Evaluating TF-IDF and Transformers-based Models for Detecting COVID-19 related Conspiracies

Rohullah Akbari[1,*]

[1]Simula Research Laboratory, Norway

Abstract

The proliferation of misinformation and conspiracy theories on online social media platforms has become a significant concern for public health and safety. To effectively combat this issue, a new generation of data mining and analysis algorithms is essential for early detection and tracking of these information cascades. In this paper, we employed a multifaceted approach for detecting and identifying conspiracy theories and misinformation spreaders related to the Coronavirus pandemic. Specifically, we utilized Text-Based Detection (Task 1) through a combination of TF-IDF-based and Transformers-based methods, Graph-Based Detection (Task 2) through a graph convolutional network, and alternative Transformers-based methods to improve the results of Task 1. Our efforts have yielded promising results, with our best models achieving an impressive MCC score of 0.705 for Task 1, 0.041 for Task 2, and 0.698 for Task 3.

1. Introduction

The COVID-19 pandemic and the associated lockdown formed the basis for many false news stories and conspiracy myths. Spontaneous and intentional digital FakeNews wildfires over online social media can be as dangerous as natural fires. The *FakeNews* Task at the MediaEval challenge 2022 targeted the detection of misinformation and its spreaders in tweets. More precisely, this task focuses on analyzing tweets, public user properties, and their connections related to Coronavirus conspiracy theories to detect conspiracies and misinformation spreaders. The description of the task and more information about the dataset can be found in [1]. The detection and verification of COVID-19-related misinformation using machine and deep learning techniques have been addressed in a number of papers [2, 3, 4, 5, 6]. An overview of previous work shows that COVID-Twitter-BERT (CT-BERT) is best suited for building the most successful model for COVID-19-related misinformation and conspiracy detection [4, 7].

2. Text-Based Misinformation and Conspiracies Detection

2.1. The TF-IDF approach

In this section, we will create nine distinct TF-IDF models for each of the nine categories. We are interested to see if the TF-IDF technique can outperform the CT-BERT model, and if not, how close it can come. This approach is based on using *TfidfVectorizer* and Stochastic Gradient Descent classifier (SGD) from the *scikit-learn* framework [8]. SGD is a simple but very efficient approach to fit linear classifiers such as linear Support Vector Machines (SVM). SGD does not belong to any particular family of machine learning models; it is only an optimization

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online

[*] Corresponding author.

[†] These authors contributed equally.

rohullaa@uio.no (R. Akbari)

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

technique. Often, an instance of *SGD Classifier* has an equivalent estimator in the Scikit-learn API, potentially using a different optimization technique. For example, logistic regression is produced when **SGDClassifier(loss='log loss')** is used. The TF-IDF approaches in previous works have been only executed with unigrams [7]. This leads to mislaid learning since there could be important information in the bigrams and trigrams. We can see in Table 2 that N-grams such as "bill gate" and "new world order" could be very important for the classification of the conspiracies. Based on this, we have chosen to implement the TF-IDF with various N-grams including unigrams, bigrams, trigrams, and other ranges. In addition to that, we have also chosen to implement the SGD with different loss functions and penalties (see Table 1 for the parameters).

Table 1
Chosen parameters for the TF-IDF approach. Note that SGD with *hinge* loss is equivalent to linear SVM, SGD with *log* loss is equivalent to LogReg, etc.

Name of parameter	Parameter values
Ngrams	(1,1),(1,2), (1,3),(1,4) (2,2),(2,3), (2,4), (3,4)
SGD loss	hinge, log, modified_huber, squared_hinge, perceptron
SGD penalty	L1 , L2, Elastic net and none

Table 2
Top 5 most common bigrams and trigrams in the dataset of Task 1. These sequences have been generated after the removal of stopwords.

N-grams	Occurences
deep state	222
world order	103
new world	102
bill gate	98
population control	78
new world order	101
years ago cbs	13
cbs show 60	13
interview retired cdc	13
qr code system	12

2.2. Transformers-based approaches

The first Transformers approach (*One-for-All*) is based on training one CT-BERT model for classifying all of the conspiracy categories at once (see Figure 1). The CT-BERT is fine-tuned with nine different weighted Cross Entropy loss functions. The weights are computed by taking into account the number of samples in a specific category and dividing it by the numbers of each of the subcategories in that category. The optimizer used in this approach is AdamW [9]. Before feeding the text data into the model, we preprocessed it by converting the emojis into their textual meaning. Furthermore, the training of the model was done with 5-fold Cross validation and the model with the best test MCC score was chosen. The *One-for-One* approach is based on training nine separate CT-BERT models for the nine categories (the approach is shown in Figure 2). In this approach, we are not using any weighted loss function. Other than that, we are applying the same loss function, optimizer, and preprocessing method. The training of the model was done with stratified 5-fold cross-validation and the model with the best MCC score was chosen.

3. Graph-Based Conspiracy Source Detection

For this task, we applied a simple node classification where the nodes are representing the user's label for whether they are a misinformation spreader or not. We created a network for each of the users that had a label. The network consisted of all of the other users that had an

Figure 1:
The One-for-All approach for Task 1.

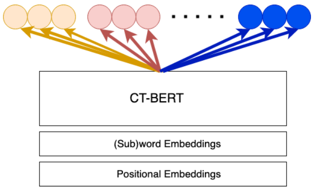


Figure 2:
The One-for-One approach for Task 1.

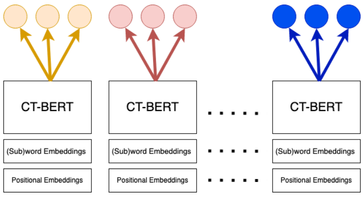
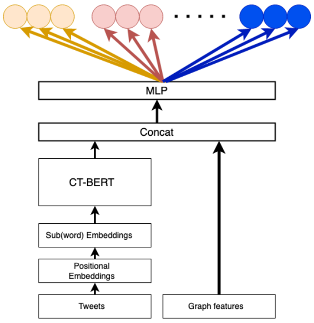


Figure 3:
The CT-BERT-Graph approach for Task 3.



edge directed to the main user and the users with low-weight values were removed. We chose to work with graph convolutional network (GCN) [10]. The implementation was done by using the *GCNConv* class from the *torch_geometric* library with PyTorch.

4. Graph and Text-Based Conspiracy Detection

In this section, we will examine whether we can improve the results from Section 2 by combining the data from Section 2 and Section 3. The output of the classifiers will be enriched by combining text with numerical features. We are proposing an approach that consists of training the CT-BERT with the text data and concatenating the last layer of the CT-BERT with the user information such as **verified_account**, **description_length**, **num_favourites**, **num_followers**, **num_statuses**, **num_friends** and **location_country**. The concatenating layer is then driven through a multilayer perceptron (MLP) and then processed into an output layer (see Figure 3). Our second approach is based on extending the text data with tweeters' statistics and then feeding it into the One-for-All approach 2.2. The numerical features that have been inserted in the text are separated with [SEP] token, e.g.

```
Tweet_text [SEP] 0 [SEP] 159 [SEP] 2812 [SEP] 566
[SEP] 1426 [SEP] 1041 [SEP] 3
```

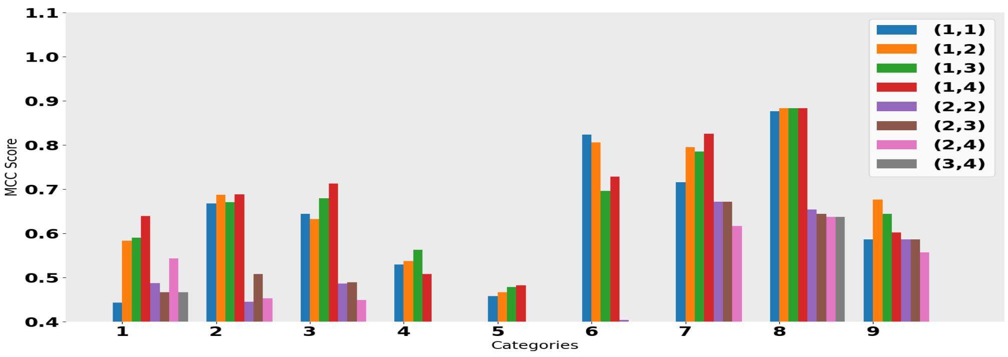
5. Results

As expected, the TF-IDF approach obtained a lower MCC score than the Transformers-based approaches (see Table 3). The One-for-One approach achieved the best score from all submitted runs. The TF-IDF approach does quite well for some of the categories, especially for the **Population reduction** and the **New World Order**. Bigrams such as "population control" and "bill gate" are very important for Population reduction, and "world order" and "new world" are obviously talking about the New World Order category (Table 2). Furthermore, we can see that the N-range such as (2,3), (2,4), and (2,4) did not do well and the dominating range is (1,4) (Figure 4). As a result, unigrams are crucial for the classification of conspiracies since the N-gram ranges without it performed poorly. We submitted only one run for Task 2 which resulted in an MCC score of **0.041** and clearly states that our implementation was not successful. The main reason for the poor performance could be the fact that we removed all the neighbors of the main user node that had low edge values. The combination of CT-BERT with numerical

Table 3
Official MCC scores per category for Task 1 and Task 3. Note that the One-For-All (Task 3) is the same as described in Section 1 but with extended data as described in Section 4.

Category	TF-IDF	One-for-All	One-For-One	One-For-All (Task 3)
Suppressed cures	0.484	0.737	**0.793**	0.563
Behaviour and Mind Control	0.504	0.698	0.700	**0.706**
Antivax	0.529	**0.726**	**0.726**	0.616
Fake Virus	0.378	**0.644**	0.628	0.628
Intentional Pandemic	0.353	0.545	0.592	**0.616**
Harmful Radiation/ Influence	0.617	0.723	**0.729**	0.695
Population reduction	0.710	0.825	0.795	**0.887**
New World Order	0.731	0.778	0.738	**0.850**
Satanism	0.414	0.638	0.663	**0.715**
Average	0.524	0.702	**0.705**	0.698

Figure 4: The plot is showing the performance of the different N-grams ranges. The MCC scores in this plot are from the validation dataset.



features resulted in an MCC score of **0.423**. This approach worsened the predictions, as the test MCC score went below the scores of Table 3. The One-for-All with extended text features achieved an MCC score of **0.698**. None of the approaches in this task improved the outcome of Task 1. However, the One-for-All technique in Task 3, was able to perform better for some of the categories (see Table 3).

6. Discussion and Outlook

We successfully implemented three approaches for Task 1; one TF-IDF approach and two Transformers-based approaches. We experimented with different N-gram ranges and found out that the N-gram range (1,4) was best suited for most of the categories. The best MCC score (**0.705**) was found with the One-for-One approach. We presented two approaches for improving the Task 1 results but none of them improved the results from Task 1.

References

[1] K. Pogorelov, D. T. Schroeder, S. Brenner, J. Moe, A. Maulana1, J. Langguth, Combining tweets and connections graph for fakenews detection at mediaeval 2022, in: roceedings of MediaEval 2022 CEUR Workshop, 2022.

[2] M. S. Al-Rakhami, A. M. Al-Amri, Lies kill, facts save: Detecting covid-19 misinformation in twitter, IEEE Access 8 (2020) 155961–155970. doi:10.1109/ACCESS.2020.3019600.

[3] A. Wani, I. Joshi, S. Khandve, V. Wagh, R. Joshi, Evaluating deep learning approaches for covid19 fake news detection, in: Combating Online Hostile Posts in Regional Languages during Emergency Situation, Springer International Publishing, 2021, pp. 153–163. URL: https://doi.org/10.1007%2F978-3-030-73696-5_15. doi:10.1007/978-3-030-73696-5_15.

[4] A. Glazkova, M. Glazkov, T. Trifonov, g2tmn at constraint@AAAI2021: Exploiting CT-BERT and ensembling learning for COVID-19 fake news detection, in: Combating Online Hostile Posts in Regional Languages during Emergency Situation, Springer International Publishing, 2021, pp. 116–127. URL: https://doi.org/10.1007%2F978-3-030-73696-5_12. doi:10.1007/978-3-030-73696-5_12.

[5] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Fighting an infodemic: COVID-19 fake news dataset, in: Combating Online Hostile Posts in Regional Languages during Emergency Situation, Springer International Publishing, 2021, pp. 21–29. URL: https://doi.org/10.1007%2F978-3-030-73696-5_3. doi:10.1007/978-3-030-73696-5_3.

[6] G. K. Shahi, D. Nandini, FakeCovid- A Multilingual Cross-domain Fact Check News Dataset for COVID-19, ICWSM, 2020. URL: https://doi.org/10.36190/2020.14. doi:10.36190/2020.14.

[7] Y. Peskine, G. Alfarano, H. Ismail, P. Papotti, R. Troncy, Detecting covid-19-related conspiracy theories in tweets (2021). URL: https://2021.multimediaeval.com/paper65.pdf.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, the Journal of machine Learning research 12 (2011) 2825–2830.

[9] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2017. URL: https://arxiv.org/abs/1711.05101. doi:10.48550/ARXIV.1711.05101.

[10] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2016. URL: https://arxiv.org/abs/1609.02907. doi:10.48550/ARXIV.1609.02907.