Wall-NUT: An Ensemble of BERT-based Models Approach to Identifying Informative COVID-19 English Tweets

Thai Hoang
University of Washington
qthai912@cs.washington.edu

Phuong Vu
University of Rochester
pvu3@u.rochester.edu

Abstract

As of 2020 when the COVID-19 pandemic is full-blown on a global scale, people's need to have access to legitimate information regarding COVID-19 is more urgent than ever, especially via online media where the abundance of irrelevant information overshadows the more informative ones. In response to such, we proposed a model that, given an English tweet, automatically identifies whether that tweet bears informative content regarding COVID-19 or not. Using primarily an ensemble of BERT-based models, we have achieved competitive results that are only shy of those by top performing teams by less than 0.1%. In the post-competition period, we have also experimented with various other approaches that potentially boosts generalization to a new dataset.

1 Introduction

People use social network a lot =¿ Can help spreading information in the case of a natural disaster/calamity =¿ May self-built crow sourcing tools rely on social networks =¿ Highlight the need for an automatic system : informative/uninf.

The task we try to tackle in this paper is da Shared task 2: bla bla by bla bla. Goal of the challenge is : ... (cite here).

Informativeness is defined as ¡...¿. All other =¿ uninformative. Help people to filter the giant mess of tweets they encounter everyday.

Nowadays people use social network a lot. Besides serving as a platform for various types of entertainment, social media is particularly helpful in spreading information, and we can leverage such to keep the majority of its user well-informed of the most updated news amidst a natural disaster or calamity. During the height of COVID-19, have been built (e.g. The John Hopkins Coronavirus Dashboard) as a means to trace and record the development of the outbreak. These systems rely on crowdsourcing and manual search for updates.

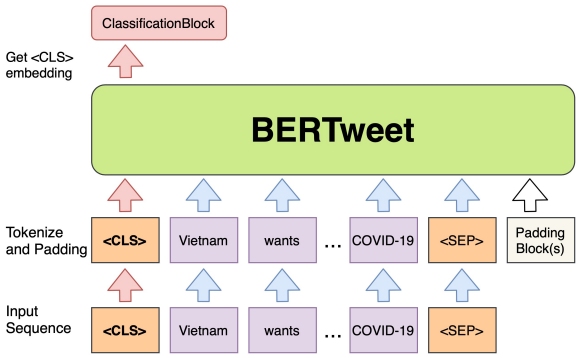

Figure 1: An overview of our model for identify Informative COVID-19 English Tweets

2 Related work

- THAI - Basic classifier: SVM, LR. - BERT
- XLNet
- Roberta

3 System Description

We use the pre-trained language model BERTweet as the core for our system. To accomplish the task of identifying Informative/Uninformative COVID-19 English tweets, we attach a classification block on top of our Transformer block. Figure 1 indicates the high level detail of our system.

3.1 BERTweet

BERTweet (**?**) is a large-scale language model pre-trained for English Tweets. Because of its nature of being a domain-specific model, BERTweet has achieved state-of-the-art performances on many downstream Tweet NLP tasks of Part-of-speech tagging, Named entity recognition and text classification, outperformed top models such as RoBERTa-base (**?**) and XLM-R-base (**?**). Trained on 845M

Tweets streamed from 01/2012 to 08/2019 and 5M Tweets related the COVID-19 pandemic as pre-training resources, BERTweet has an advantage compares to other models for our particular task of classifying COVID-19 related English Tweets.

3.1.1 Embedding Extractions

Each Transformer layer learns different information. We experiments different ways of extracting the pooled token from our BERTweet model to analyze the performance on this downstream task. More detail would be discussed in the "Experiments" section.

3.1.2 Global Local BERTweet

Due to the informal nature of writing Tweets, many tweets have noteworthy information at particular part of the tweets. Therefore, besides reading the whole Tweets, paying more attention to local parts of the Tweets is also important. Inspired by that idea, we propose a method to concurrently training 3 BERTweet models: one for reading the whole sequence, one for reading the first part of the sequence, and one for reading the remaining part. The pooled token from each model would then extracted and concatenated together for the system to learn both global and local information of the Tweets.

3.2 Classification Block

The classification block contains one or more linear layers stacked onto each other. The final layer is then used to classify whether a Tweet is informative or not.

4 Experiments

4.1 Datasets

We use the Dataset released by the competition organizer, consisting of 10,000 COVID-19 English Tweet. Each Tweet in the dataset is annotated by 3 annotators independently, and the overall inter-annotator agreement score of Fleiss' Kappa is 0.818. The dataset is then divided into 3 distinct set for training, validation, and testing, with the ratio of 70/10/20, respectively.

Table 1 shows the division of the dataset.

4.1.1 Re-splitting Data

During the Evaluation Phrase, we re-splitting our dataset by combining Training and Validation sets then dividing randomly with the ratio of 90/10.

	Informative	**Uninformative**
Training Set	3303	3697
Validation Set	472	528
Test Set	944	1056

Table 1: Dataset

	Informative	**Uninformative**
Training Set	...	...
Validation Set	...	...

Table 2: Dataset

Table 2 shows the division of the dataset (not including the Test set).

4.2 Implementation

4.2.1 Main Library and Framework

We use the `transformers` library (**?**) with `PyTorch` framework (**?**) to run our codes.

4.2.2 Two-Phrases Training

During Training progress, we follow the Two-phrases training, in which we freeze all BERTweet parameters during the first phrase and start with high learning rate to focus on training the Classification block. Because the Classification block is a combination of Linear layers, the training stage only takes a small amount of time to reach convergence.

4.2.3 Optimizer

Phuong: ADAM.

4.2.4 Hyperparameters Configuration

By examining the dataset, we observe that the longest sequence in the dataset has 89 words and 110 tokens (VERIFY). Therefore, we set the max length for padding and truncating before feeding into the BERTweet model to be 256, which helps decreasing training time and memory used.

We train our models on 1 NVIDIA DGX V100 and 1 NVIDIA RTX 2080. To fit the machine, we alternatively use batch size of 16 and 32.

Phuong: Within the Two-phrase training that we have discussed above, we use 10-12 epochs to train the classification block with the learning rate of 5^{-4} and 4-6 epochs to finetune the whole system with the learning rate of 10^{-5}.

Model	F1
1	0.9028

Table 3: Caption

4.3 Model Performance

4.3.1 Baselines

Phuong

4.3.2 BERTweet with different embedding selections

We experiment different ways to extract embedding from the Transformer model. Table 3 shows the evaluation of these implementation.

The reported results are the F1 Score of predictions on the original validation set.

4.3.3 Ensembling

Phuong

4.4 Additional Works

...

5 Future work

6 Conclusion

PHUONG