Challenge [11] and this year's task of FakeNews: Corona Virus and Conspiracies Multimedia Analysis [10, 12].

A critical role in developing accurate methods for the automatic detection of misleading tweets (and any other text or multimedia item) plays the amount of annotated training samples. Due to the relatively small training dataset provided to deal with the challenge of detecting corona virus conspiracies, our approach follows a two-stage pipeline built on ensemble classification. In the first stage the task is converted to a binary classification problem that classifies the tweets in COVID Conspiracy tweets (involving both promoting and discussing a conspiracy) and non-Conspiracy tweets. In the second stage, the COVID conspiracy tweets are further classified in promoting conspiracy (tweets that promotes, supports, claim, insinuate some connection between COVID-19 and various conspiracies) and discussing conspiracy (just mentioning the existing various conspiracies connected to COVID-19). The final output of the methods is a three-class prediction.

MediaEval'21, December 13-15 2021, Online

2 RELATED WORK

Several works have been introduced dealing with the detection and verification of COVID 19-related misinformation utilizing machine and deep learning approaches [1, 3, 16]. An overview of CONSTRAINT 2021 Shared Tasks: Detecting English COVID-19 Fake News and Hindi Hostile Posts [8] shows that BERT or its variations was used for building the most successful models.

A significant contribution to combat misinformation is the creation of large enough annotated datasets which will serve to build more accurate models. Patwa et al. [9] released a dataset of 10,700 social media posts and articles of real and fake news on COVID-19. In Shahi et al. [14], the first multilingual cross-domain dataset of 5,182 fact-checked news articles for COVID-19 was introduced.

3 APPROACH

We first utilized the approach that we had developed in last year's task of FakeNews: Coronavirus and 5G conspiracy [7]. We adapted the method by corresponding the *5G conspiracy* class to the *Promote Conspiracy* class of this year's task, the *Other Conspiracy* to *Discuss Conspiracy* and the *Non-Conspiracy* to *Non-Conspiracy*. In short, it is a two-step classification approach that first applies an initial classification based on ensemble learning in order to provide a first-level classification of the Conspiracy and Non-conspiracy tweets and then a second step that predicts the classified Conspiracy tweets whether they are promoting conspiracy or discussing a conspiracy. For further details about the approach, the reader is referred to last year's working notes [7].

In addition, we run a set of complementary experiments based on the proposed approach, i.e. leveraging on the two-stage classification, experimenting with different feature representations and classify them using machine learning algorithms. In the following, we first describe how we deal with the imbalanced dataset, then we list the different combinations of features and models that we used in our experiments and we conclude with the results of the proposed runs on the provided testing set of unseen tweets.

3.1 Dealing with the imbalanced dataset

The provided dataset consist of 1,554 tweets in total for which 516 promotes COVID-related conspiracies (*Promote*), 271 discusses COVID-related conspiracies (*Discuss*) and 767 do not refer to COVID-related conspiracies (*Non-Conspiracy*). Johnson el al. [4] published a survey on deep learning with class imbalance showing that machine and deep learning approaches are essentially affected in terms of prediction accuracy when trained with imbalanced samples. 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, the classifiers of the first stage were trained with

540 samples of Conspiracy tweets (270 random samples of *Promote* class and 270 random samples of *Discuss* class) and 540 samples of Non-Conspiracy tweets. In the second stage, we trained a binary classifier with positive class the *Promote* class and negative class the *Discuss* class and a three-class model (*Promote*, *Discuss*, *Non-Conspiracy*) by randomly selecting 270 samples from each class for balance.

3.2 Feature representation and machine learning algorithms

In our additional experiments, we extracted five feature representations: i) **BoW**: A simple and effective model for text representation is the Bag-of-Words (BoW) Model. The model throws away all of the order information in the words and focuses on the occurrence of words in a tweet. ii) **TFIDF**: term frequency–inverse document frequency reflects how important a word is to a tweet in a collection of tweets. iii) **BERT**: We employ the bert-base-uncased version of BERT [2], which is a compact transformer model, trained on lower-cased English text. iv) **Distil**: We employ DistilBERT [13], which is a small, fast, cheap and light Transformer model trained by distilling BERT base. v) **Roberta**: We employ the RoBERTa model [5], which is built on BERT and and modifies key hyperparameters, removing the next-sentence pretraining objective and training with much larger mini-batches and learning rates. Each feature representation is fed in an SVM and a LR and we conclude with a set of multiple classifiers (Bow + SVM, BERT + LR, etc.)

3.3 Runs

The classifiers are trained on the samples presented in Section 3.1. In the first stage the predictions of the binary classifiers and fused using majority voting and provided to the second stage where the final predictions are calculated. We submitted four runs based on different combinations of the feature representations and machine learning algorithms.

- *Run 1*: In the first stage we build a ensemble of binary models combining all feature representations and both machine learning algorithms. The predictions of the models are fused using majority voting and in the second stage the tweets classified as Conspiracy are further fed in an ensemble of three-class models and binary classifiers (Promote vs Discuss) again trained on all combinations.
- *Run 2*: The first stage is the same as with Run 1 and in the second stage we fuse the predictions of binary classifiers trained on *Promote* vs *Discuss* classes and the *Conspiracy* vs *Non-Conspiracy* classes. For the *Conspiracy* vs *Non-Conspiracy* models we use all training samples.
- *Run 3*: In the first stage we select a combination of feature representations and machine learning algorithms which derived as the best combination in terms of accuracy based on cross validation. In the second stage we follow the combinations of *Run 2*.
- *Run 4*: In the first stage, we select only the combinations of BoW and BERT feature representations and LR and SVM. For each combination, we train N models with sub-samples of the training set. Similarity, the second stage fuses the predictions on models trained on the same combinations on the three classes.
- *Run 5*: This run is the method proposed in [7].

Table 1: Evaluation results

Run id	Run 1	Run2	Run 3	Run 4	Run 5
MCC	0.257	0.268	0.238	0.292	**0.294**

4 RESULTS AND ANALYSIS

The proposed approach of Papadopoulou et al. [7] achieved the best score (among our runs) of 0.294 in terms of MCC on the provided testing set of unseen tweets for the task of FakeNews: Corona Virus and Conspiracies Multimedia Analysis Task. In Table 1, the evaluation results in terms of MCC on the unseen tweets are presented for the five submitted runs. We observed that the accuracy of the four additional runs compared to the approach of Papadopoulou et al. [7] is slightly worse. The fact that traditional feature representation such as BoW and TFIDF combined with emebeddings achieve similar results to more complex deep learning approaches highlights the challenge of the limited training data. We assume that with a significantly larger training set the approach of Papadopoulou et al. [7] will achieve much better predictions.

5 DISCUSSION AND OUTLOOK

The proposed method achieves fairly accurate results in the task of FakeNews: Corona Virus and Conspiracies Multimedia Analysis Task. We followed our approach introduced in the MediaEval 2020 Challenge and based on the proposed pipeline we experimented with different setups by extracting several feature representations and using them to train traditional machine learning algorithms. We noticed that fusing the predictions of different feature representations and classification models we achieved almost the same results as with fine-tuning pre-trained BERT, one of the most popular transformer models. We observed that the limitation or the relatively small training set affects the prediction accuracy of the models negatively and augmentation techniques to create more samples of the minority classes could be a step to improve the predictions in future implementations.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Mabrook S Al-Rakhami and Atif M Al-Amri. 2020. Lies kill, facts save: detecting COVID-19 misinformation in twitter. *Ieee Access* 8 (2020), 155961–155970.
[2] 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).
[3] Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. 2020. Detecting misleading information on covid-19. *Ieee Access* 8 (2020), 165201–165215.
[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] 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. *arXiv preprint arXiv:1907.11692* (2019).

[6] Salman Bin Naeem, Rubina Bhatti, and Aqsa Khan. 2021. An exploration of how fake news is taking over social media and putting public health at risk. *Health Information & Libraries Journal* 38, 2 (2021), 143–149.

[7] Olga Papadopoulou, Giorgos Kordopatis-Zilos, and Symeon Papadopoulos. 2020. MeVer Team Tackling Corona Virus and 5G Conspiracy Using Ensemble Classification Based on BERT. (2020).

[8] Parth Patwa, Mohit Bhardwaj, Vineeth Guptha, Gitanjali Kumari, Shivam Sharma, Srinivas Pykl, Amitava Das, Asif Ekbal, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*. Springer, 42–53.

[9] Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: Covid-19 fake news dataset. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*. Springer, 21–29.

[10] Konstantin Pogorelov, Daniel Thilo Schroeder, Stefan Brenner, and Johannes Langguth. Online, 13-15 December 2021. FakeNews: Corona Virus and Conspiracies Multimedia Analysis Task at MediaEval 2021. In *MediaEval 2021 Workshop*.

[11] 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*.

[12] Konstantin Pogorelov, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, and year=2021 Johannes Langguth, booktitle=Proc. of the 2021 Workshop on Open Challenges in Online Social Networks, pp. 21-25. WICO Text: A Labeled Dataset of Conspiracy Theory and 5G-Corona Misinformation Tweets.

[13] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[14] Gautam Kishore Shahi and Durgesh Nandini. 2020. FakeCovid–A multilingual cross-domain fact check news dataset for COVID-19. *arXiv preprint arXiv:2006.11343* (2020).

[15] Sander van Der Linden, Jon Roozenbeek, and Josh Compton. 2020. Inoculating against fake news about COVID-19. *Frontiers in psychology* 11 (2020), 2928.

[16] Apurva Wani, Isha Joshi, Snehal Khandve, Vedangi Wagh, and Raviraj Joshi. 2021. Evaluating deep learning approaches for covid19 fake news detection. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*. Springer, 153–163.