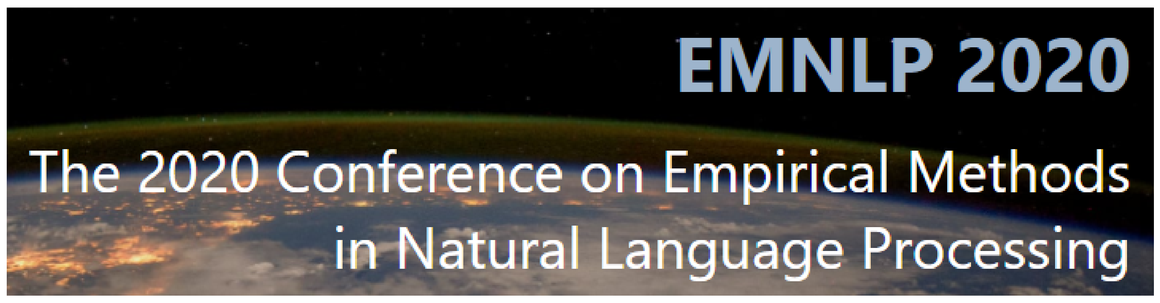


CIA_NITT at WNUT-2020 Task 2: Classification of COVID-19 Tweets Using Pre-trained Language Models

Yandrapati Prakash Babu; Rajagopal Eswari

National Institute of Technology Trichy, India

Introduction

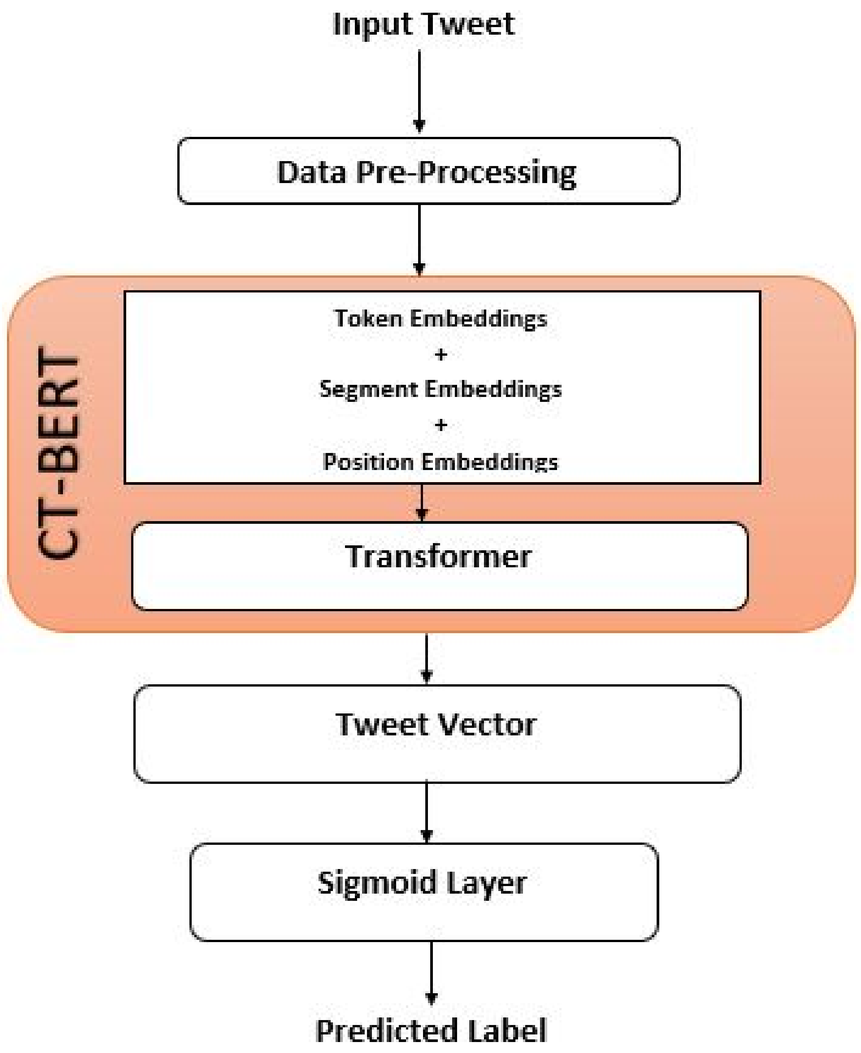
This task involves identification of COVID-19 related informative tweets. We treat this as binary text classification problem and experiment with pre-trained language models. Our first model which is based on CT-BERT achieves F1-score of 88.7% and second model which is an ensemble of CT-BERT, RoBERTa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

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.
- 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.

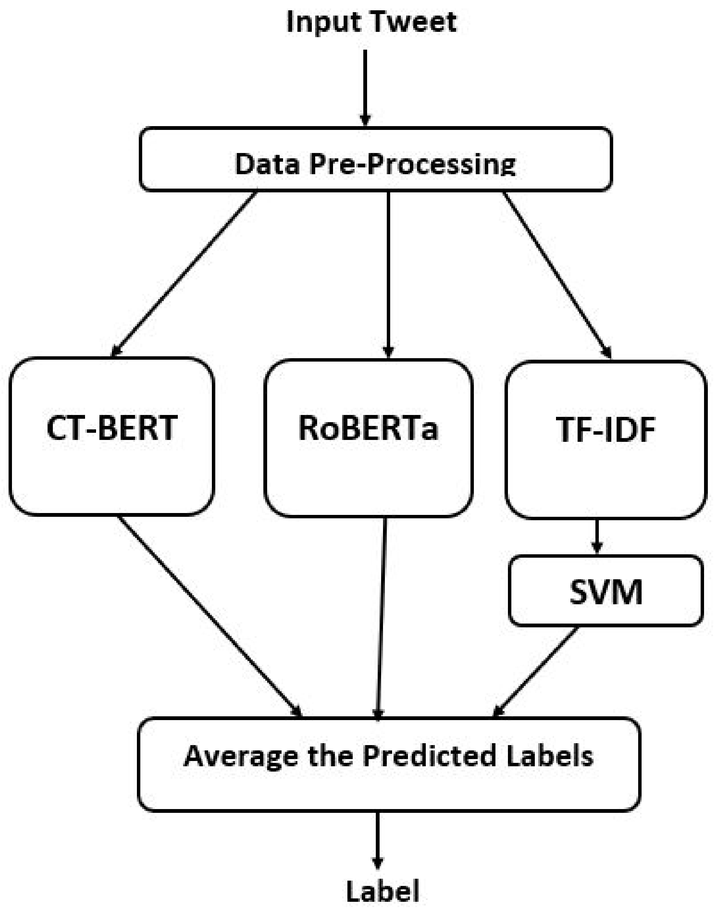
Model-1

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.



Model-2

This model is ensemble of CT-BERT, RoBERTa and TF-IDF with SVM. Each model is individually trained using the training set. The final prediction is obtained from the average of predictions of all these models.



Results

Model	F1 Score	Precision	Recall
CT-BERT	96.03	93.8	98.26
CT-BERT+ RoBERTa+ (TFIDF+SVM)	95.68	93.96	97.47

Table 1: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 2:F1-score, Precision, and Recall of proposed models on Validation data without using pre-processing steps

Conclusion

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%.