COVID-19 Fake News Detection by Using BERT and RoBERTa models

Tashko Pavlov and Georgina Mirceva
Faculty of computer science and engineering, Ss. Cyril and Methodius University in Skopje, Skopje, Macedonia
tashko.pavlov@students.finki.ukim.mk, georgina.mirceva@finki.ukim.mk

***Abstract* - We live in a world where COVID-19 news is an everyday occurrence with which we interact. We are receiving that information, either consciously or unconsciously, without fact-checking it. In this regard, it has become an enormous challenge to keep only true COVID-19 news relevant. People are exposed to these stories on a daily basis, and not all of them are true and fact-checked reports on the COVID-19 pandemic, which was the primary reason for our research. We accepted the challenge that fake news is extremely common and that some people take these news as they are. Knowing the true power of the most recent NLP achievements, in this research we focus on detecting fake news regarding COVID-19. Our approach includes using pre-trained BERT and RoBERTa models, which we then fine-tune on real and fake news about the COVID-19 pandemic. By using pre-trained BERT and RoBERTa models on tweet data, we explore their capabilities and compare them to previous research in regard to fine-tuned BERT models for this task in which we achieve better accuracy, recall and f1 score.**

Keywords - COVID-19; fake news; deep learning; transformer models; BERT; RoBERTa

I. INTRODUCTION

The COVID-19 pandemic is the first global pandemic since the Spanish flu that took place between 1918 and 1920. The world that we live in today during the COVID-19 pandemic is much more different than the previous pandemic. The main difference is that people can share their knowledge, interests, achievements, etc. through a variety of media such as images, texts, videos, and they are available to everyone. Every person can state something, that they feel is right, online. We live in a world where communicating with people all over the world is as simple as clicking two buttons, which is fascinating considering that during the previous pandemic the greatest achievement was the first transcontinental telephone call in 1915 [1]. The way that we came from the first transcontinental telephone call to communicating with other people on the other side of the world is amazing. We can easily share our opinion with the public and get responses for that.

Since the first ever COVID-19 case occurred in December 2019, the abuse of the world online communication started to arise. People started to share their thoughts on the new virus and in these communications fake news started to take place. COVID-19 fake news started to get a lot of attention to everyone. Even more active when vaccines and new variants arrived. The World Health Organization started to take actions in regard to stopping the misleading and false information available online, they called it the Infodemic [2].

We took an initiative to actively try and fight against fake news using Machine Learning (ML). Using the recent achievements in the field of Natural Language Processing (NLP) and ML we try to detect if a given information regarding COVID-19 is real or fake. For that reason, we use a pre-trained transformer models and fine-tune them on news data.

As fake news started to arise, much research has been done in this field of automatic fake news detection. However, data are required to train such state-of-the-art models, therefore the members of the NLP community created various fake news datasets. Some of these datasets are: CoAID [3], FakeCovid [4], ReCOVery [5] and CMU-MisCOV19 [6]. As many datasets were developed, also different models for detecting COVID-19 fake news were introduced.

Elhadad et al. [7] constructed a voting ensemble machine learning classifier with ten machine learning algorithms and seven feature extraction techniques for detecting misleading information related to COVID-19. Rutvik et al. [8] created a two-stage automated pipeline using state-of-the-art machine learning models for natural language processing for COVID-19 fake news detection. Hamid et al. [9] tackled this challenge on the MediaEval 2020 task [10], named “Fake Multimedia Twitter-Data-Based Analysis”. This challenge consists of two sub-tasks, one of which is text-based fake news detection, for which they proposed six different solutions limited on Bag of Words and BERT embeddings. Another benchmark dataset and results were developed in Hossain et al., NLP-COVID19 2020 [11] research. Namely they provide a benchmark dataset called COVIDLies [12], which consists of expert-annotated tweets to evaluate models on 86 different bits of COVID-19 related misinformation. They divide the misinformation detection in two sub-tasks. For an initial benchmark and for identification of key challenges that future models may come upon, they use a variety of different NLP models such as BERT, BiLSTM, SBERT and many more. Gundapu and Mamidi [13] tackled this challenge using different approaches such as LSTM model, BiLSTM with Attention model, CNN, CNN + BiLSTM, and they also performed this task on different transformer models such as BERT, XLNet, ALBERT, and finally they developed an Ensamble model from the aforementioned three transformer models. For evaluating the performance of these models, they used the same dataset as we use for our performance evaluation.

In this research paper we focus on the problem of detecting fake news regarding COVID-19 using pre-trained BERT and RoBERTa models on tweet data. The novelty in our approach is that we took advantage of the transformer models that were pre-trained on tweet data and therefore we examine the performance of such models. Using a COVID-19 related social media news gathered from tweets, Instagram posts, Facebook posts, or any other popular social media content, we managed to measure the performance of our approach and compare to the previous similar research and results.

The rest of this paper is organized in the following order. In Section II we describe the data that we used to fine-tune the transformer models. Section III represents the training methodology that we used to train the transformer models used to detect misleading COVID-19 information. Section IV denotes the discussion of our results and in Section V we take a look at our conclusions.

II. Dataset

The dataset that we used for training our models is the ConstraintAI's 2021 COVID-19 fake news dataset in English [14], [15], [16] containing 10,700 data samples. The data are divided into 3 disjoint datasets, including training set, validation set and testing set. The training set contains 6,420 data samples, the validation set which is used for tuning the models' hyperparameters contains 2,140 data samples and the remaining 2,140 are used for testing purposes. Each of these datasets contain 3 columns: id, tweet, label. The id represents a unique identifier of the data sample, the tweet column is someone's opinion found on social-media platforms and the third label column classifies the tweet in one of the two classes real or fake.

Table 1 shows the distribution of real and fake samples in the three datasets. We can see that the datasets are fairly balanced. In Table 2 we can see examples of data samples from the training set for both classes.

TABLE I. The Distribution of Real and Fake Famples in Each Dataset

Dataset	**Real**	**Fake**
train	3360	3060
validation	1120	1020
test	1120	1020

TABLE II. Examples of Real and Fake Samples from the Training Dataset

Tweet	**Label**
Pregnant women with COVID have a 25% higher rate of premature births.	Real
CDC Recommends Mothers Stop Breastfeeding To Boost Vaccine Efficacy.	Fake

Below we illustrate what are the most common words occurring in a news that are real or fake. In Fig. 1, which represents the real news word cloud, we can see that some of the words do not occur in the fake news word cloud. Such words are "amp", "new", "case", "confirmed case", "data" etc. On the other hand, in Fig. 2 we can clearly see that the most frequent words do not occur in the real news word cloud. Words such as "coronavirus", "vaccine", "donald trump", "bill gates", "government" etc. These frequency words can give us an interesting insight on what we can expect to occur in fake or real news.

III. Training and Testing Methodology

In this section, we'll look at how we trained the transformer models to determine whether a given statement is false or true news. On Fig. 3, we present the pipeline that is used. First, we preprocess the data. Then, pre-trained models are loaded, and then they are fine-tuned for solving the task at hand. Once the models are created, then a given query sample (news from the test set) is classified by using the models that are obtained.

First, we'll take a deeper look at how we prepared the data. Then, we will give details about the pre-trained models that are used in this research. Next, we fine-tune these pre-trained models for the purpose of our research. In this way, we used transfer learning in order to generate our final models.

The full code for the fine-tuning the models will be available on GitHub [17].

A. Data Preprocessing

In this subsection we will be talking about how we preprocessed the text data and what was the final form of the data that we used to feed into the transformer models.

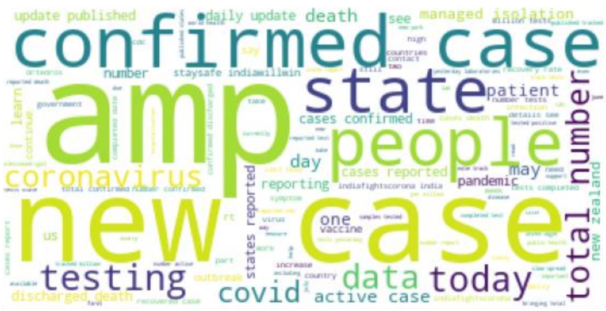

Figure 1. Real news word cloud

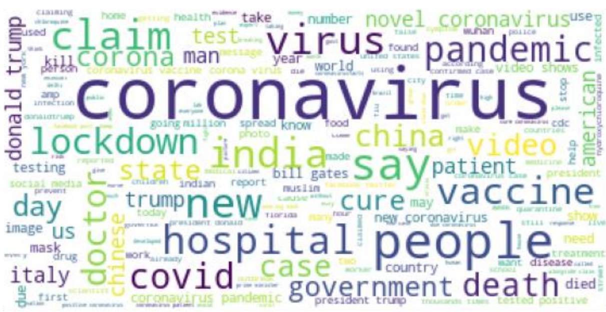

Figure 2. Fake news word cloud

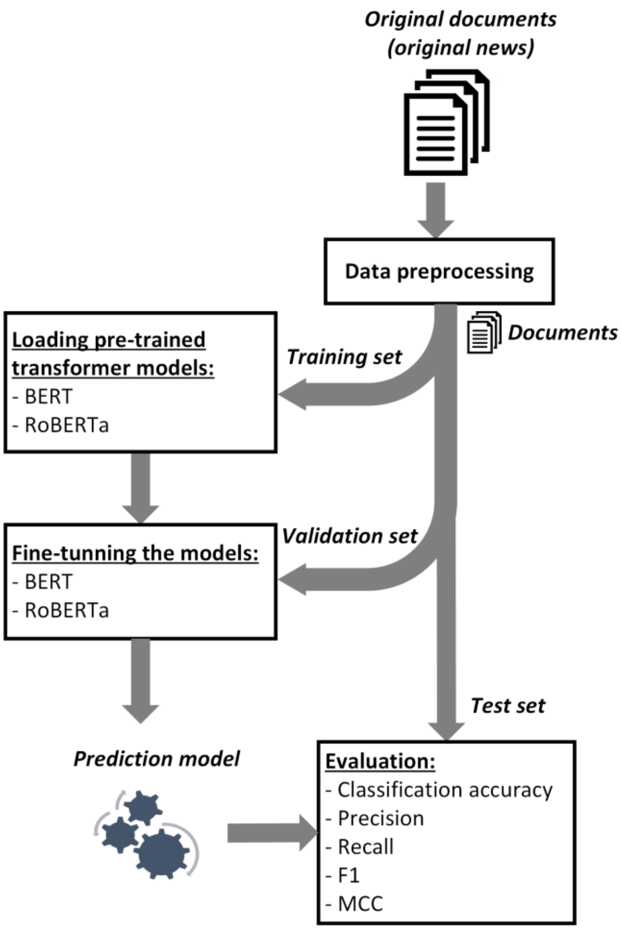

Figure 3. Our approach for COVID-19 fake news detection

On Fig. 4 we can see how we managed to preprocess the data in order to get to the final form of the text.

Firstly, we start with the initial tweet and move through the pipeline. The first filtering that we do is check whether the given word in our initial tweet sentence is an URL, if it is an URL we convert it into an URL token which is in the form of 'URL'. We do this because there are a lot of URLs in the tweets and they can mislead the model into believing that the URL carries important information, in this regard we treat each URL the same.

The next filtering that we do is check for emojis in the tweet. If emoji is present in the tweet, we convert it into an emoji token. For example, we convert 👍 into ':thumbs_up:' token. As a final step to the data preprocessing, we have the lowercase filtering. We simply transform the tweet into lower case letters. At the end of this pipeline, we get the final form of the tweet, which is free of urls and emojis. The tweets in this final form can then be fed into our transformer models.

We also preprocessed the label column, which has only two values: 'real' and 'fake'. We converted these values into 0 for 'fake' and 1 for 'real'.

B. *Transformer Models*

In this subsection we give details on how we managed to train the two transformer models for detecting COVID-19 fake news. As we mentioned previously, we did data preprocessing and we input the final form of the tweet sentence into our models. We used a pre-trained RoBERTa model on tweets (twitter-roberta-base-sentiment) [18] and pre-trained BERT model on COVID-19 related tweets (covid-twitter-bert-v2) [19]. The twitter-roberta-base-sentiment is a RoBERTa model [20] trained on ~58M tweets and then fine-tuned on for sentiment analysis downstream task. On the contrary, the covid-twitter-bert-v2 is a BERT-large-uncased model [21], which is pre-trained on a corpus of tweets regarding COVID-19 news. We favored using the pre-trained transformer models on tweets since we wanted to take advantage of the fact that the models have seen enough tweets and can be better at understating the tweets that we used for fine-tuning on our downstream task of detecting COVID-19 fake news. On Fig. 5 and Fig. 6 we can see how the training and validation loss change through each epoch of training for both models. We can note that the training loss decreases by each epoch.

Next, we will examine the models' hyperparameters, which are used to fine-tune the models. We will discuss more about why the specific values are used and how do they affect the results. As seen in Table 3, we fine-tuned both models with almost identical hyperparameters. The only difference is in the length of the input sequence to the model. In the process of fine-tuning, different hyperparameters were chosen and then validated on the test set. The best results were acquired by applying the hyperparameters that are displayed in Table 3.

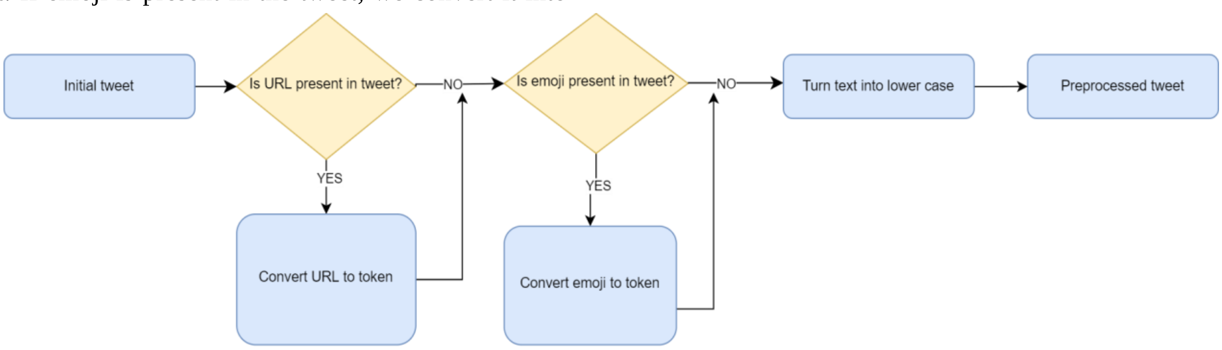

Figure 4. Data preprocessing pipeline

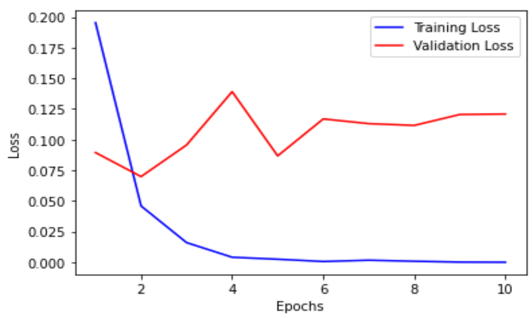

Figure 5. Training and validation loss for BERT model through 10 epochs.

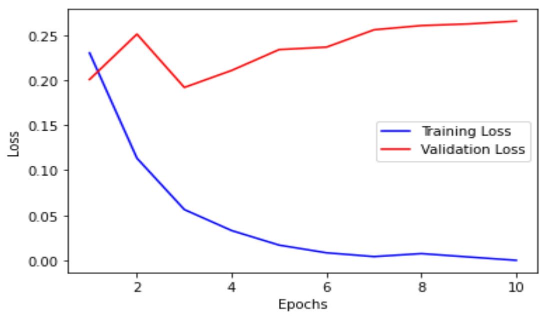

Figure 6. Training and validation loss for RoBERTa model through 10 epochs.

TABLE III. HYPERPARAMETERS USED FOR FINE-TUNING THE MODELS

Hyperparameter	**Model**	
	BERT	**RoBERTa**
Learning Rate	1e-05	1e-05
Batch Size	12	12
Optimizer	Adam	Adam
Max Length	128	286
Epochs	10	10

IV. RESULTS AND DISCUSION

In this section we present the results that are achieved, and we compare the different models that are used. For the purpose of comparing the performance of each model we used several evaluation measures, i.e., accuracy, precision, recall, F1-score and Matthew's correlation coefficient (MCC).

In Table 4, we present the results that are obtained when testing the pre-trained models (without fine-tuning). With these results, we aim to demonstrate the impact of fine-tuning the models. On the other hand, in Table 5, we show the results that are obtained for classifying the test samples using the fine-tuned models. If we take a look at Table 4 and Table 5 and compare the results, the impact of fine-tuning is inevitable. Namely, the values for all evaluation metrics are significantly improved with fine-tuning. Interesting thing to note here is that the pre-trained RoBERTa model performs better than the pre-trained BERT model, but things change when we fine-tune them. As can be seen, the fine-tuned BERT model on COVID-19 related tweets outperforms on all evaluation metrics. The reason for that we believe is that the BERT model was pre-trained solely for this purpose, to detect fake news from tweets. On the contrary, the RoBERTa model was pre-trained on general tweets, hence it does not handle COVID-19 related tweets with such performance.

TABLE IV. RESULTS OBTAINED BY USING THE PRE-TRAINED MODELS

Evaluation metric	**Model**	
	BERT	**RoBERTa**
Accuracy	0.3457	0.5504
Precision	0.2971	0.6167
Recall	0.1830	0.5504
F1	0.3250	0.5743
MCC	-0.3125	0.1792

TABLE V. RESULTS OBTAINED BY USING THE FINE-TUNED MODELS

Evaluation metric	**Model**	
	BERT	**RoBERTa**
Accuracy	**0.9831**	0.9752
Precision	**0.9796**	0.9708
Recall	**0.9883**	0.9821
F1	**0.9831**	0.9752
MCC	**0.9663**	0.9504

Finally, in Table 6 we give evidence about the time that is needed for training and testing both models. As it can be seen, when BERT transformer model is used, it takes more time for both training the models and classifying the test samples compared with the case when RoBERTa model is used. However, the time needed for classifying the test samples is not high, so the fine-tuned BERT model would be better choice because it outperforms the RoBERTa models based on the results obtained for all evaluation measures.

As mentioned in the introduction, many other research papers tackle this challenge. Many of them are also taking advantage of the Transformer models' capabilities. We made an analysis where we compared the evaluation metrics for the fine-tuned BERT models on COVID-19 fake news detection, see Table 7. The first BERT model is suggested in Gundapu and Mamidi's

TABLE VI. TRAINING AND TESTING TIMES USING BOTH MODELS

Time	**BERT**	**RoBERTa**
Training time	3h 7m 0s	1h 58m 7s
Testing time	1m 54s	0m 34s

TABLE VII. RESULTS FOR DIFFERENT FINE-TUNED BERT MODELS

Evaluation metric	Model	
	BERT [13]	**OurBERT**
Accuracy	0.9813	**0.9831**
Precision	**0.9813**	0.9796
Recall	0.9813	**0.9883**
F1	0.9813	**0.9831**

work [13], where they fine-tune a pre-trained BERT model on the same dataset. The second BERT model is the one that we are proposing, which is outperforming the previously mentioned model in the accuracy, recall and f1 metrics. The difference in the results is certainly not significant, but it can be said that using a pre-trained BERT model on data that is related to the challenge that we are trying to tackle can give advantages in performing better results, rather than using the pre-trained model on general data.

V. CONCLUSION

In this paper, we discussed how we may use state-of-the-art pre-trained transformer models to prevent spreading fake news regarding the COVID-19 pandemic. As we mentioned in the introduction, the use of online communication has made it easier to spread fake news and more difficult to determine whether a given news is real or fake. To solve this problem, we used pre-trained BERT and RoBERTa models, and then we fine-tuned them for solving the particular task. The results showed that both models are able to detect fake news very accurately. However, the fine-tuned BERT model outperformed the RoBERTa model.

Even though the models could almost perfectly predict fake news, there is still cause for concern because new things happen around COVID-19 every day, such as new variants or vaccines, making it difficult for the models to predict tweets containing recent statements about COVID-19 simply because the models have not seen any of these new statements. As a result, the models should be fine-tuned on new data periodically in order to produce good results on new COVID-19 statements.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of computer science and engineering at the “Ss. Cyril and Methodius University in Skopje”, Skopje, Macedonia.

REFERENCES

[1] https://en.wikipedia.org/wiki/First_transcontinental_telephone_call.

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

[3] L. Cui and D. Lee, “CoAID: COVID-19 Healthcare Misinformation Dataset,” ArXiv, abs/2006.00885, 2020.

[4] G.K. Shahi and D. Nandini, “FakeCovid - A Multilingual Cross-domain Fact Check News Dataset for COVID-19,” ArXiv, abs/2006.11343, 2020.

[5] X. Zhou, A. Mulay, E. Ferrara, and R. Zafarani, “ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research,” In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020.

[6] S.A. Memon and K.M. Carley, “Characterizing COVID-19 Misinformation Communities Using a Novel Twitter Dataset,” ArXiv, abs/2008.00791, 2020.

[7] M.K. Elhadad, K. Li, and F. Gebali, “Detecting Misleading Information on COVID-19,” IEEE Access, 8, pp. 165201-165215, 2020.

[8] R. Vijjali, P. Potluri, S. Kumar, and S. Teki, “Two Stage Transformer Model for COVID-19 Fake News Detection and Fact Checking,” 2020.

[9] A. Hamid, N. Sheikh, N. Said, K. Ahmad, A. Gul, L. Hassan, and A. I. Al-Fuqaha, “Fake News Detection in Social Media using Graph Neural Networks and NLP Techniques: A COVID-19 Use-case,” CoRR, abs/2012.07517, 2020.

[10] P. Patwa, M. Bhardwaj, V. Guptha, G. Kumari, S. Sharma, S. Pykl, A. Das, A. Ekbal, S. Akhtar, and T. Chakraborty, “Overview of CONSTRAINT 2021 Shared Tasks: Detecting English COVID-19 Fake News and Hindi Hostile Posts,” In Proceedings of the First Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation (CONSTRAINT), Springer, 2021.

[11] T. Hossain, R. Logan, A. Ugarte, Y. Matsubara, S. Young, S. Singh, “COVIDLies: Detecting COVID-19 Misinformation on Social Media,” In Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2), EMNLP 2020, Association for Computational Linguistics, 2020.

[12] https://ucinlp.github.io/covid19/

[13] S. Gundapu and R. Mamidi, “Transformer based Automatic COVID-19 Fake News Detection System,” ArXiv, abs/2101.00180, 2021.

[14] K. Pogorelov, D. T. Schroeder, L. Burchard, J. Moe, S. Brenner, P. Filkukova, and J. Langguth, “FakeNews: Corona Virus and 5G Conspiracy Task at MediaEval 2020,” In MediaEval 2020 Workshop, 2020.

[15] https://constraint-shared-task-2021.github.io/

[16] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M.S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, “Fighting an Infodemic: COVID-19 Fake News Dataset,” 2020.

[17] https://github.com/taskop123/COVID-19FakeNewsDetection

[18] F. Barbieri, J. Camacho-Collados, L. Espinosa-Anke, L. Neves, “TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification,” In Proceedings of Findings of EMNLP, 2020.

[19] M. Müller, M. Salathané, P. Kummervold, “COVID-Twitter-BERT: A Natural Language Processing Model to Analyze COVID-19 Content on Twitter,” arXiv preprint arXiv:2005.07503, 2020.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” 2019.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” CoRR, abs/1810.04805, 2018.