COVID-19 Fake News Detection Using Bidirectional Encoder Representations from Transformers Based Models

Yuxiang Wang, Yongheng Zhang[⋆], Xuebo Li, and Xinyao Yu[⋆⋆]

Johns Hopkins Unversity, Baltimore MD 21218, USA
{ywang594,yzhan470,xli248,xyu63}@jhu.edu

Abstract. Nowadays, the development of social media allows people to access the latest news easily. During the COVID-19 pandemic, it is important for people to access the news so that they can take corresponding protective measures. However, the fake news is flooding and is a serious issue especially under the global pandemic. The misleading fake news can cause significant loss in terms of the individuals and the society. COVID-19 fake news detection has become a novel and important task in the NLP field. However, fake news always contain the correct portion and the incorrect portion. This fact increases the difficulty of the classification task. In this paper, we fine-tune the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model as our base model. We add BiLSTM layers and CNN layers on the top of the fine-tuned BERT model with frozen parameters or not frozen parameters methods respectively. The model performance evaluation results showcase that our best model (BERT fine-tuned model with frozen parameters plus BiLSTM layers) achieves state-of-the-art results towards COVID-19 fake news detection task. We also explore keywords evaluation methods using our best model and evaluate the model performance after removing keywords.

Keywords: COVID-19 Fake news · BERT · Fine-tune · BiLSTM · CNN.

1 Introduction

In December 2019, COVID-19 pandemic outbroke in Hubei, Wuhan, China[17]. In January 2020, the World Health Organization recognized it as Public Health Emergency of International Concern[9]. The whole world has gone through the COVID-19 pandemic and the spread of COVID-19 caused big challenge for the world, and the pandemic related news brought many attentions in the society. As the situation becomes worse, the rate of those news is growing exponentially. For example, Twitter, Facebook, Instagram and many other social platforms update news on the pandemic everyday. However, it is noticeable that there exist some misinformation about COVID-19. Those news not only led the number of cases

[⋆] The first two authors have equal contribution.

[⋆⋆] The last two authors have equal contribution.

increases, but also let public feel anxious and take wrong actions. Therefore, it is necessary to build models that can distinguish between real news and fake news.

In order to fight against the spreading disinformation, in this paper, we use and fine-tune the pre-trained Bidirectional Encoder Representations from Transformers (BERT) based models to train social media news posts, which are already known for truth or fake, for better recognizing those possible false news that may appear in the future. BERT is a Google developed technique which is transformer-based, and it is used for Natural Language Processing[16]. We start with fine-tune the BertForSequenceClassification model as our base model. Then we add additional layers upon the base model combining with frozen parameters approaches, and achieve very good results by our fine-tune BERT + BiLSTM with frozen parameters model. This model obtains high accuracy and high F1 score, which can help the public to better distinguish the COVID-19 fake news. Additionally, we explore a method to evaluate the key words in fake news aiming to find potential important words in fake news to help us better identify fake news in the future.

Our structure for this paper can be summarized as follows. In section two, we introduce publications that are relevant to our work. Those publications are also the works we refer while working on our own work. In section three, we introduce the generate method for our work, including the choice of the dataset, the set up before the training procedure, and the training and evaluation part. In section four, we dive into detail to the experiments. We introduce five different models used for this paper and provide the experiments' result and analysis by introducing different evaluation metrics. We also write about the keywords experiment that can find keywords in fake news and real news. In section five, we discuss our results in terms of dataset, model hyperparameters, model structure, model evaluation, and keywords evaluation. In the last section, we make conclusion and write about possible future works.

2 Related Work

Adhikari [1] presents the first application of BERT to document classification. Aggarwal [2] demonstrates classification of fake news by fine-tuning deep bidirectional transformers based language Model. Devlin [3] introduces a new language representation model called BERT. Gundapu [4] introduces an ensemble of three transformer models (BERT, ALBERT, and XLNET) to detect fake news. Gupta [5] builds a model that makes use of an abusive language detector coupled with features extracted via Hindi BERT and Hindi FastText models and metadata. Kaliyar [6] proposes a BERT-based deep learning approach by combining different parallel blocks of the single-layer deep Convolutional Neural Network having different kernel sizes and filters with the BERT. Kula[7] presents a hybrid architecture connecting BERT with RNN. Liu [8] treats fake news detection as fine-grained multiple-classification task and use two similar sub-models to identify different granularity labels separately. Pham [10] explores encoding news title pairs and transforms into new representation space. Pham-Hong [11] uses a stack of BERT and LSTM layers to evaluate multilingual offensive language identification in social media. Safaya [12] describes approach to utilize pre-trained BERT models with Convolutional Neural Networks for sub-task of the Multilingual Offensive Language Identification shared task. Sun [13] investigates different fine-tuning methods of BERT on text classification task and provides a general solution for BERT fine-tuning. Tang [14] mentions keyword extraction using Attention-based Deep Learning models with BERT. Vijjali [15] leverages a novel fact checking algorithm that retrieves the most relevant facts concerning user claims about particular COVID-19 claims, and verifies the level of"truth" in the claim by computing the textual entailment between the claim and the true facts.

3 Method

3.1 Dataset

The COVID-19 Fake News Detection Dataset comes from the Kaggle website[1]. We have the balanced training data, which contains 6,420 data entries with variable id, tweet and label. We also have balanced testing data which contains 2,140 data entries with variable id and tweet. There are three main variables in our training dataset: 'id' indicates the id number of the tweet; 'tweet' means the actual context of the tweet/post; lastly, 'label' describes whether the news is real or fake. We combine those two datasets together and randomly split data into training set (90%) and test set (10%) using the set seed method.

3.2 Setup

Data pre-processing is essential for feeding the data into BERT. The pre-processing steps can be summarized as the following steps: First, we load the dataset. Second, we perform tokenization and Encoding. We use BertTokenizer and our own tokenizer to tokenize the tweets. [SEP] and [CLS] tokens need to be added at the end and beginning of every sentence. Then, we map tokens to ids. Third, we apply Padding and Truncation. BERT requires that all sentences must have the same fixed length and the max length of 512 tokens per sentence. We found out that only 10 out of 8560 rows has length that is over 512. Therefore, we set up our max length to 512. Last, we use Attention Masks. The purpose of adding the masks is to not incorporate the padded tokens into the interpretation of the sentences.

3.3 Training and Evaluation

Our training and evaluation procedure can be summarized as the following steps. First, we apply the BertForSequenceClassification model. Second, we fine-tune

[1] https://www.kaggle.com/elvinagammed/covid19-fake-news-dataset-nlp

the BERT model. Third, we add additional layers after the fine-tuned model, including CNN and Bidirectional LSTM, for both with and without freezing the parameters in the fine-tuned model. Then, we perform training, hyperparameter tuning, and testing. Last, we investigate key words that affect the authenticity of the news. The procedure can be visualized in Fig. 1.

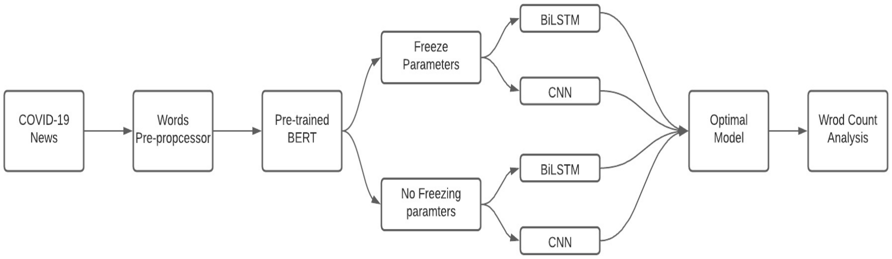

Fig. 1. Pipeline

4 Experiments

We build five different models to evaluate and compare the performance of fake news classification. We define Model 1 as the BERT fine-tuned model. Next, we define Model 2 as the BERT fine-tuned model with frozen parameters plus CNN layer(s). Then, we define Model 3 as the BERT fine-tuned model without frozen parameters plus CNN layer(s). Besides, we define Model 4 as the BERT fine-tuned model with frozen parameters plus BiLSTM layer(s). Lastly, we define Model 5 as the BERT fine-tuend model without frozen parameters plus BiLSTM layer(s).

4.1 Design Architecture and Hyperparameters

BERT fine-tune. For model 1, we use 2e-5 as the learning rate, and use 4 epochs to fine-tune the model.

BERT fine-tune + CNN. For model 2 and model 3, after the fine-tuned BERT model, we apply two convolutional layers with kernel size (1,768) and (2,768) and ReLU activation function. Then we follow a max pooling layer with the previous output size as the kernel size and the previous height of the output as the stride. After this, we add a dropout layer with the rate 0.1. Finally, we

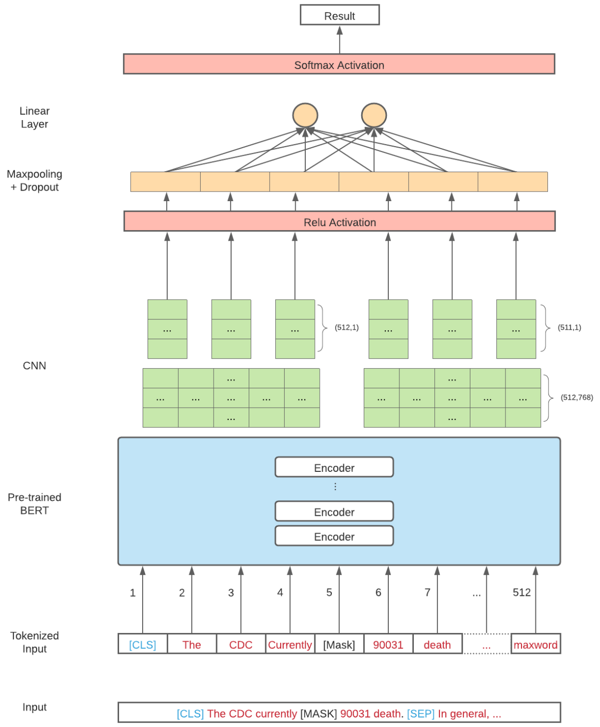

Fig. 2. BERT+CNN Model Architecture

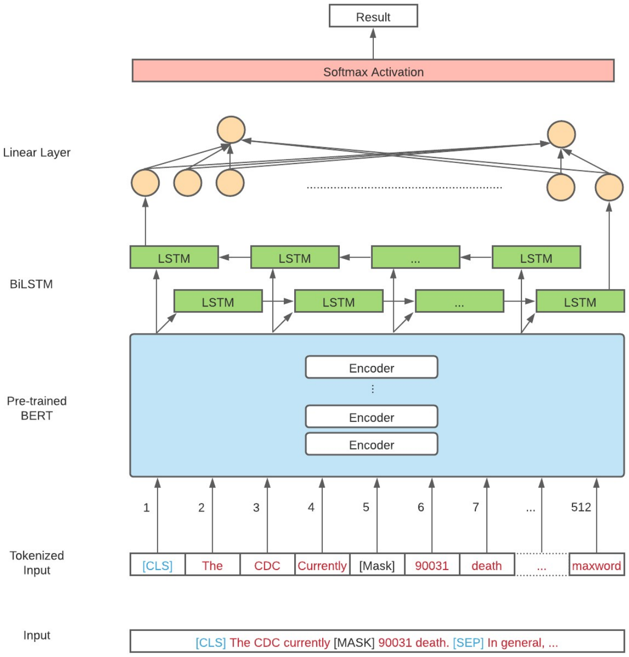

Fig. 3. BERT+BiLSTM Model Architecture

Table 1. Hyperparameters of BERT-based models

Model	LR	Epoch	# of Addition Layers	Kernel Size
Fine-tuned BERT (Model 1)	2e-5	4	1	-
Fine-tuned BERT (Freeze parameters) + CNN (Model 2)	5e-5	4	2	(1,768),(2,768)
Fine-tuned BERT + CNN (Model 3)	5e-5	4	2	(1,768),(2,768)
Fine-tuned BERT (Freeze parameters) + BiLSTM (Model 4)	5e-5	10	2	-
Fine-tuned BERT+ BiLSTM (Model 5)	5e-5	6	1	-

apply a linear layer with softmax activation function. The softmax activation function can be expressed as (1) [20]:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{1}$$

Where K = 2 because we have two classes, fake and real news. z indicates input. The classification result is the class with the highest output of the softmax activation function. Besides, the learning rate for those two models are both 5e-5 and the epoch for those two models are both 4. The architecture can be visualized in Fig. 2.

BERT fine-tune + BiLSTM. For model 4 and model 5, after the fine-tuned BERT model, we apply 2 BiLSTM layers to model 4 and 1 BiLSTM layers to model 5. After this, we apply a linear layer with softmax activation function. Besides, the learning rate for model 4 and model 5 are both 5e-5, and the epoch used for model 4 is 10 and for model 5 is 6. The architecture can be visualized in Fig. 3.

Architecture and Hyperparameter Summary The summarized hyperparameter settings for all of the models can be seen in Table 1.

4.2 Experimental Results

Evaluation Criteria To test our classifiers' prediction results on fake news dataset, we use the following metrics. First, we use test accuracy as our primary metric. In our task, the test accuracy is defined as (2):

$$\frac{Number\ of\ Correctly\ Classified\ News}{Total\ Number\ of\ News} \tag{2}$$

Table 2. Model Evaluation

Model	Test acc	Training loss	ROC AUC	F1 score
Fine-tuned BERT (Model 1)	0.9579	**0.0036**	0.9586	0.9607
Fine-tuned BERT (Freeze parameters) + CNN (Model 2)	0.9591	0.0200	0.9589	0.9622
Fine-tuned BERT + CNN (Model 3)	0.9439	0.0211	0.9449	0.9474
Fine-tuned BERT (Freeze parameters) + BiLSTM (Model 4)	**0.9614**	0.0197	**0.9607**	**0.9646**
Fine-tuned BERT+ BiLSTM (Model 5)	0.9346	0.0227	0.9351	0.9389

The second metric we use is training loss. We use cross-entropy loss can be defined as (3). The cross-entropy compares the model's prediction with the label which is the true probability distribution [18].

$$L = -(y \log(p) + (1 - y) \log(1 - p)) \tag{3}$$

The third metric we use is ROC AUC score. The ROC AUC stands for the area under the curve of ROC. The range of ROC AUC score is 0 to 1, and a large AUC value for a model indicates a good performance of the prediction.

The last metric we use is F1 score. It ranges from 0 to 1 and is calculated from the precision and the recall of our test results. The precision is the number of true positive results divided by the number of all positive results, and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive. The higher the score, the better performance it indicates. The formula of F1 score can be written as (4) [19]:

$$F1\ score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4}$$

Performance Analysis From the results in Table 2, model 4 has the highest test accuracy, ROC AUC and F1 score. The performance of model 2 is better than that of the BERT fine-tuned model as well. It is in our expectation that model 4 performs the best since BiLSTM considers the context before and after the target words. We also find out that adding additional layers will improve the accuracy because the new model can capture more information on the dataset. Overall, the performance of adding BiLSTM layer(s) is better than that of adding CNN layers under the same condition. We believe that the result is reasonable because BiLSTM consider the context of the sentence where CNN does not.

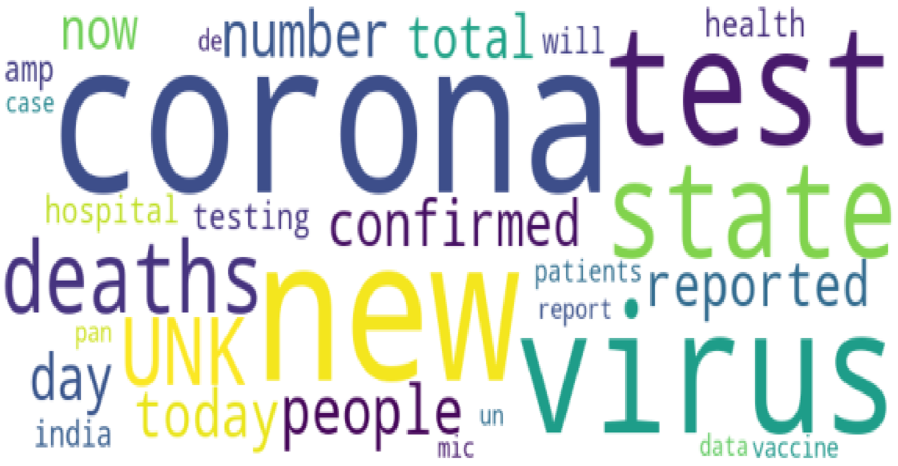

Fig. 4. Most Frequent Words in Our Prediction

4.3 Keywords in Fake News

Word count We count and sort the words in sentences which are classified as fake news by our best model to obtain keywords. For example, excluding some commonly used prepositions, some of the keywords can be visualized in Fig. 4.

Frequent words and model performance We delete those top frequent words listed above in our inputs, and see if the model performance changes after removing those words. As a result, the model performance does not change. This indicates that top frequent words do not usually sololy contribute to the overall performance.

5 Discussion

5.1 Dataset Limitation

In terms of the size of the dataset, one possible reason for the unsatisfactory performance of not frozen paramerters models could be that the dataset scale is unable to support training such large models. In this case, we could collect more fake news data from those platforms so that the model can be better trained. Besides, we could try different data set split ways to find a more reasonable one.

5.2 Model Hyperparameters

In this paper, we have tried some combinations of hyperparameters. While there is still space for trying different values of hyperparameters, such as learning rate and number of additional layers. Also we can examine different kernel size in CNN layer(s) and different number of recurrent layers in BiLSTM to further explore the model performance.

5.3 Model Structure

According to our experimental findings, the combination of not pre-trained model with additional layers could possibly improve the performance. Pre-trained model is representative for general tasks but might not for this specific case. Besides, we can pre-train our own BERT-based model using COVID-19 news corpus to make the model more specific towards our task. In addition, we could try different additional layers other than BiLSTM and CNN. One example could be GRU.

5.4 Model Evaluation

We realize that the performance of models with frozen parameters in the fine-tuned model improves, and the performance of models without frozen parameters in the fine-tuned model does not improve. The reason could be that the size of the dataset does not have enough support to learn those architectures without frozen parameters. In addition, the performances among five models are relatively similar. In order to achieve a stronger true conclusion, we need to consider use other experimental methods such as cross validation.

5.5 Keywords Evaluation

In this paper, we propose one of the possible ways to analyze keywords in COVID-19 fake news dataset. There are plenty of other ways to find keywords that contribute to the fake news detection. For example, we can find keywords by analyzing the Attention Layer in BERT. Moreover, we can also find keywords by tracing the gradient value during the backpropagation procedure.

6 Conclusion and Future Work

In this paper, we examine five BERT-based deep learning models for the COVID-19 fake news detection task. The Fine-tune BERT model is our base model, and we add additional layers at the top of the base model combined with and without frozen parameters approaches. It is shown that combining BERT with BiLSTM using freezing parameters approach yields better than the other four models and this model achieves very good results in accuracy, ROC AUC and F1 score. Additionally, we find that adding additional layers can potentially improve the models' results. While if we do not freeze parameters for Fine-tune BERT, it is relatively hard to train the model and reach a good performance.

For the future work, we can extend our dataset size and try more combinations of hyper-parameters in models. Besides, we can try other impressing additional layers other than what we use. Additionally, we can explore more ways to evaluate key words in COVID-19 fake news.

References

1. Adhikari, A., Ram, A., Tang, R., Lin, J.: Docbert: BERT for document classification. CoRR **abs/1904.08398** (2019), http://arxiv.org/abs/1904.08398
2. Aggarwal, A., Chauhan, A., Kumar, D., Mittal, M., Verma, S.: Classification of fake news by fine-tuning deep bidirectional transformers based language model. EAI Endorsed Transactions on Scalable Information Systems **7**(27) (2020)
3. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805
4. Gundapu, S., Mamidi, R.: Transformer based automatic COVID-19 fake news detection system. CoRR **abs/2101.00180** (2021), https://arxiv.org/abs/2101.00180
5. Gupta, A., Sukumaran, R., John, K., Teki, S.: Hostility detection and covid-19 fake news detection in social media. CoRR **abs/2101.05953** (2021), https://arxiv.org/abs/2101.05953
6. Kaliyar, R.K., Goswami, A., Narang, P.: Fakebert: Fake news detection in social media with a bert-based deep learning approach. Multimedia Tools and Applications **80**(8), 11765–11788 (2021)
7. Kula, S., Choraś, M., Kozik, R.: Application of the bert-based architecture in fake news detection. In: Conference on Complex, Intelligent, and Software Intensive Systems. pp. 239–249. Springer (2020)
8. Liu, C., Wu, X., Yu, M., Li, G., Jiang, J., Huang, W., Lu, X.: A two-stage model based on bert for short fake news detection. In: International Conference on Knowledge Science, Engineering and Management. pp. 172–183. Springer (2019)
9. Organization, W.H., et al.: Covid 19 public health emergency of international concern (pheic). global research and innovation forum: towards a research roadmap (2020)
10. Pham, L.: Transferring, transforming, ensembling: the novel formula of identifying fake news. In: Proceedings of the 12th ACM International Conference on Web Search and Data Mining, Melbourne, Australia. pp. 11–15 (2019)
11. Pham-Hong, B.T., Chokshi, S.: PGSG at SemEval-2020 task 12: BERT-LSTM with tweets' pretrained model and noisy student training method. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. pp. 2111–2116. International Committee for Computational Linguistics, Barcelona (online) (Dec 2020), https://www.aclweb.org/anthology/2020.semeval-1.280
12. Safaya, A., Abdullatif, M., Yuret, D.: KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. pp. 2054–2059. International Committee for Computational Linguistics, Barcelona (online) (Dec 2020), https://www.aclweb.org/anthology/2020.semeval-1.271
13. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune BERT for text classification? CoRR **abs/1905.05583** (2019), http://arxiv.org/abs/1905.05583
14. Tang, M., Gandhi, P., Kabir, M.A., Zou, C., Blakey, J., Luo, X.: Progress notes classification and keyword extraction using attention-based deep learning models with BERT. CoRR **abs/1910.05786** (2019), http://arxiv.org/abs/1910.05786
15. Vijjali, R., Potluri, P., Kumar, S., Teki, S.: Two stage transformer model for covid-19 fake news detection and fact checking. arXiv preprint arXiv:2011.13253 (2020)
16. Wikipedia contributors: Bert (language model) — Wikipedia, the free encyclopedia (2021), https://en.wikipedia.org/w/index.php?title=BERT$_{(}language_{m}odel)oldid =$ 1029082047, $[Online; accessed24 - June - 2021]$

17. Wikipedia contributors: Covid-19 — Wikipedia, the free encyclopedia (2021), https://en.wikipedia.org/w/index.php?title=COVID-19oldid=1029578763, [Online; accessed 24-June-2021]
18. Wikipedia contributors: Cross entropy — Wikipedia, the free encyclopedia (2021), https://en.wikipedia.org/w/index.php?title=Cross$_{e}ntropyoldid = 1024917567, [Online; accessed24 - June - 2021]$
19. Wikipedia contributors: F-score — Wikipedia, the free encyclopedia (2021), https://en.wikipedia.org/w/index.php?title=F-scoreoldid=1027663591, [Online; accessed 24-June-2021]
20. Wikipedia contributors: Softmax function — Wikipedia, the free encyclopedia (2021), https://en.wikipedia.org/w/index.php?title=Softmax$_{f}unctionoldid = 1024992584, [Online; accessed24 - June - 2021]$