CIA_NITT at WNUT-2020 Task 2: Classification of COVID-19 Tweets Using Pre-trained Language Models

Yandrapati Prakash Babu
Department of Computer Applications
NIT Trichy, India
prakash.babu23@gmail.com

Rajagopal Eswari
Department Computer Applications
NIT Trichy, India
eswari@nitt.edu

Abstract

This paper presents our models for WNUT 2020 shared task2. The shared task2 involves identification of COVID-19 related informative tweets. We treat this as binary text classification problem and experiment with pre-trained language models. Our first model which is based on CT-BERT achieves F1-score of 88.7% and second model which is an ensemble of CT-BERT, RoBERTa and SVM achieves F1-score of 88.52%.

1 Introduction

As of September 07,2020 COVID-19 Coronavirus infected 27.3M people and caused 887K deaths[1]. Real time updates regarding the number of infected cases and death cases is given in dashboards. These dashboards make use of information from social networking sites like twitter. As majority of the tweets posted online are uninformative, it is necessary identify the informative tweets which include useful information related to recovered, suspected, confirmed and death cases as well as location or travel history of the cases.

The WNUT 2020 shared task2 involves identification of informative tweets. We treat this as binary text classification problem. Prior to 2018, most of the text classification models are based on Convolutional Neural Network (CNN) or Recurrent Neural Network(RNN). These models are shallow in nature and cannot learn more informative features from the input. Moreover as these models are to trained from scratch, they require more number of training instances (Kalyan and Sangeetha, 2020a,b).

Recently pre-trained language models like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) achieved significant improvements in many of the natural language processing tasks (Qiu et al., 2020). BERT is a transformer encoder based language model trained using 16 GB text corpus using language modeling and next sentence prediction objectives. The 16GB text corpus includes 3.5B words from Wikipedia articles and 0.8B words from Books. BERT model is available in two versions namely BERT-base (consists of 12 transformer encoder layers with 768 hidden vector size) and BERT-large (consists of 24 transformer encoder layers with 1024 hidden vector size). As BERT models are trained using generic less noisy text corpus, these may not be effective for noisy text like tweets. Moreover,these models don't include any domain specific information. A common strategy is to adapt BERT model to a specific domain is to further pre-train the model or train the model from scratch using domain specific text.

In this paper, we propose two models to identify informative COVID-19 tweets. First model is based on Covid-Twitter-BERT (CT-BERT) which is a BERT-Large based model which is further trained on 160M Corona virus related tweets (Müller et al., 2020). Second model is ensemble of CT-BERT, RoBERTa and SVM (Islam et al., 2017). As CT-BERT is initialized from BERT-large weights and further pre-trained on COVID tweets, it has two advantages compared to BERT-large which is pre-trained on generic less noisy texts. First advantage is, CT-BERT includes domain as well as specific information and second advantage is, CT-BERT can better handle noisy texts like tweets. Our CT-BERT based model achieves F1-score of 88.87% and ensemble model achieves F1-score of 88.52%

2 Related work

Text classification is one of the core NLP tasks. It involves assigning labels to text sequences like phrases, sentences or documents. It has applica-

[1] https://www.worldometers.info/coronavirus/

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 471–474
Online, Nov 19, 2020. ©2020 Association for Computational Linguistics

Label	Training	Validation	Test
INFORMATIVE	3303	472	944
UNINFORMATIVE	3697	528	1056

Table 1: Original Split of Dataset

Label	Training	Validation
INFORMATIVE	3024	751
UNINFORMATIVE	3376	849

Table 2: After Splitting the Dataset

tions in various NLP tasks like sentiment analysis, spam classification, abusive text detection etc (Minaee et al., 2020). The use of deep learning models for text classification started with using models like Convolutional Neural Network or Recurrent Neural Network (Kim, 2014; Nowak et al., 2017). These models are used on the top of word embeddings. To over the issue of Out of Vocabulary (OOV) words, char level CNN or RNN are used (Zhang et al., 2015). As these models are shallow in nature and need to be trained from scratch, it requires more number of training instances to train these models. Recently, with the introduction of deep pre-trained language models like BERT, RoBERTa , there is no need to train the downstream model from scratch. To adapt the model to downstream task, it is enough to add task specific layers and fine-tune the model for few epochs (Devlin et al., 2019; Liu et al., 2019).

3 Methodology

3.1 Dataset and Pre-Processing

The dataset contains 20K tweets each of which is labeled as 0 (uninformative tweet) or 1 (informative tweet) (Nguyen et al., 2020). The dataset is divided into train, validation and test sets(Actual 2k test tweets mixed with the 10K tweets), total test set size 12K. The statistics of the original dataset is reported in Table 1 and the dataset splitted into 80% and 20% reported in Table 2.

As tweets are noisy in nature, we do the following pre-processing steps

- remove unnecessary punctuation and non-ASCII characters.
- standardize words with repeating characters (e.g. coooool → cool)
- replace emoji characters with their text descriptions[2]
- replace interjection words with their meanings (e.g. oww → pain)
- replace contraction with full form (e.g., I'm → I am)
- replace twitter slang words with related words (e.g., 2morrow → tomorrow

3.2 Model Description

We treat the problem of identification of informative tweets as binary text classification. Following the recent trend of using pre-trained language models in NLP, we propose models based on BERT and RoBERTa.

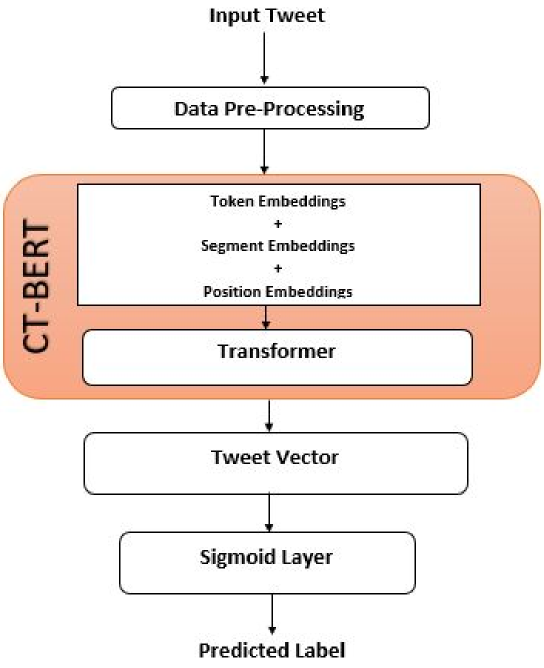

Figure 1: Overview of Model-1

Model-1 This model is based on COVID-Twitter-BERT (CT-BERT). CT-BERT is initialized from BERT-Large weights and further pre-trained on 160M Corona virus related tweets. As it is binary classification, a fully connected sigmoid layer is included on the top of CT-BERT. The entire model (CT-BERT + fully connected sigmoid layer) is then fine-tuned using the training dataset. The original tweet is added with the special tokens [CLS] and [SEP] and then tokenized using word-piece tokenizer. The embedding of each token is obtained by the summation of word-piece, position

[2] We gather list of emojis and corresponding descriptions from https://emojipedia.org/

and segment embeddings. A sequence of 24 transformer encoder layers is applied on these token embeddings to get the final hidden state vectors. Following , we treat $e_t \in R^h$ the final hidden vector of [CLS] token as the representation of tweet. Then, e_t is passed through fully connected sigmoid layer to get the required label $\hat{p} \in [0, 1]$(as shown in figure 1).

$$e_t = CTBERT(tweet) \tag{1}$$

$$\hat{p} = Sigmoid(W^T e_t + b) \tag{2}$$

Model-2 This model is ensemble of CT-BERT, RoBERTa and TF-IDF with SVM. In this model we used base model of Roberta and TF-IDF is used for to extract the features from the tweets which were used in the SVM. Each model is individually trained using the training set. In case of CT-BERT and RoBERTa, task-specific classifier layer having fully connected sigmoid layer is added and the entire model is fine-tuned. In case of SVM, the model is trained using the tf-idf vectors of training tweets and we use kernel as sigmoid. The final prediction is obtained from the average of predictions of all these models(as shown in figure 2).

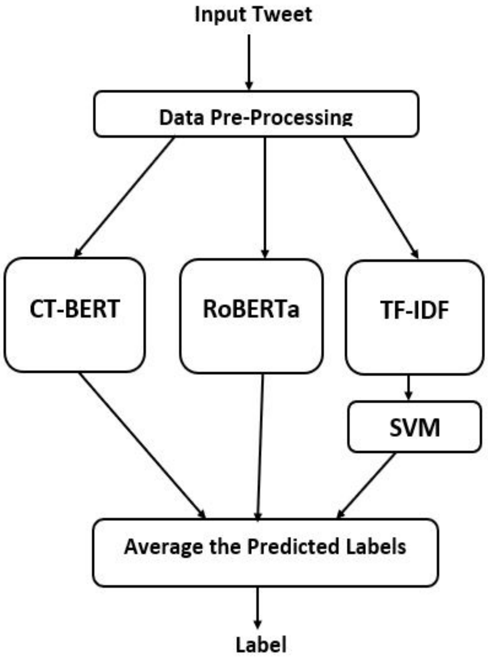

Figure 2: Overview of Model-2

Model	F1 Score	Precision	Recall
CT-BERT	96.03	93.8	98.26
RoBERTa	93.85	92.16	95.60
TFIDF+SVM	82.90	87.90	79.09
CT-BERT+ RoBERTa	95.64	96.35	94.94
CT-BERT+ (TFIDF+SVM)	85.62	97.12	76.56
RoBERTa+ (TFIDF+SVM)	84.94	95.66	76.43
CT-BERT+ RoBERTa+ (TFIDF+SVM)	95.68	93.96	97.47

Table 3: F1-score, Precision, and Recall on Validation data

Model	F1 Score	Precision	Recall
CT-BERT	95.17	92.14	98.40
CT-BERT+ RoBERTa+ (TFIDF+SVM)	95.31	94.50	96.13

Table 4: F1-score, Precision, and Recall of proposed models on Validation data without using preprocessing steps

3.3 Evaluation Metrics

The model is officially evaluated using precision, recall and F1-score metrics.

$$Precision = \frac{T_{positive}}{T_{positive} + F_{positive}}$$

$$Recall = \frac{T_{positive}}{T_{positive} + F_{negative}}$$

$$F1 - Score = 2X\frac{Precision * Recall}{Precision + Recall}$$

3.4 Implementation Details

Task organizers provided training and validation sets with labels . We merged both training and validation set and split into 80% train and validation sets with 80% and 20% of instances. We set batch size = 32, learning rate = 3e-5 and epochs=3 after doing random search over the hyperparameter space. All our models are implemented using tranformers library in PyTorch (Wolf et al., 2019).

4 Results

To identify informative tweets related to Corona virus, we experimented with two models. First

Model	F1 Score	Precision	Recall
CT-BERT	88.87	87.72	90.04
CT-BERT+ RoBERTa+ (TFIDF+SVM)	88.52	89.24	87.82

Table 5: F1-score, Precision, and Recall of proposed models on Test data

model is based on CT-BERT and second model is an ensemble of CT-BERT, RoBERTa and SVM with TF-IDF. The pre-processing steps improved the results. The performance on the validation set using pre-processing steps is reported in Table 3, the performance on the validation set without using pre-processing steps is reported in Table 4 and on test set is reported in Table 5. As reported in Table 3, a) SVM with TF-IDF features performed poorly. This is because as text is noisy, TF-IDF based features are less informative. b) CT-BERT outperformed RoBERTa as CT-BERT is trained on COVID- 19 related tweets. As reported in Table 5, CT-BERT based model achieved F1-score of 88.87% and ensemble model achieved F1-score of 88.52%. From the Table 5, it is clear that CT-BERT based model achieved slightly better results compared to ensemble model.

5 Conclusion

In this work, we present our models to identify COVID-19 related informative tweets. We treat this as binary text classification problem. We propose two models based on pre-trained language models for this task. Our model based on CT-BERT achieved F1-score of 88.87%.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

M. S. Islam, F. E. M. Jubayer, and S. I. Ahmed. 2017. A support vector machine mixed with tf-idf algorithm to categorize bengali document. In *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 191–196.

Katikapalli Subramanyam Kalyan and S Sangeetha. 2020a. Bertmcn: Mapping colloquial phrases to standard medical concepts using bert and highway network. Technical report, EasyChair.

Katikapalli Subramanyam Kalyan and S Sangeetha. 2020b. Secnlp: A survey of embeddings in clinical natural language processing. *Journal of biomedical informatics*, 101:103323.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.11692*.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2020. Deep learning based text classification: A comprehensive review. *arXiv preprint arXiv:2004.03705*.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Dat Quoc Nguyen, Thanh Vu, Afshin Rahimi, Mai Hoang Dao, Linh The Nguyen, and Long Doan. 2020. WNUT-2020 Task 2: Identification of Informative COVID-19 English Tweets. In *Proceedings of the 6th Workshop on Noisy User-generated Text*.

Jakub Nowak, Ahmet Taspinar, and Rafał Scherer. 2017. Lstm recurrent neural networks for short text and sentiment classification. In *International Conference on Artificial Intelligence and Soft Computing*, pages 553–562. Springer.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *arXiv preprint arXiv:2003.08271*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.