# NIT_COVID-19 AT WNUT-2020 TASK 2: DEEP LEARNING MODEL ROBERTA FOR IDENTIFY INFORMATIVE COVID-19 ENGLISH TWEETS

**Jagadeesh M S,P J A Alphonse**
**Department of Computer Applications,**
**National Institute of Technology,**
**Thiruchirapalli,TamilNadu,India.**

## ABSTRACT

This paper presents the model submitted by NIT COVID-19 team for identified informative COVID-19 English tweets at WNUT-2020 Task2. This shared task addresses the problem of automatically identifying whether an English tweet related to informative (novel coronavirus) or not. These informative tweets provide information about recovered, confirmed, suspected, and death cases as well as location or travel history of the cases. The proposed approach includes pre-processing techniques and pre-trained RoBERTa with suitable hyperparameters for English coronavirus tweet classification. The performance achieved by the proposed model for shared task WNUT 2020 Task2 is 89.14% in the F1-score metric.

## DEEP LEARNING MODEL

The goal of this WNUT-2020 task 2 is to identify a given COVID-19 tweet that is INFORMATIVE or UNINFORMATIVE. In this task, we have used a pre-trained deep learning model RoBERTa (Robustly Optimized BERT Approach), which is an optimized model for BERT. RoBERTa has features like

- Train the data up to 160 GB.
- Increase the number of iterations up to 500k.
- Train the model with batch size 8k.
- Larger byte-level BPE vocabulary with 50k sub word units.
- Dynamically changing the masking pattern applied to the training data.

We trained the RoBERTa model with different combinations of hyperparameters for the given dataset. Finally, we got better metrics for below hyper parametric values.

## ROBERTA MODEL HYPERPERAMETERS

- Used 'roberta-base'.
- Maximum learning rate is equal to 1e-5.
- We used batch size is equal to 16.
- Maximum sequence length of tweet in the dataset is 143.
- Avoid over-fitting we set hidden dropout is equal to 0.05.
- Hidden size for 'roberta-base' is equal to 768.
- An 'adam' is used for optimizer.
- Trained for 50 epochs.

## TABLE 0

| Dataset | Informative | UnInformative |
|---|---|---|
| Training | 3303 | 3697 |
| Validation | 472 | 528 |
| Test | 944 | 1056 |

COVID-19 English Tweets Data Set Details

## RESULTS ANALYSIS

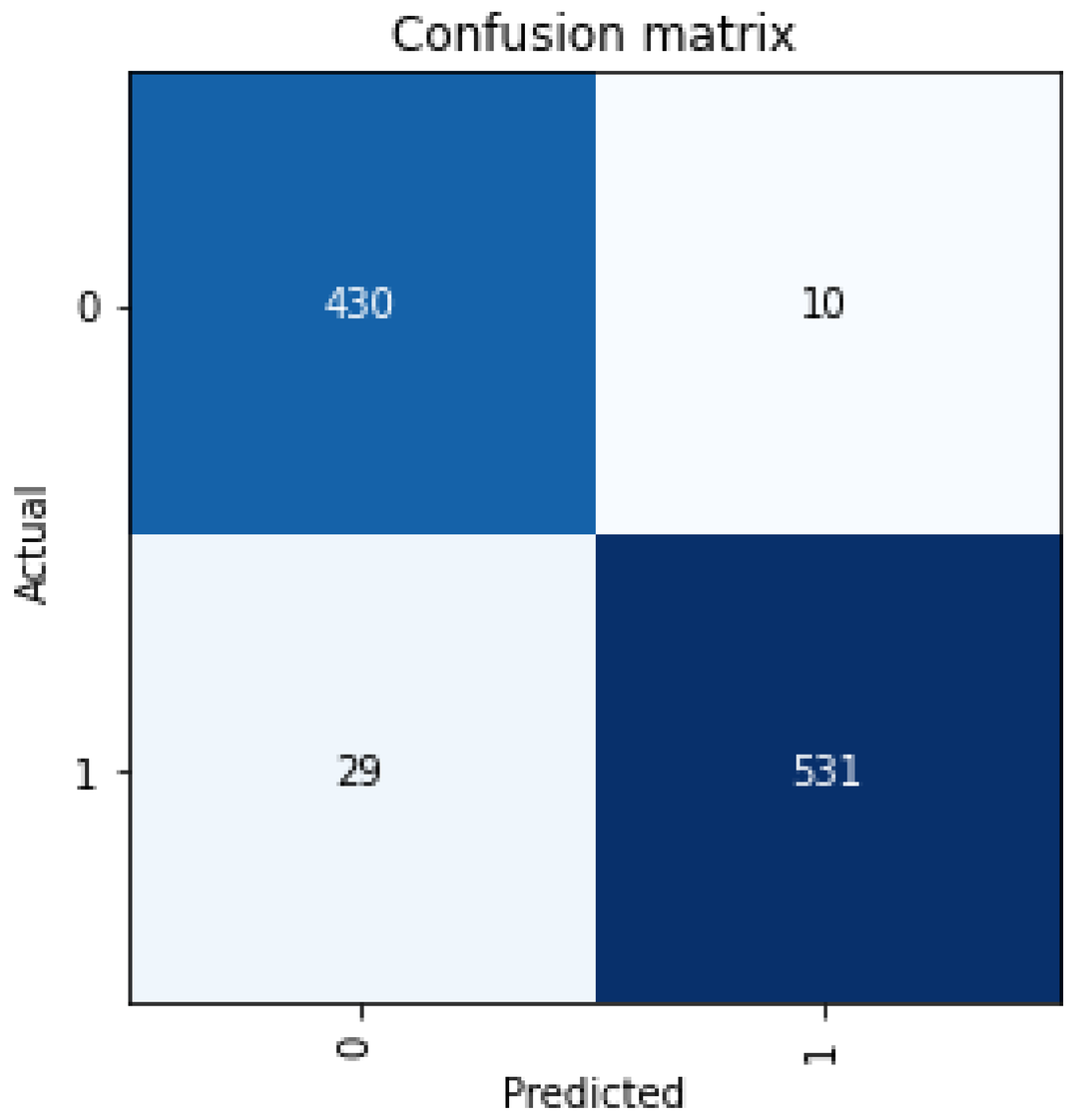
RoBERTa outperformed well,compare with BERT, by 0.20% F1-score. The baseline model results, however,worse than the RoBERTa and ,probably due the fact that the RoBERTa model requires fine-tuning for more task specific representations. The majority of RoBERTa errors are in INFORMATIVE class

## TABLE 0

| Model | F1-score | Accuracy |
|---|---|---|
| Random Forest | 81.5894 | 82.0688 |
| SVM | 82.7105 | 82.0488 |
| CNN | 83.7370 | 83.0691 |
| BERT | 88.9787 | 89.1006 |
| **RoBERTa** | **89.1864** | **89.5000** |

Results obtained on the Validation Set.

## FIGURE 1




0 - Represents INFORMATIVE
1 - Represents UNINFORMATIVE

Confusion Matrix for validation set

## CONCLUSION AND FUTURE WORK

In this (WNUT-2020 shared Task2) competition, We introduced the NIT_COVID-19 team's approach to the issue of informative tweet recognition and automatic categorization from the COVID-19 tweet master dataset of informative tweets. The pre-trained deep learning RoBERTa model outperformed the remaining models, including Random Forest, SVM,CNN and BERT. Furthermore, the analysis of the results indicates the some of INFORMATIVE tweets are not identified by the model. Such deficiencies demand larger training corpa and need a prominent features for training. In future work we will concentrate on latest pre-trained deep learning models and other issues for better results.