DEEP LEARNING FOR NATURAL LANGUAGE PROCESSING

TWEET SENTIMENT ANALYSIS

ABHISHEK K

INTRODUCTION:

In recent years, people of all walks of life have chosen digital mediums to express their thoughts, concerns, and opinions on day-to-day issues occurring across the globe. Social media platforms have shrunk the world where anyone from one part of the world can communicate their thoughts and views about an incident that is happening on the other end of the globe. It is this ability to express views and freedom of expression that has also given rise to an aspect of rules and regulations to restrict users from sharing sensitive and delicate information and commenting in the absence of netiquette.

Tweet sentiment analysis is one such means and measure to identify inappropriate and derogatory comments that get posted on social media such as Reddit, Twitter, Facebook, and many others, and eradicate them at their root level. In order to do so, our primary objective should be to analyze and predict each and every tweets' underlying tone and sentiment. Further, the tweets can be classified as extremely positive, positive, neutral, negative, and extremely negative. The classification is done mainly based on keywords that project their positive or negative meaning.

This work is an attempt to use various machine learning techniques and follow data preprocessing approaches to train an ML text-based model with the given dataset containing the tweets and their nature. The trained model is then used to classify an unlabelled dataset containing only tweets.

DATASET DESCRIPTION :

The training dataset contains 6 columns namely:

- UserName - A primary key to uniquely identify every data point.
- ScreenName - id value of screen
- TweetAt - time of the tweet
- Location - the location of the tweet
- Tweet - the tweet
- Sentiment - the sentiment

The testing dataset contains 5 columns namely:

- UserName - A primary key to uniquely identify every data point.
- ScreenName - id value of screen
- TweetAt - time of the tweet
- Location - the location of the tweet
- Tweet - the tweet

Testing data contains one less column as compared to training data because it does not contain the target column, which has to be predicted by our model.

OVERVIEW OF CODE:

The following packages and libraries have been used during the course of this work.

- NLTK: The Natural Language Toolkit (NLTK) contains an array of libraries and programs for symbolic and statistical natural language processing for English, written in the Python programming language. Models belonging to this library were trained and used for classification.

- Sklearn: Sklearn is a machine learning library primarily used for supervised and unsupervised regression and classification tasks. Feature selection for this dataset was done using this library.

- Numpy: Numpy is a python library for performing large array-based and matrix-based calculations which also provides in-built functions for performing tasks associated with the data.

- Pandas: Pandas is a python library used for handling tabular-like data structures and it also provides tools to perform in-depth analysis on the data.

- Matplotlib: Matplotlib is a python library that facilitates functions to visualize data in various formats and intervals.

Feature selection :

- Feature selection is a process of considering only the required features based on correlation metrics.

- It may even increase the performance of models exponentially as the dataset becomes huge.

- Selecting features is an important task before performing any machine learning techniques to avoid the curse of dimensionality as much as possible.

Unlike numerical or categorical data, textual data should be processed in a different approach in order to make it suitable for training models and to test them.

The first step towards the classification of text is to tokenize it (break down the tweet into smaller chunks of words, symbols, links, etc based on spaces and punctuations).

The next step is to clean the data and remove unwanted and junk data such as hyperlinks, symbols, random characters and stop words so as to increase the accuracy and efficiency of the classification process.
Stop words are words in the English language that don't have any value if considered without any context (words like ‘and’, ‘is’, ‘in’).

After cleaning the data and getting rid of all unwanted characters, it is time to improve the quality of data by normalizing it. Normalization generally means the process of converting various canonical forms of an entity to its base form so as to reduce ambiguity among different words which actually belong to the same word. Lemmatization has been used to achieve normalization.

For example, the word ‘running’ will be converted into ‘run’.

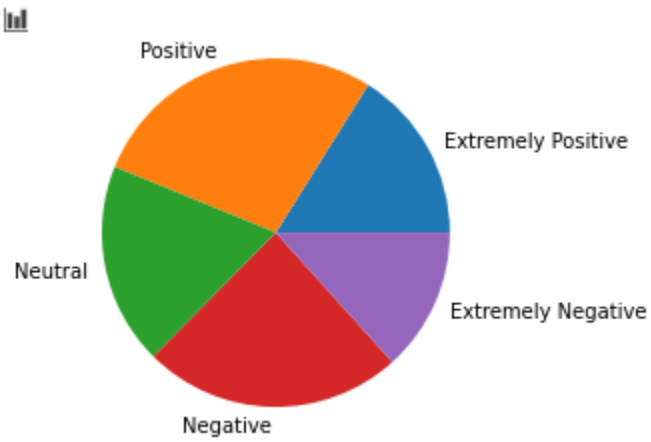
Before moving on to the actual training part, using the feature selection technique, it was found that of all of 5 feature vectors in the training dataset, the tweet contributes the most in classifying the target vector (‘sentiment’). Hence, all other feature columns other than the actual tweet are dropped from the main dataset.

Finally, naive bayes classifier model from NLTK library was trained using the processed dataset and then later the trained model was used to predict the sentiment of tweets present in the test dataset.

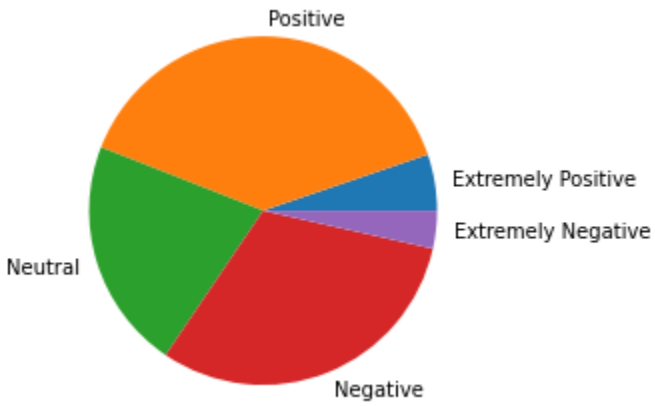
METHODS ATTEMPTED :

The classifier model being used here is naive bayes classifier from NLTK library which was trained using train() function with the processed clean train dataset.

VISUALIZATION :



Distribution of 5 unique sentiments within the given train dataset.

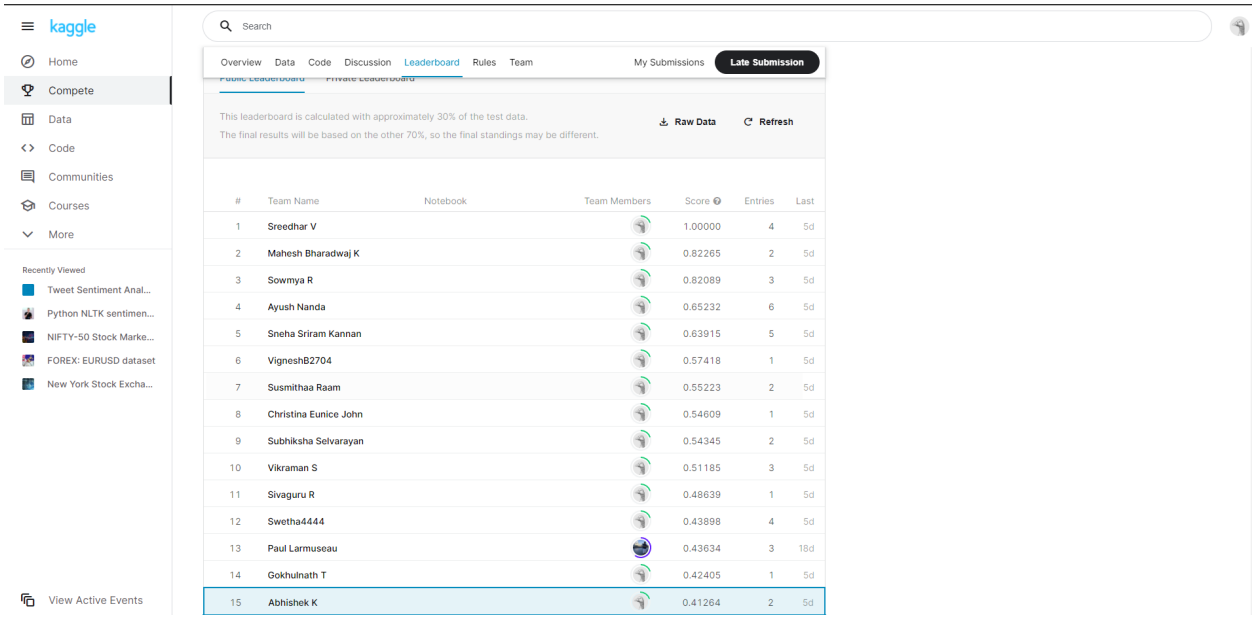


Distribution of 5 predicted sentiments within the given test dataset.

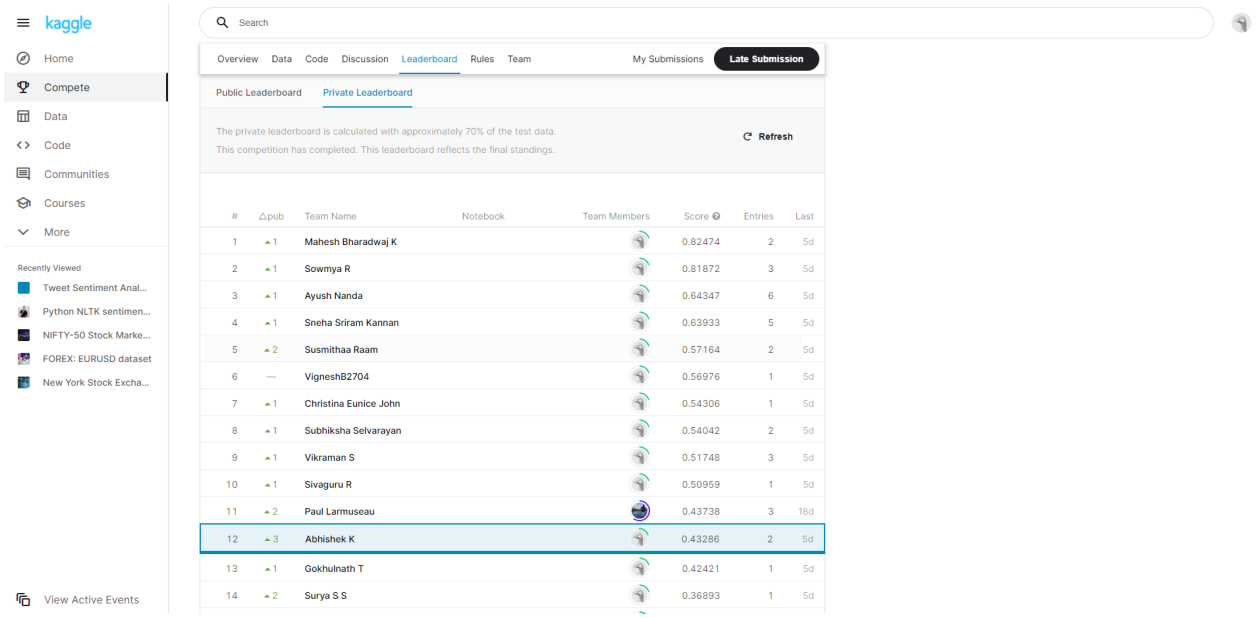
CONCLUSION :

This work was done with the aim of classifying the tweets as one among the given five sentiment labels. This work can be taken up by many others who would like to explore different paths and experiment with the dataset for better results. The scope of this work extends beyond just tweets where the applications of this project include filtering unwanted and derogatory comments and content on websites.

LEADERBOARD SCREENSHOTS



PUBLIC LEADERBOARD



PRIVATE LEADERBOARD