Detection of COVID-19 informative tweets using RoBERTa

Sirigireddy Dhana Laxmi Rohit Agarwal Aman Sinha

Indian Institute of Technology (Indian School of Mines) Dhanbad, India

Abstract

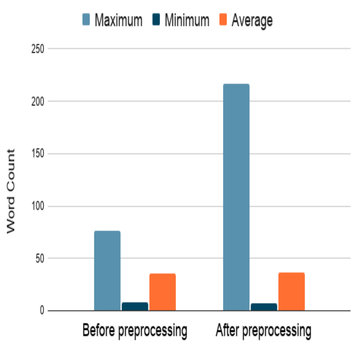
Social media such as Twitter is a hotspot of user-generated information. In this ongoing Covid-19 pandemic, there has been an abundance of data on social media which can be classified as informative and uninformative content. In this paper, we present our work to detect informative Covid-19 English tweets using RoBERTa model as a part of the W-NUT workshop 2020. We show the efficacy of our model on a public dataset with an F1-score of 0.89 on the validation dataset and 0.87 on the leaderboard.

WNUT Shared Task 2 Challenge

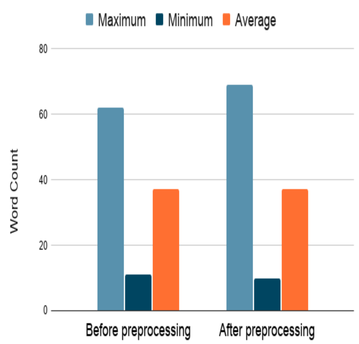
- A dataset consisting of english tweets subjecting Covid-19 is provided, which contains 7000 train data , 1000 valid data and 12000 test data points.We need to classify these tweets as informative or uninformative.
- In the context of this shared task, a tweet is considered informative if it is about recovered, suspected,confirmed, and death cases and location or travel history of the cases, and all the other tweets fall into the category of uninformative class.

Dataset Analysis

Data Cleaning is the primary step which includes lowercasing, conversion of emojis to text , removal of url and non-ascii characters. The word count of train and valid dataset before and after preprocessing are represented in histogram as below.



(a) Train Data word count statistics



(b) Valid Data word count statistics

Methodology

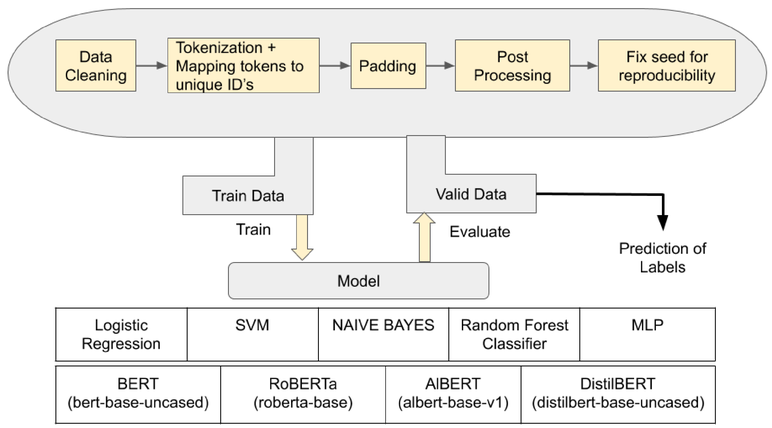


Figure: Overview pipeline of the experiment

Implementation

Processing part

Clean the dataset. Tokenize the tweets and map them to unique Ids

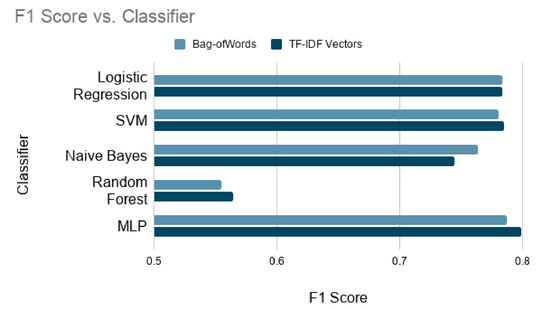
Conventional approach

Vectorize the tweets using Bag of words and Tf-Idf vectorizers. Apply Classifiers such as Logistic Regression, SVM, Naive Bayes, Random Forest and MLP.

Transformer based approach

1. Special tokens are applied at the start and end of each sentence.
2. Input tokens are padded and length of each tweet is truncated to 100 tokens and attention mask is applied
3. Training set is splited into train and Dev
4. Choose batch size and an Optimizer (AdamW considered)
5. Torch seed and numpy seed are set as zero to maintain reproducibility
6. Train dataset is trained using the pre-trained model
   - Choose number of epochs of training, train and perform a backward pass and evaluate the Dev set
7. Labels are predicted for Valid/ Test dataset

Results



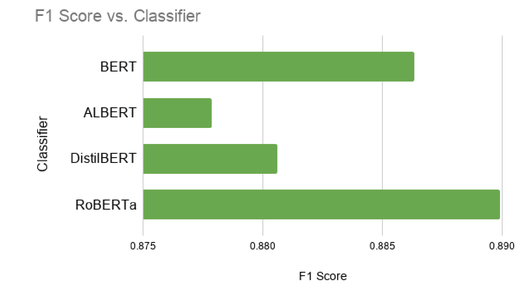


Figure: F1 Scores of Conventional approaches and Transformer approaches

Analysis and Future Work

- We have extensively compared the performance of various methods for this task. We applied conventional approaches and the latest state-of-the-art transformer-based methods.
- Since Twitter is primarily a microblogging media, short text classification using topic modeling and topic-enhanced embedding-based approach can also be useful for tweet classification.

References

[1] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.
Roberta: A robustly optimized bert pretraining approach.
arXiv preprint arXiv:1907.11692, 2019.

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