COVID-19 Fake News Detection Using BERT

Yuxiang Wang, Yongheng Zhang[*], Xinyao Yu, Xuebo Li
Whiting School of Engineering, Johns Hopkins Unversity
ywang594@jhu.edu, yzhan470@jhu.edu, xyu63@jhu.edu, xli248@jhu.edu

Abstract

COVID-19 fake news detection has become a novel and important task in the NLP field. In this paper, we fine tune the pre-trained Bidirectional Encoder Representations from Trans-formers (BERT) model as our base model. We add BiLSTM layers and CNN layers on the top of the finetuned BERT model with frozen parameters or not frozen parameters methods respectively. The model performance evaluation results showcase that our best model(BERT finetuned 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.

1 Introduction

1.1 Background

From the past year, the whole world has gone through the COVID-19 pandemic. Twitter, Facebook, Instagram and many other social platforms update news on pandemics every day. In this project, we will use and fine tune the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model 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.

1.2 Related Work

Adhikari [1] presents the first application of BERT to document classification. Devlin [2] introduces a new language representation model called BERT. Gundapu [3] introduces an ensemble of three transformer models (BERT, ALBERT, and XLNET) to detect fake news. Pham-Hong [4] uses a stack of BERT and LSTM layers to evaluate multilingual offensive language identification in social media. Safaya [5] describes approach to utilize pre-trained BERT models with Convolutional Neural Networks for subtask of the Multilingual Offensive Language Identification shared task. Sun [6] investigates different fine-tuning methods of BERT on text classification task and provides a general solution for BERT finetuning. Tang [7] mentions keyword extraction using Attention-based Deep Learning models with BERT.

2 Methods

2.1 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

[*] The first two authors have equal contribution.

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

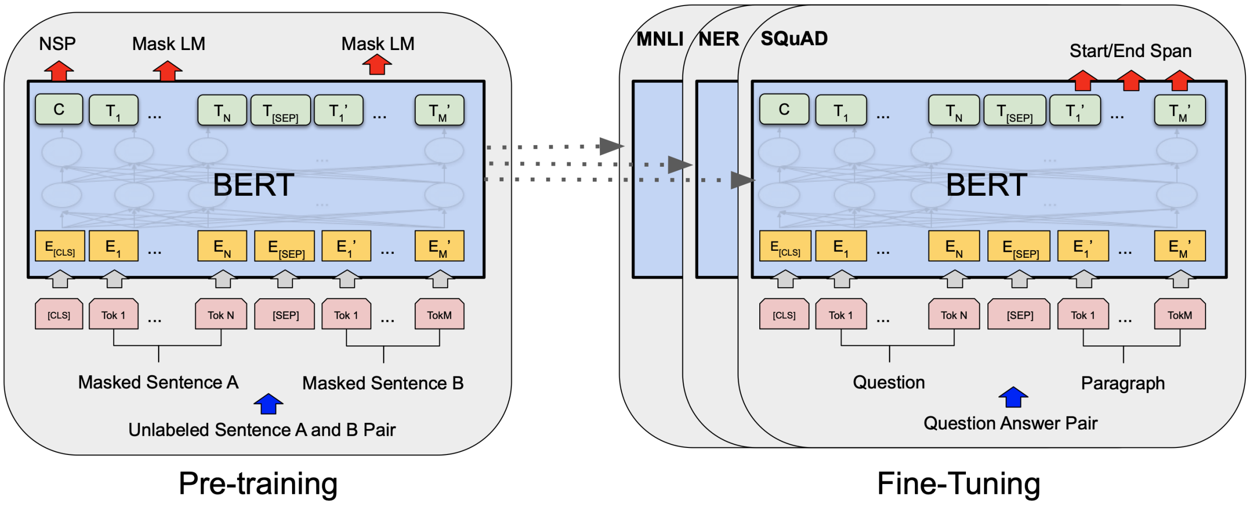

Figure 1: Architecture for the BERT Model[2]

together and randomly split data into training set (90%) and test set (10%) using the set seed method.

2.2 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 pad 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.

2.3 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 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.

3 Model

We build five different models to evaluate and compare the performance of fake news classification.We define Model 1 as the BERT finetuned model, which can be visualized in Figure 1. Next, we define Model 2 as the BERT finetuned model with frozen parameters plus CNN layer(s).Then, we define Model 3 as the BERT finetund model without frozen parameters plus CNN layer(s). Besides, we define Model 4 as the BERT finetuned model with frozen parameters plus BiLSTM layer(s). Lastly, we define Model 5 as the BERT finetuend model without frozen parameters plus BiLSTM layer(s).

3.1 Design Architecture

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

For model 2 and model 3, after the 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 apply a linear layer with softmax activation function. Besides, the learning rate for those two models are both 2e-5 and the epoch for those two models are both 4.

For model 4 and model 5, after the 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.

4 Results

4.1 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 the number of news which are correctly classified divided by the total number of news in the test dataset. The second matric 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 third matric 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 (1).

$$F1 \text{ score} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{1}$$

Table 1: Model Evaluation

Model	Test acc	Train loss	ROC AUC	F1 score
Model 1	0.9579	0.0036	0.9586	0.9607
Model 2	0.9591	0.0200	0.9589	0.9622
Model 3	0.9439	0.0211	0.9449	0.9474
Model 4	0.9614	0.0197	0.9607	0.9646
Model 5	0.9346	0.0227	0.9351	0.9389

4.2 Performance Analysis

From the results, 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.

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 Figure 2.

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 solo contribute to the overall performance.

5 Discussion

5.1 Dataset Limitation

In terms of the size of the dataset, we could collect more fake news data so that the model can be better

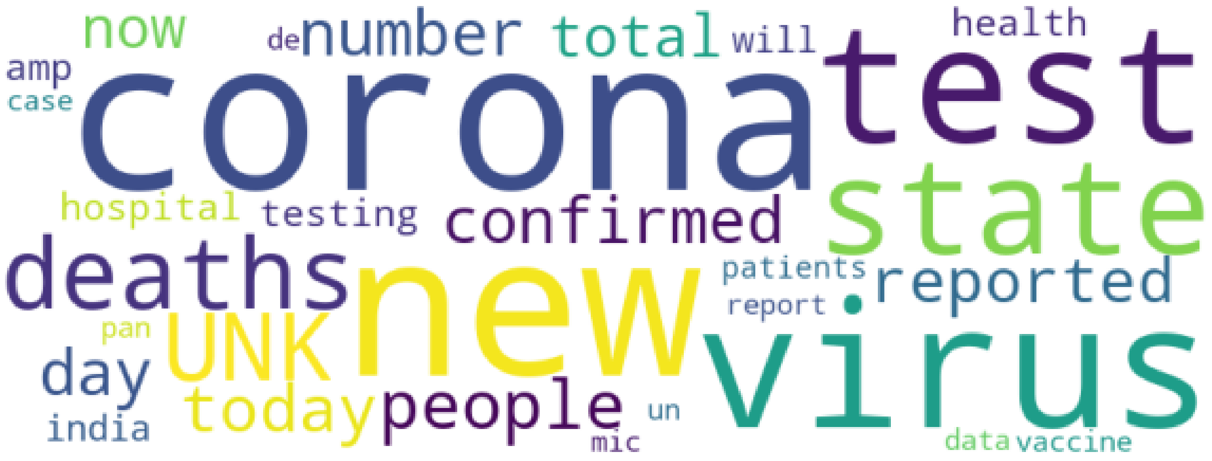

Figure 2: Most Frequent Words in Our Prediction

trained. We also could try different data set split ways to find a more reasonable one.

5.2 Model Hyperparameters

In the future, there is still space for trying different values of hyperparameters, such as learning rate and number of additional layers.

5.3 Model Structure

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. 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. The performances among five models are relatively similar. In order to achieve a stronger true conclusion, we need to consider use other methods such as cross validation.

5.5 Keywords Evaluation

There are plenty of 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.

References

[1] Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. Docbert: BERT for document classification. *CoRR*, abs/1904.08398, 2019.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[3] Sunil Gundapu and Radhika Mamidi. Transformer based automatic COVID-19 fake news detection system. *CoRR*, abs/2101.00180, 2021.

[4] Bao-Tran Pham-Hong and Setu Chokshi. 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*, pages 2111–2116, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[5] Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[6] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583, 2019.

[7] Matthew Tang, Priyanka Gandhi, Md. Ahsanul Kabir, Christopher Zou, Jordyn Blakey, and Xiao Luo. Progress notes classification and keyword extraction using attention-based deep learning models with BERT. *CoRR*, abs/1910.05786, 2019.