OCHADAI at SMM4H-2021 Task 5: Classifying self-reporting tweets on potential cases of COVID-19 by ensembling pre-trained language models

Ying Luo[1], **Lis Kanashiro Pereira**[1], **and Ichiro Kobayashi**[1]

[1]Ochanomizu University, Japan

1 Introduction

Since the outbreak of coronavirus at the end of 2019, there have been numerous studies on coronavirus in the NLP arena. Meanwhile, Twitter has been a valuable source of news and a public medium for the conveyance of information and personal expression. This paper describes the system developed by the Ochadai team for the Social Media Mining for Health Applications (SMM4H) 2021 Task 5, which aims to automatically distinguish English tweets that self-report potential cases of COVID-19 from those that do not. We proposed a model ensemble that leverages pre-trained representations from COVID-Twitter-BERT (Müller et al., 2020), RoBERTa (Liu et al., 2019), and Twitter-RoBERTa (Glazkova et al., 2021). Our model obtained F1-scores of 76% on the test set in the evaluation phase, and 77.5% in the post-evaluation phase.

2 System Overview

In this section, we overview the pre-processing steps, pre-trained language models and training prodedure used by our system.

2.1 Text pre-processing

We follow (Müller et al., 2020) for pre-processing the dataset. First, we lowercase the text. Then, we replace user tags (e.g. @ScottGottliebMD) with the token "@USER", and replace urls with the token "URL". All unicode emoticons are replaced with textual ASCII representations (e.g. dog for 🐕) using the Python emoji library [1]. We also remove the unicode symbols (e.g. & for &), control characters and accented characters (e.g. shyapu for shyápu).

[1]https://pypi.org/project/emoji/

2.2 Pre-trained Models

We mainly experimented with three transformer-based pre-trained language models as follows:

COVID-Twitter-BERT (CT-BERT) (Müller et al., 2020): This is a $\text{BERT}_{\text{LARGE}}$ model trained on a large corpus of Twitter messages on the topic of COVID-19, collected during the period from January 12 to April 16, 2020.

$\text{RoBERTa}_{\text{LARGE}}$ (Liu et al., 2019): We use the $\text{RoBERTa}_{\text{LARGE}}$ models released by the authors. Similar to $\text{BERT}_{\text{LARGE}}$, $\text{RoBERTa}_{\text{LARGE}}$ consists of 24 transformer layers, 16 self-attention heads per layer, and a hidden size of 1024.

Twitter-RoBERTa (Glazkova et al., 2021): This is a $\text{RoBERTa}_{\text{BASE}}$ model pre-trained on a large corpus of English tweets. This corpus includes tweets from 2020, possibly covering the COVID-19 topic as well.

2.3 Training Procedure

We fine-tuned each pre-trained language model on the training set with 5-fold cross-validation. We ran each model using three different random seeds, and selected the best performing model on the validation set or averaged the prediction probabilities obtained after softmax. Then, we further combined the outputs of the models generated by each fold by again taking an average of the prediction probabilities obtained after softmax. We also experimented on max-voting on the predicted labels.

We further experimented on ensembling CT-BERT, RoBERTa-large, and Twitter-RoBERTa.

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 123–125

Index	Training Models	Ensemble			Cross Validation	Ensemble Method		F1-Score	
						average probability	max voting	Validation	Test
1	Covid-Twitter-BERT							78.00	
2	Covid-Twitter-BERT				✔		✔	78.40	
3	Covid-Twitter-BERT	★	★		✔			92.00	
4	Covid-Twitter-BERT	★	★	★	✔	✔		79.00	
5	Twitter-RoBERTa-base							86.00	
6	Twitter-RoBERTa-base				✔		✔	92.00	
7	Twitter-RoBERTa-base				✔	✔		**93.00**	
9	RoBERTa-large							83.00	
10	RoBERTa-large				✔		✔	92.00	
11	RoBERTa-large				✔	✔		**93.00**	
12	*Twitter-RoBERTa-base+Covid-Twitter-BERT+RoBERTa-large*	★	★			✔		94.00	
13	*Twitter-RoBERTa-base+Covid-Twitter-BERT+RoBERTa-large*	★		★	✔	✔		77.00	
14	Ensemble 1*	○					✔	**97.00**	76.00
15	Ensemble 2 ‡		○			✔		93.00	**77.50**
16	Ensemble 3 ‡			○			✔	93.00	76.67

Table 1: Comparison of different text encoders and different ensemble methods. Best results are highlighted in bold. ★ indicates each model that was used in the Ensemble 1, Ensemble 2, and Ensemble 3 models, respectively indicated in each column by ○. * indicates the models submitted during the evaluation phase, and ‡ indicates the models submitted during the post-evaluation phase.

3 Experiments

3.1 Implementation Details

In this work, we used the PyTorch implementation released by huggingface[2] of $\text{RoBERTa}_{\text{LARGE}}$, Covid-Twitter-BERT, and Twitter-RoBERTa. We used AdamW as our optimizer, with a learning rate in the range $\in \{9 \times 10^{-6}, 1 \times 10^{-5}, 2 \times 10^{-5}\}$ and a batch size $\in \{16, 32\}$. The maximum number of epochs was set to $\in \{5, 10\}$. A linear learning rate decay schedule with warm-up over 0.01 was used. All the texts were tokenized using wordpieces and were chopped to spans no longer than 512 tokens.

The performance of the models were measured in terms of F1-score, and the model with the highest performance on the validation set was selected.

3.2 Main Results and Analysis

Our results are shown in Table 1. First, we observe that performing cross-validation and averaging the results of each fold yields to better performance on the validation set than max-voting. For instance, the Covid-Twitter-BERT could improve the F1-score from 78.00% to 78.40% to 79.00% on the validation set in lines 1,2,4 of the table. The same tendency can be observed on the F1-score of the validation set in the Twitter-RoBERTa-base (from 86% to 92% and 93% in lines 5,6,7 of the table) and RoBERTa-large (from 83% to 92% and 93% in lines 9,10,11 of the table) models. Moreover, another observation is that combining the outputs of models by taking an average of the prediction probabilities obtained after softmax instead of max-voting on the predicted labels leads to higher performance on the validation set. For instance, the improved F1-score on validation set was observed from the table in the Twitter-RoBERTa-base (from 92% to 93% in lines 6,7 of the table) and RoBERTa-large (from 92% to 93% in lines 10,11 of the table) models.

Finally, ensembling different pre-trained models leads to better performance on the test set. For instance, the Covid-Twitter-BERT model submitted in the evaluation phase obtained an F1-score of 69% which is not referred in the table, while the Ensemble 1, Ensemble 2, and Ensemble 3 models obtained F1-scores of 76%, 77.5%, and 76.66%, respectively.

4 Conclusion

We presented the Ochadai system submitted to the SMM4H-2021 Task 5. We proposed an ensemble model that leverages pre-trained representations from COVID-Twitter-BERT (Müller et al., 2020), RoBERTa (Liu et al., 2019), and Twitter-RoBERTa (Glazkova et al., 2021). Our best performing model obtained an F1-score of 77.5%. In future efforts, we plan to further improve our model by exploring other pre-trained language models and ensemble techniques.

[2]https://huggingface.co/models

References

Anna Glazkova, Maksim Glazkov, and Timofey Trifonov. 2021. g2tmn at constraint@aaai2021: Exploiting ct-bert and ensembling learning for covid-19 fake news detection.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter.