Detecting COVID-19-Related Conspiracy Theories in Tweets

Youri Peskine, Giulio Alfarano, Ismail Harrando, Paolo Papotti, Raphael Troncy
EURECOM, France
firstName.lastName@eurecom.fr

ABSTRACT

Misinformation in online media has become a major research topic the last few years, especially during the COVID-19 pandemic. Indeed, false or misleading news about coronavirus have been characterized as an *infodemic*[1] by the World Health Organization, because of how fast it can spread online. A considerable vector of spreading misinformation is represented by conspiracy theories. During this challenge, we tackled the problem of detecting COVID-19-related conspiracy theories in tweets. To perform this task, we used different approaches such as a combination of TFIDF and machine learning algorithms, transformer-based neural networks or Natural Language Inference. Our best model obtains a MCC score of 0.726 for the main task on the validation set and a MCC score of 0.775 on the test set making it the best performing method among the challenge competitors.

1 INTRODUCTION

The full description of the task is detailed in [7] and more information about the dataset can be found in [8]. Text classification is a problem widely studied in many different fields for various applications such as sentiment analysis or topic modeling. Standard machine learning based approaches used in combination with Term Frequency - Inverse Document Frequency (TFIDF) are considered decent baselines [2] for performing text classification tasks. However, the recent introduction of transformer-based architectures like BERT [1], RoBERTa [3] or DistilBERT [9] has allowed significant improvement in various text-based problems [5].

2 APPROACH

In order to tackle this challenge, we studied three different kind of approaches. The first uses a combination of TFIDF and machine learning algorithms. The second approach uses Natural Language Inference (NLI) combined with metadata from Wikipedia. The third approach aims at fine-tuning transformer-based models. In the following sections, we discuss the experiments we pursued for each of these approaches. In order to ease reproducibility, we release all our code at https://github.com/D2KLab/mediaeval-fakenews.

2.1 TFIDF-based approach

TFIDF is one of the most widely used feature extraction techniques in the field of text processing, often used in parallel with preprocessing techniques such as tokenization, capitalization and stop word removal, which we also applied to the dataset in question. The derived features are fed to different supervised machine learning methods that allow us to obtain a first baseline. Several algorithms have been tested: Decision Tree, Naive Bayes classifier (Gaussian and Bernoullian), AdaBoost, Ridge and Logistic Regression. In the case of Task 1, these were used in a multi-class asset. In the multi-label case of Task 2, we used a multi-output classifier with the different methods listed above as estimators: in this scenario the algorithm instantiates a binary model for each conspiracy theory. Finally, in the Task 3, only the strictly tree-based algorithms were tested, since they are the only ones to allow a multi-label and multi-class output.

2.2 NLI-based approach

This approach relies on leveraging pre-trained language models that are then fine-tuned on the task of NLI. Put simply, given two statements (a *premise* and a *hypothesis*), these models are trained to classify the logical relationship between them: *entailment* (agreement or support), *contradiction* (disagreement), or *neutrality* (undetermined). Since these models are trained to project statements that share similar opinions into close points in their embedding space, our hypothesis was to identify the tweets that support/discuss the same conspiracies using this common embedding space.

For the Task 1, we need to differentiate between the different stances (agreement, discussion, neutrality) regarding conspiracy theories. Therefore, we generate an embedding for all the tweets using the fine-tuned model, and we then classify them using a K-nearest Neighbor classifier, the idea being that tweets sharing similar stances would be embedded close to each other. This approach can also be applied as is for the second task.

For the Task 2, we provided as a premise to the model a definition of each conspiracy theory, relying mostly on the Wikipedia articles describing them, thus classifying a tweet as pertaining to one of the listed conspiracies if the pre-trained model predicts that there is a entailment relationship between the definition of the conspiracy (premise) and the tweet text (hypothesis).

Finally, as a combination of both methods, we also used some annotated tweets related to a specific conspiracy as a premise instead of a definition of the conspiracy, and proceed to classify them by whether an entailment relationship is found.

2.3 Transformer-based approach

Transformer-based models have been performing remarkably well on text classification tasks in the last few years. For our submissions, we used RoBERTa [3] large pre-trained models and COVID-Twitter-BERT (CT-BERT) [6] pre-trained models in different ways to tackle each sub-task. The CT-BERT model is a BERT-large model pre-trained on COVID-related tweets. For this approach, we decided to use some pre-processing on the input tweets. We replaced all emojis with their textual meaning, and removed all the '#' characters.

The simplest strategy when working with transformer-based models for the first task is to approach it as a 3-class classification

[1] https://www.who.int/health-topics/infodemic

MediaEval'21, December 13-15 2021, Online

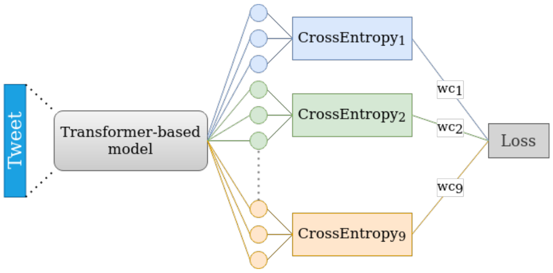

Figure 1: Training framework of transformer-based models performing Task 3. Each color represent a conspiracy theory.

problem. Both RoBERTa and CT-BERT are fine tuned on the data to perform classification with a weighted Cross Entropy loss function. We approach the second task as a multi-label binary classification problem. Both models are fine tuned to perform this objective with a weighted Binary Cross Entropy loss function.

The third and main task can be performed with different strategies. We first try to combine our results of the first two tasks, by labeling the tweets with the level detected in Task 1 for the conspiracy theories detected in Task 2. While this approach obtains convincing results, it is not able to deal properly with cases where tweets discuss about one conspiracy theory but support another one. An alternative approach is to train both transformer-based models to perform the main task directly. These models are fine tuned for nine different classification problems with nine Cross Entropy loss functions, one for each conspiracy theory. The final loss is the weighted sum of the nine losses. This training framework is illustrated in Figure 1. The advantage of such approach is that a single model is trained to perform all the tasks at once, because the first two tasks are just simplifications of the main task.

In our experiments, the weights of all our loss functions are proportional to the inverse frequency of each class or sub-class they are related to, and all our models are trained using the AdamW [4] optimizer.

3 RESULTS AND ANALYSIS

Our results for this challenge are presented in Table 1. All the models have been first evaluated on a stratified 5-fold cross-validation set and then evaluated on the test set.

Transformer-based approaches obtained the most competitive results. First we notice that RoBERTa models are under-performing compared to CT-BERT models on all the tasks. The latter models are more suited to this dataset because it contains tweets that use plenty of COVID-related vocabulary that would not be understood with the former models. It is also worth mentioning that models trained on the main task (Task 3) perform better on Task 1 than their task-specific counterpart.

For the TFIDF approach, the best performing method for Task 1 is the Support Vector Machine with a MCC score of 0.461. For Task 2, the Decision Tree gave the best result with 0.585 using the multi output classifier. On Task 3, the best result was given by the Decision Tree with an MCC of 0.497.

For the NLI approach, the observed results of these methods on cross-validation did not measure up to the fully-trained models. They remain an interesting alternative in the case where annotated data is lacking: a few tweets of each class or, minimally, just the definition of the classes are enough to provide some decent results.

Table 1: MCC results for each task, based on stratified 5-fold cross-validation set and then on the test set

	Models	Evaluation MCC	Test MCC
Task 1	TFIDF (SVC)	0.461	0.498
	NLI transformer	0.426	X
	RoBERTa	0.624	X
	CT-BERT	0.676	X
	RoBERTa-task3	0.667	X
	CT-BERT-task3	0.700	0.720
	Ensembling Models	0.716	**0.733**
Task 2	TFIDF (Multi output clf)	0.585	0.317
	NLI Wikipedia	0.310	X
	RoBERTa	0.731	X
	CT-BERT	0.780	**0.774**
	RoBERTa-task3	0.734	X
	CT-BERT-task3	0.743	0.719
	Ensembling Models	0.781	0.768
Task 3	TFIDF (DT)	0.497	0.186
	RoBERTa-task1+task2	0.675	X
	CT-BERT-task1+task2	0.717	**0.775**
	RoBERTa-task3	0.690	X
	CT-BERT-task3	0.706	0.713
	Ensembling Models	0.726	0.676

Looking at conspiracy theories, our worst results on Tasks 2 and 3 are about the *Intentional Pandemic* theory, even though it is the most represented class in the dataset. Instead, our best results are obtained with the *Harmful Influence* and *New World Order* theories. One possible explanation is that both conspiracies can be represented with very specific keywords ('5g' or 'NWO' for example).

We also performed late fusion ensembling through majority voting with different combination of transformer-based models to further improve our results on all the tasks. While this was our best results on a stratified 5-fold cross-validation set, this was less competitive on the test set.

4 DISCUSSION AND OUTLOOK

In this paper, we presented three different methods to perform COVID-19-related conspiracy theories detection in tweets. The first approach consist of a combination of standard machine learning based algorithm with TF-IDF. The second approach uses NLI with transformer-based models and Wikipedia enrichment. The last approach aims at fine-tuning transformer-based models on the given dataset. Our best model obtains a MCC score of 0.775 for the main task on the test set which outperforms by a large margin all the other competitors in this challenge.

ACKNOWLEDGMENTS

This work has been partially supported by CHIST-ERA within the CIMPLE project (CHIST-ERA-19-XAI-003).

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2019). arXiv:cs.CL/1810.04805

[2] Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, and Brown. 2019. Text Classification Algorithms: A Survey. *Information* 10, 4 (2019).

[3] 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. (2019). arXiv:cs.CL/1907.11692

[4] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. (2019). arXiv:cs.LG/1711.05101

[5] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep Learning Based Text Classification: A Comprehensive Review. (2021). arXiv:cs.CL/2004.03705

[6] Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. (2020). arXiv:cs.CL/2005.07503

[7] Konstantin Pogorelov, Daniel Thilo Schroeder, Stefan Brenner, and Johannes Langguth. 2021. FakeNews: Corona Virus and Conspiracies Multimedia Analysis Task at MediaEval 2021. In *Multimedia Benchmark Workshop*.

[8] Konstantin Pogorelov, Daniel Thilo Schroeder, Petra Filkukova, and Johannes Langguth. 2021. WICO Text: A Labeled Dataset of Conspiracy Theory and 5G-Corona Misinformation Tweets. In *Workshop on Open Challenges in Online Social Networks (OASIS)*.

[9] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. (2020). arXiv:cs.CL/1910.01108