Assignment 3

Kishen N Gowda
17110074

Sentiment Analysis of Hin-Eng mixed tweets

Dataset: link

Methodology:

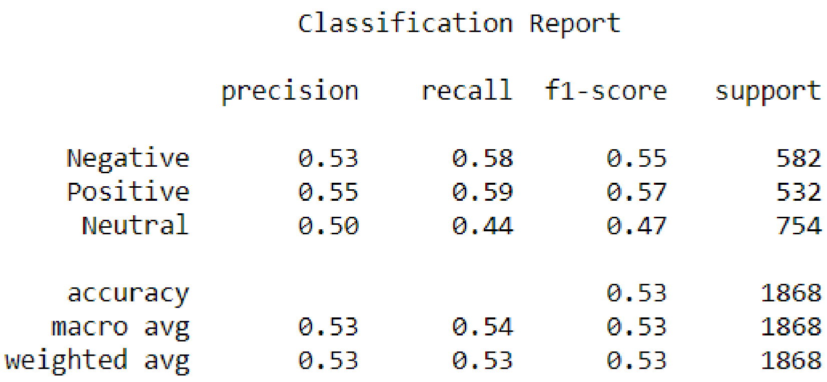
Github: link

For this task, I propose the following novel architecture based on SVM. As we know, SVMs are one of the best for classification-based tasks. But, as these are Hindi-English mixed tweets, some pre-processing has to be done. So, my method can be described in three phases:

1. **Phase I (Pre-processing):** First, I clean the unwanted text from the tweets (like the ones with labels "O", links, etc.). Next, on keen observation, I noticed that the labels of "Hin" and "Eng" were not proper in most of the tweets, especially the ones tagged "Hin". So, I followed the following procedure:
    - Check if the word exists in Wordnet, if yes just return, else keep it for further processing.
    - Check if the word is a bad word using this dictionary, if yes, translate, else keep for further processing
    - Finally, translate this word using Google's Cloud Translate API
2. **Phase II:** In the first phase, the aim is to classify the tweets if they are of neutral stance or non-neutral stance. For that, I used the Weighted MPQA Subjectivity-Polarity Classification. Based on the subjectivity score, if the cumulative score is either < 2 or > 2, they are classified as non-neutral, else neutral. Also, if there is an adjective in a tweet, it generally implies subjectivity. Hence, using Wordnet based potential adjective recognition I classify between non-neutral and neutral.
3. **Phase III:** In this phase, I classify between positive and negative stances. For that, I use Sentiwordnet to fetch positive and negative scores of the words, and then consider cumulative scores. With this as feature, I use Count Vectorizer (One Hot Encodings) on the tweets and concatenate it to form the feature vector. Then I use this as input to the SVM model.

[**Note:** I tried tf-idf vectors as well as Glove embeddings, but among them, one hot encoding had highest accuracy.]

Results:



```
                    Classification Report

              precision    recall  f1-score   support

    Negative       0.53      0.58      0.55       582
    Positive       0.55      0.59      0.57       532
     Neutral       0.50      0.44      0.47       754

    accuracy                           0.53      1868
   macro avg       0.53      0.54      0.53      1868
weighted avg       0.53      0.53      0.53      1868
```