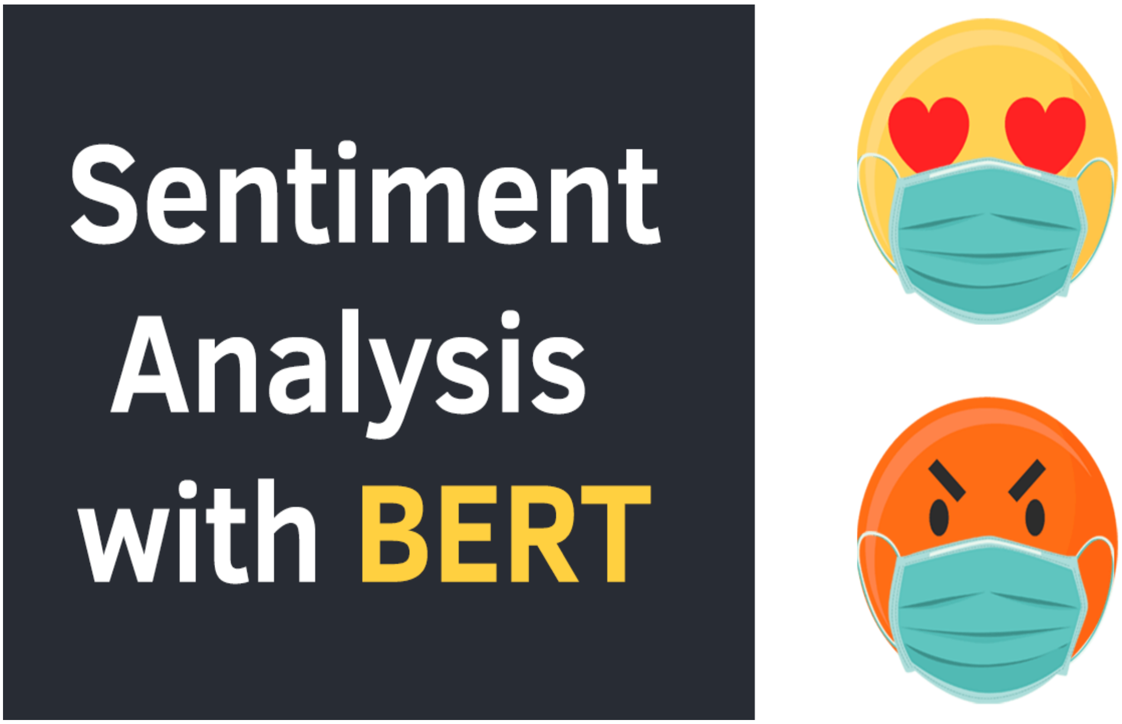


AUTOMATED HATE SPEECH ANALYSIS OF TWEETS USING BERT

ABSTRACT:

This project analyses nearly 25 thousand tweets and classifies them into three main categories:

- Hate Speech
- Offensive Language
- Neither

BERT stands for Bidirectional Encoder Representation from Transformers. The original English-Language BERT has two models:

The BERT$_{\text{base}}$:12 Encoders with 12 Bidirectional self-attention heads

The BERT $_{large}$:24 Encoders with 16 Bidirectional self-attention heads

We are using The BERT $_{base}$ model for our tweet analysis.

Introduction

The definition of the term "hate speech" as per Oxford is a 'speech that might involve abusive or threatening words which can have or can express pre-bias against a special community/ group. The pre-bias can be anything like region, race or sexual orientation, caste'. The anonymity of the internet has led to the propagation of abusive and hate content by haters from the safety of their homes. The proposed research work has attempted to classify the tweets into three main categories i.e., hate speech, offensive language or None.

Social Media platforms like Twitter, Facebook etc, use AI to identify and remove posts that offend society. Face book defines hate speech as direct attack against people – rather than concepts or institutions – on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. However, it is important to differentiate between hate-speech and offensive language as offensive language of one group may be the normal vocabulary of another group.

Literature Survey:

Automated Hate Speech Detection and the Problem of Offensive Language by T Davidson, is the paper on which we based this project. This project uses logistic regression, naïve Bayes, decision trees, random forests, and linear SVMs. And each model was tested using 5 point cross validation and found that logistic regression and linear SVMs tended to perform significantly better.

Methodology:

Data:

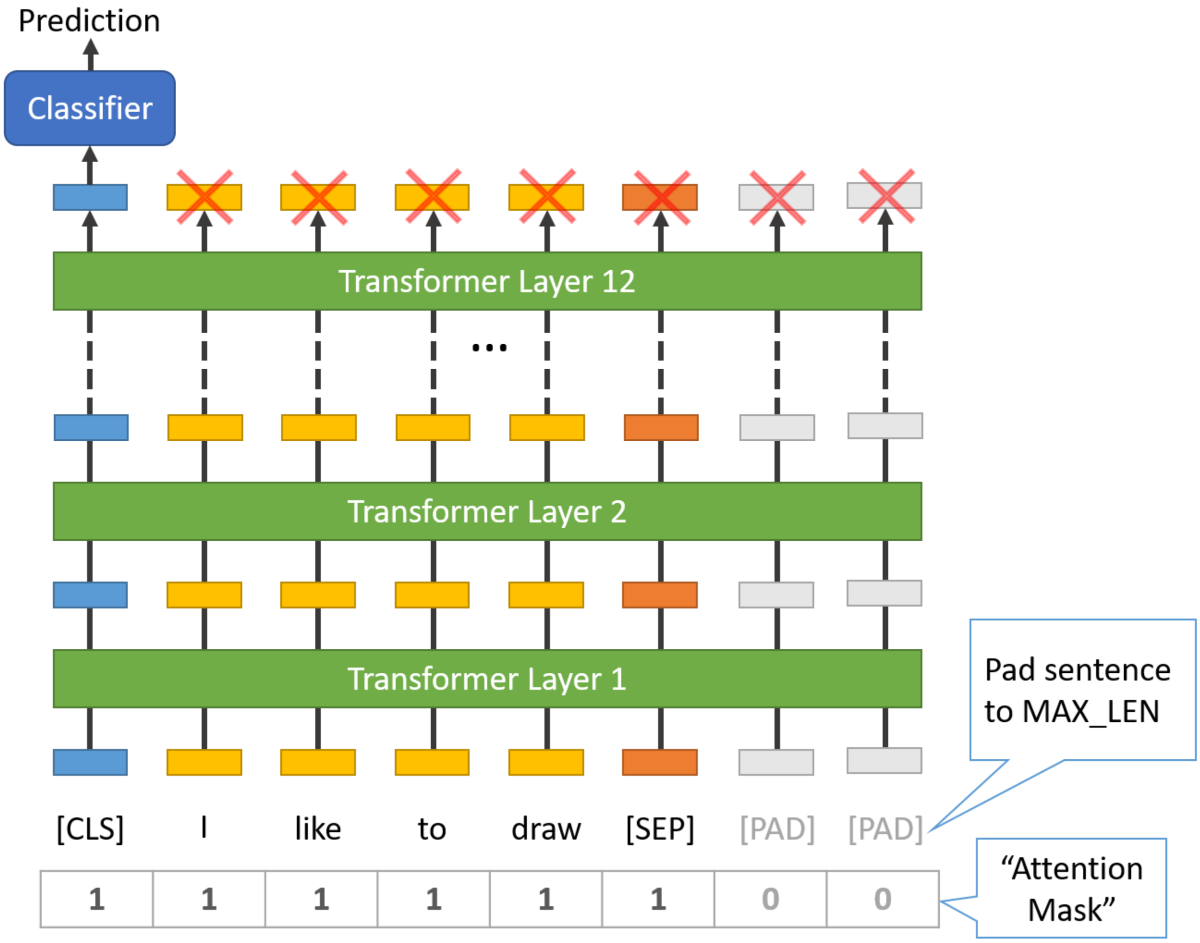
We used the original data the authors [1] had used in the original paper (T Davidson). It has 24783 tweets classified as either hate speech, offensive language or neither.

Data pre-processing:

We reduced the attributes from 7 to 2 by removing redundant attributes and we also checked for null values. We renamed the class attribute to score. We then tokenized the data using a BERT Tokenizer. We also need to add special tokens namely [CLS] token at beginning of each sentence to let it know were doing classification analysis, the padding token [PAD] to pad the sentence to maximum length which in our case is 32. The, other special token is the [UNK], the unknown token. Each token is given an integer id and an attention mask. An attention mask has either the value one or zero. 1 implies that the token matters and 0 implies that it doesn't. We split the data into 90% for training and 5% each for validation and testing. We then created torch datasets for training testing and validation data and then a data loader to load data into the model.

The Model:

We used a BERT base model that is pretrained on English Wikipedia and Book Corpus. The BERT base model has 12 encoders stacked on one another and each layer accepts input of the lower layer. For a given token, its input representation is constructed by summing the corresponding token, segment, and position embeddings [2]. Each layer applies self-attention and passes its result is through a feed-forward network before handing it over to the next encoder.

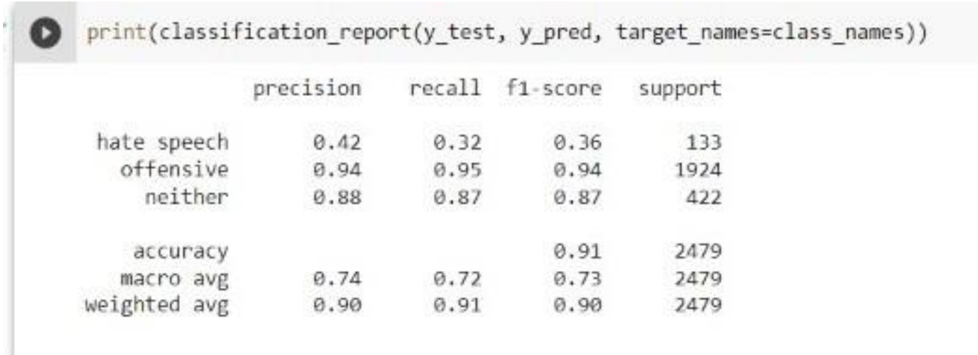


The [CLS] token contains the embedding of the whole sentence. The [CLS] representation is fed into an output layer for classification, such as entailment or sentiment analysis.

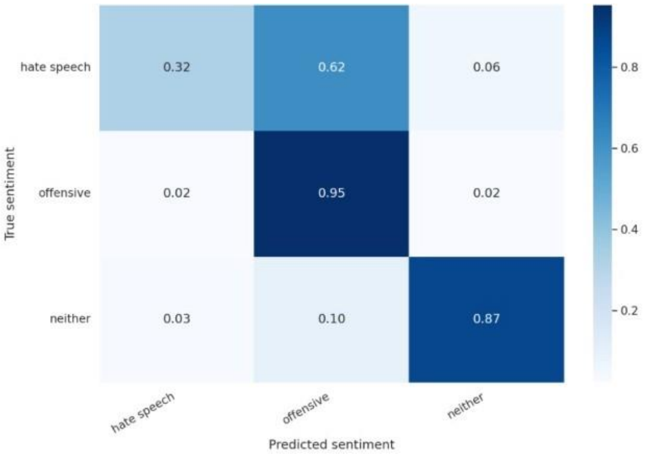
We then trained validated and tested the data.

Results:

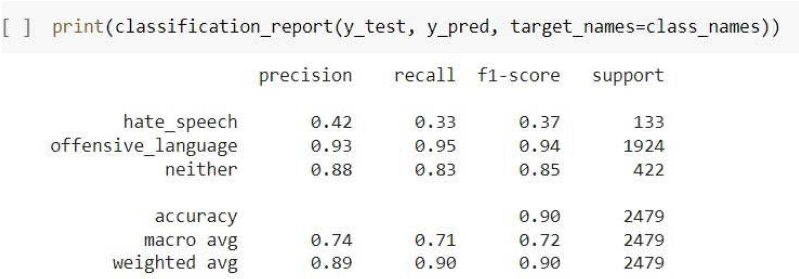
After training for 2 epochs:



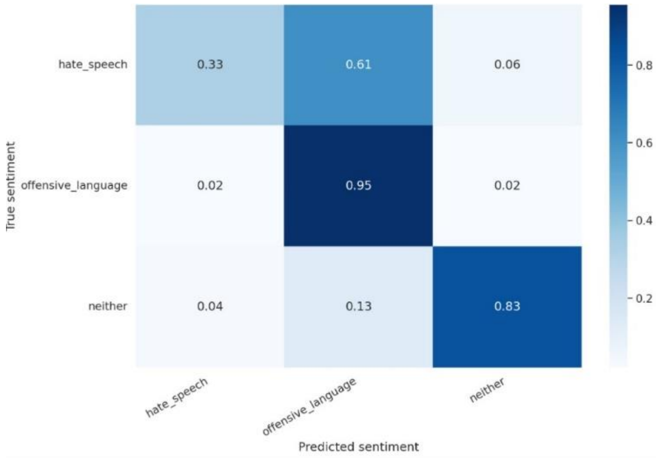
```
print(classification_report(y_test, y_pred, target_names=class_names))

               precision    recall  f1-score   support

  hate speech       0.42      0.32      0.36       133
    offensive       0.94      0.95      0.94      1924
      neither       0.88      0.87      0.87       422

     accuracy                           0.91      2479
    macro avg       0.74      0.72      0.73      2479
 weighted avg       0.90      0.91      0.90      2479
```



After Training for 4 epochs:



```
[ ]  print(classification_report(y_test, y_pred, target_names=class_names))

                    precision    recall  f1-score   support

       hate_speech       0.42      0.33      0.37       133
offensive_language       0.93      0.95      0.94      1924
           neither       0.88      0.83      0.85       422

          accuracy                           0.90      2479
         macro avg       0.74      0.71      0.72      2479
      weighted avg       0.89      0.90      0.90      2479
```



After training for 2 or 3 epochs (recommended by original BERT authors) we couldn’t get the required accuracy. 61% and 62% of hate-speech was still classified as offensive language by our model. Which is worse compared to the 32% by the logistic regression model with L2 regularization.

CONCLUSION:

Our BERT model did not perform as expected, we believe it is mainly due to the skewed nature of the data. In general BERT model outperform many models like SVM's, RNN, LSTM's, etc. in the field of text analysis due to its Bidirectionality which helps the model to grasp the context easily. Due to its faster output, it can be used to predict the type of content the post has before posting. This helps the company to censor the content beforehand. The only current limitation of BERT is that it cannot handle negation. For example:

A Labrador is a______

BERT predictions: dog, pet, companion

Negation:

A Labrador is not a______

BERT predictions: dog, pet, companion

BERT predicts the same output for both, though it is a simple straightforward thing.

References:

[1] Thomas Davidson, Debasmita Bhattacharya, Ingmar Weber Automated Hate Speech Detection and the Problem of Offensive Language, Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM '17, 2017, Montreal, Canada pp 512-515[online] ArXiv: 1905.12516

[2] Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
arXiv:1810.04805