Detecting COVID-19-Related Fake News with Neural Networks

Minzhe Feng, Haochen Liu, Bianca Gilchrist
McGill University
{minzhe.feng, haochen.liu2, bianca.gilchrist}@mail.mcgill.ca

Abstract

Incorrect information about the currently ongoing COVID-19 pandemic has caused confusion among the general public and the spreading of fake news on social media. It has come to our attention that such misinformation has impacted individuals' responses and actions towards the pandemic. While media nowadays allow for rapid information exchange, we aim to clear the platform so that valid, creditable, and official truths can be conveyed to the general public without intervention. In our study, we develop models to perform such a task using various approaches such as bag-of-words, logistic regression, and convolutional neural networks. We have found that a BERT-based approach leads to the most outstanding classification performance of approximately 95% testing accuracy, while a hybrid CNN-RNN model also shows competitiveness against other models with a test accuracy of roughly 93.6%.

1 Introduction

1.1 Motivation

Spreading misinformation on COVID-19 is a matter of life or death. Any information regarding the number of local positive cases, the progress or effectiveness of a vaccine, and ways of coping with the pandemic should be obtained from official sources only. As social media became more accepted as a source of information, the spreading of fake news about the pandemic has boosted in both speed and size. The short texts employ lurid words to catch peoples' eyes and can be sent to other readers with just a single click. While many have the ability to criticize news and question its validity, our general public does consist of a relatively vulnerable population. This phenomenon has been a striking issue, often referred to as the "infodemic" by the media. The WHO has been taking action by collaborating with countries such as the UK to manage such matters. Vulnerable populations such as the elderly and youths can be more prone to taking impulsive actions upon receiving information. Cases of bullying and discrimination have been reported regarding potential virus transmitters and their region of residence. Elderlies who do not have access to convenient transportation may feel unnecessary desperation and helplessness when presented with anxiety-inducing information such as an exaggerated number of COVID-19 cases and severity. Till today, the number of cases of COVID-19 is still increasing as well as death cases, and the general public's life has been abnormal for nearly 2 years. Helping to identify and prevent the spread of COVID-19-related fake news with the knowledge we have is a capable way of retrieving order in society.

1.2 Approach

Under this panicking time, our general public should not carry the extra burden of mistrust. Thus, our objective is to establish models that can identify fake information regarding COVID-19, cleansing the general public's source of information. No prevarication or unofficial claims should have the chance of misleading and confusing people. NLP has proved to have its advantages over large corpora and text processing. With many pre-processing methods available at our reach, we believe this task is doable through methods of NLP. Due to the limit of resources, we are unable to train a large language model for this task. Instead, we plan to use a simple bag-of-words model as the baseline, and then perform a hybrid of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) as our target model. We will compare this hybrid model's performance with the baseline to see whether this

approach is favorable. In addition, this model will be compared with a pre-trained large language model in order to see how far we are from state-of-the-art performance.

2 Related Work

There have been researches about detecting fake news using neural network methods. There has been a convolutional neural network approach that analyzes both images and text to determine whether the new is real or fake, with two separated branch handing text and image then merging them for the final result (Yang et al., 2018). This model performs better than the plain models such as LSTM on text-based corpora and CNN on text-based corpora. Another research adopted a hybrid approach for fake news detection, using CNN to extract local features from the text, and then establishing long-term dependencies based on the results from LSTM (Nasir et al., 2021). This approach achieved over 90% accuracy on various datasets. As individuals, we do not have access to advanced hardware, but using pre-trained and fine-tune it for the targeted task is potentially doable (Turc et al., 2019). The Bidirectional Encoder Representations from Transformers (BERT) is a model that made great improvement and achieved top or state-of-the-art performance, and even surpassed human-performance on certain tasks (Devlin et al., 2018). BERT has a bidirectional

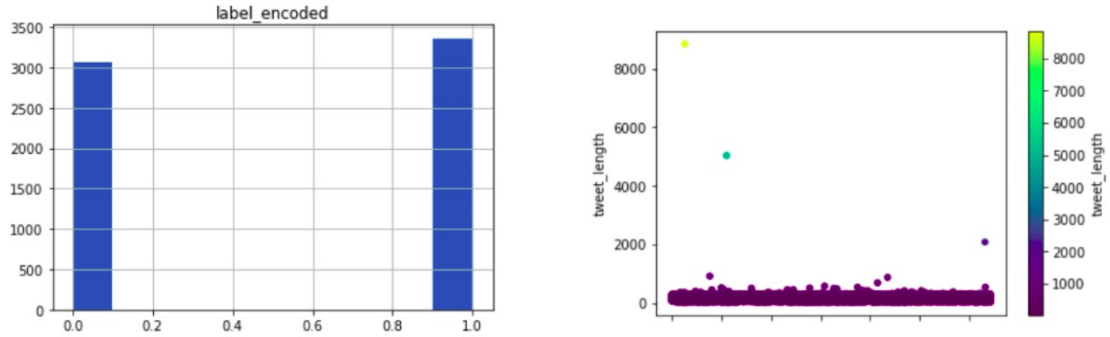

Figure 1: The distribution of label and corpus length in training set.

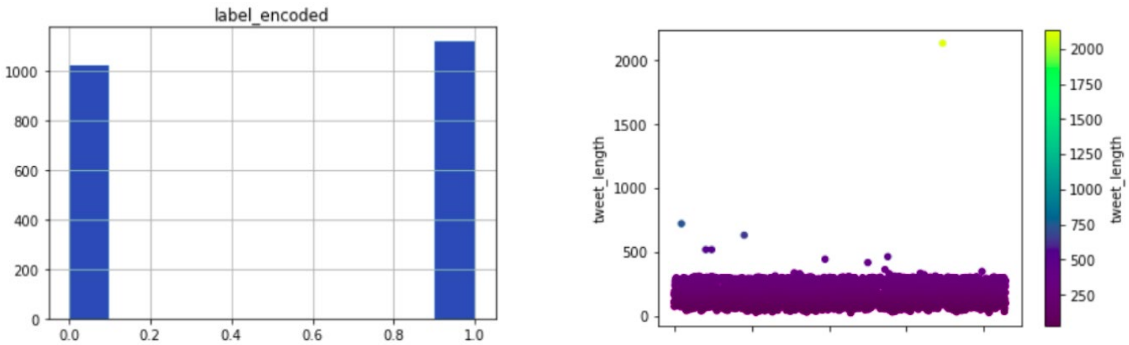

Figure 2: The distribution of label and corpus length in validation set.

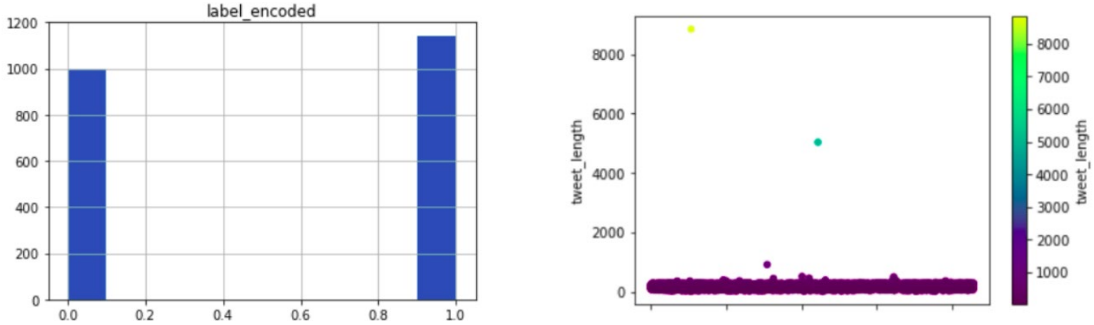

Figure 3: The distribution of label and corpus length in test set.

attention mechanism, which is based on the Transformer model (Vaswani et al., 2017).

3 Method

The dataset (Patwa et al., 2020) that we obtained contains tweets about the COVID-19 pandemic, being labeled as either real or fake. The training set consists of 6420 instances, while the validation and test set each has 2140 instances. In the total of 10700 samples, 5600 are labeled real while 5100 are fake. We prepared 3 different models to apply on the same training data and compared their performance. As a method of pre-processing, some redundant characters such as punctuations, emojis, hashtags, and URLs, are removed from each text corpora. We proceed to train the three models and compare their performances on the validation set and testing set, respectively. Some details and statistics of the dataset are displayed in Figure 1, Figure 2, and Figure 3. There are slightly more instances labeled as real compared to fake. The majority of the corpus has a length within the limit of 280 characters, but some exceeded it. We decided to keep them since the three sets all contain such outliers of extreme length.

3.1 Bag-of-Words

The simple and straightforward bag-of-words method is used in order to provide a baseline for our experiment. Compared to neural network models, bag-of-words models mainly rely on the frequency of each word but ignore the grammar and order of words in a sentence, especially for the unigram model. The cleaned corpus is converted into unigrams using CountVectorizer from the Scikit-learn package. Now, each sentence is represented by a matrix, and we use logistic regression to perform a binary classification task, with real news marked with 1, and fake news marked with 0. Since the size of the dataset that we adopted is relatively small and the content of the corpus is constrained (related to COVID-19 only), we expect this simple model to provide a solid baseline.

3.2 BERT

BERT (Bidirectional Encoder Representation from Transformers) is a machine-learning technique for natural language processing. It was published in 2018 by Google. With the transformer model being its core, the architecture also contains several encoding layers and self-attention heads. The transformer model is based on the encoder-decoder architecture, with each layer of encoder containing a self-attention layer and a feed-forward neural network, and each layer of decoder having an extra attention layer between them. Compared to RNN, CNN, and many other famous models such as ELMo (Embeddings from Language Models) and ULMFiT (Universal Language Model Fine-tuning), BERT is deeply bidirectional, which means that the meaning of a word can be determined by context on the left side and the right side rather than just the left side (unidirectional), Ewhich reduced the chances of having ambiguity. It is worth noting that BERT can also execute concurrently, extracting the relationship features of a word in the sentence on different levels, thus the semantics of a sentence can be better represented. BERT models have achieved state-of-the-art performance on various datasets, The purpose of using BERT in our study is to provide a potential higher bound for comparison. In the experiment, we followed the guideline of Dale Markowitz and applied the BERT-base model (Devlin et al., 2018) to the dataset. After setting up the pre-trained model, we trained it with the preprocessed training dataset and then evaluated its performance on validation and test set, respectively.

3.3 Hybrid CNN - RNN

CNN (Convolutional Neural Network) is a popular neural network architecture that relies on the use of "filters" to convolve through input data and works great on extracting local features from structured data, making it a great tool for tasks such as interpreting visual data. However, the existence of long-range dependencies renders CNN alone insufficient for modeling sequential text data. Thus, RNN (Recurrent Neural Network), with its ability to process sequential inputs and retain long-term memory of what comes before the part being processed, becomes a great complement to make up for this exact deficiency of CNN.

The combination of CNN and RNN has been proven successful in several classification and regression tasks since they can capture both local and sequential characteristics of input data. In practice, after producing word embedding through a dedicated embedding layer, a CNN layer is first applied to extract local features from the data. An

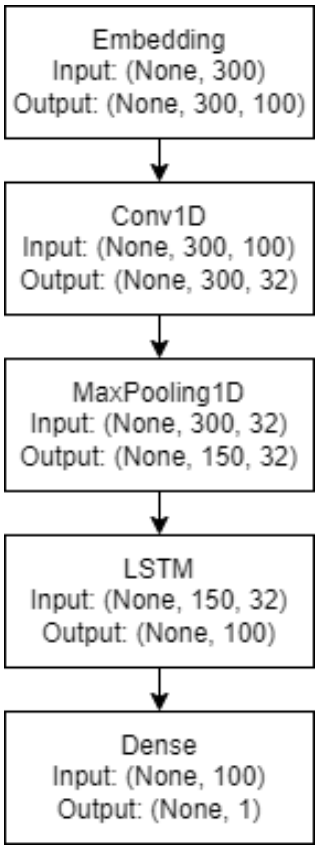

Figure 4: The architecture of the hybrid CNN-RNN model

RNN layer is then employed to capture the required long-range dependencies.

Our experimentation consists of two parts. First, combinations of various pre-processing techniques like stemming, lemmatization & removing stop words are benchmarked, and the best-performing configuration of removing stop words & special characters without any stemming or lemmatization) is selected. Then, our hybrid CNN-RNN model (architecture given below) is applied to the processed & tokenized data. With some minor tweaks in network architecture and parameters, our final setup manages to achieve compelling performance on distinguishing real covid-related news from fake ones.

4 Results

To our surprise, the unigram bag-of-words model achieved accuracy above 0.92 on both the validation and test set. The BERT-based model achieved the best performance on the test set, with an accuracy of 0.9519. The model we adopted also performed better than the baseline bag-of-words model, with a test accuracy of 0.9360.

	Training	Validation	Test
Bag-of-Words	0.9914	0.9280	0.9206
BERT-based	0.9986	0.9575	0.9519
Hybrid CNN-RNN	0.9829	0.9262	0.9360

Table 1: Results of 3 models

5 Discussion and Conclusion

5.1 Focusing on Result

The surprisingly competent performance of our Bag-of-Words approach, consisting basically of a unigram model only, might stem from the fact that covid-related fake news tends to be particularly polarizing, which in turn leads to word choice easily distinguishable from that of real news. Thus, this approach might not generalize well with text data from a different context.

The exceptional capacities of BERT should not come as a surprise. First, it has long been a tried-and-true model for dealing with text data. Second, we believe that the self-attention mechanism of BERT ensures that it understands the meaning of words better. The bidirectional feature further improves this advantage. When performing tasks on larger corpora with longer texts, the advantages of these features will be displayed more clearly when compared with traditional CNN and RNN models.

As for the Hybrid CNN-RNN model, the simple motive of combining them to capture both local features and long-range dependencies really pays off in practice, as can be seen from its prediction results. Even though it is still not as accurate as BERT-base, the small gap on this dataset can be filled up with more fine-tuning. The randomness of dropout sometimes leads to a better result, and sometimes it does not, but the results are usually above the baseline. Comparing this model to other models, it finds a middle ground of model complexity, but performance is just a little not as

good as the most outstanding one, which means that it should be the relatively efficient model.

In summary, detecting COVID-19-related fake news from social media such as Twitter is not as hard as we believed since a unigram bag-of-words model can have an accuracy of over 90%. Models with neural network approaches generally perform better than simpler models. However, pre-trained large language models such as BERT still have a great advantage over others since they are trained on larger datasets. For example, when the task is to classify more generalized fake news rather than just focusing on COVID-19, the pre-trained BERT models will be more likely to outperform other models, but that will be another story. With these methods adopted, we believe that the spread of COVID-19-related fake news on social media will be successfully identified and eliminated, which would potentially push this pandemic to an end.

5.2 Future Leads

All of our models have adopted the same pre-processing methods, though it was not discussed in our study, we hope to explore additional pre-processing techniques that may have a superior impact on such classification models.

Through the comparison of our three models, we have discovered a particularly outstanding approach to identifying the validity of COVID-related news. In the future, we wish to apply the BERT-based model on similar classification tasks but with a differing theme. For example, prenatal education alone is a field from which many jobs and careers have derived, it has been noted that an increasing amount of advertisement and social influencers have emerged. Such activities are often profit-making and specifically target inexperienced parents. With the increasing variety of non-traditional sources to absorb prenatal information, it is important that young parents receive the correct guidance on how to endure and adapt to this stage.

6 Statement of Contributions

We have distributed our work evenly. Minzhe Feng focused on implementing and testing the hybrid CNN-RNN model. Haochen Liu contributed to the baseline bag-of-words model and the BERT-base model according to the guideline (Devlin et al., 2018) mentioned above. Bianca Gilchrist composed our research proposal, implemented the pre-processing of texts, and searching through related papers.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805. Version 2

Jamal A. Nasir, Osama S. Khan, and Iraklis Varlamis. 2021. *Fake news detection: A hybrid CNN-RNN based deep learning approach. International Journal of Information Management Data Insights,*vol. 1, no. 1, https://doi.org/10.1016/j.jjimei.2020.100007.

Parth Patwa, Mohit Bhardwaj, Vineeth Guptha, Gitanjali Kumari, Shivam Sharma, Srinivas PYKL, Amitava Das, Asif Ekbal, Shad Akhtar, and Tanmoy Chakraborty. 2021. *Overview of CONSTRAINT 2021 Shared Tasks: Detecting English COVID-19 Fake News and Hindi Hostile Posts. Proceedings of the First Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation,* 2021, pp. 42-53.

Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md S. Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. *Fighting an Infodemic: COVID-19 Fake News Dataset*. arXiv: 2011.03327. Version 4

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *Well-Read Students Learn Better: On the Importance of Pre-training Compact Models.* arXiv:1908.08962. Version 2

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *Attention Is All You Need.* arXiv:1706.03762.

Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S. Yu. 2015. *TI-CNN: Convolutional Neural Networks for Fake News Detection.* arXiv:1806.00749