Mid Semester Presentation

TEAM NAME

Hardly Humans

Hardi Kadia

AU1841059

Tejas Chauhan

AU1841093

Khush Kalavadia

AU1841115

Jimil Desai

AU1841147

Mid Semester Presentation

Depression detection using Twitter data

Depression is a mood disorder that cannot be easily recognized. Due to the lockdown, there is a surge in the cases of depression.

Twitter, unlike other social media platforms, gives emphasis on the interaction through "tweets" which majorly includes text. Analyzing the text is simpler than analyzing audio, video, or photo.

Problem Statement

We are using the Twitter dataset to identify whether the user's tweet is reflecting positive or negative sentiment.

We are using a dataset of 1.6 million tweets which are assigned a label of positive or negative sentiment. We are applying the machine learning model to the words that present in the tweets after preprocessing the same. This way we would be able to classify the sentiment of our tweet.

Existing body of work

- A lot of research and time has gone into Sentiment Analysis and it is handled as NLP task at many levels of granularity. Sentiment analysis traditionally has been considered as a 3 layered approach. Starting from document-level classification, it has been handled at a sentence level and more recently at an abstract/phrase level.
- As far as the sentiment classification techniques are concerned there have been research going around mainly 3 topics:
  - Machine Learning
  - Lexicon Based Approach
  - Combination/Hybrid Approach

Our approach

Dataset identification

Using Twitter API to seek tweets

Using dictionary to find sentiment of the tweets

Using dataset which had 1.6 million tweets with sentiment labels

Preprocessing data

Removal of URL, tags, non-alphabetic characters, etc

Stopwords removal

Word tokenization

Lemmatization

TF-IDF Vectoriser

A process that identifies the importance of each word

Assigning weight to each word based on the frequency of the word in tweets.

These weights are used in the prediction

Our approach

Naive Bayes Classification

Calculate prior probability of class label

Likelihood Probability of each attribute for each class

Posterior probability using Bayes Formula

Higher probability class would be our prediction

Logistic Regression

Preparing optimisation function logistic regression

Achieving parameters from optimization function

Finding probability using weight of tweets and parameters

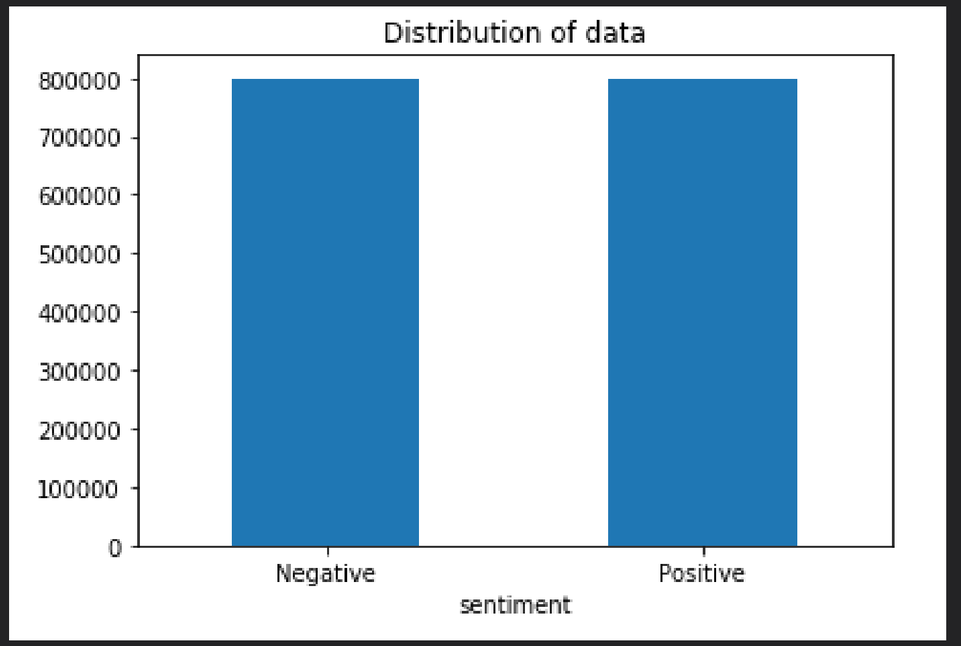
Classifying sentiment based on the probability

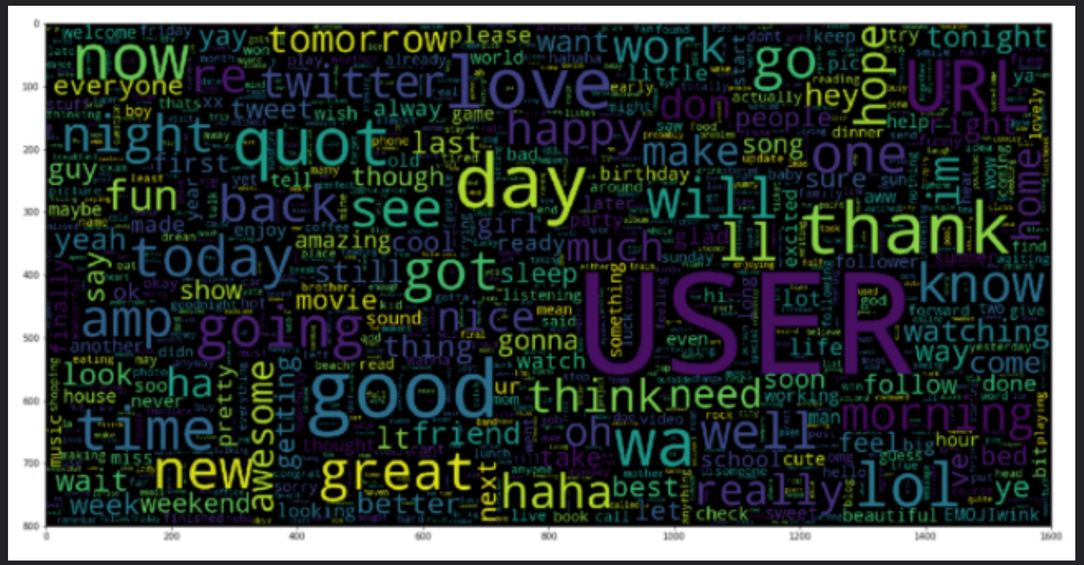
Depression detection using Twitter data

Result 1

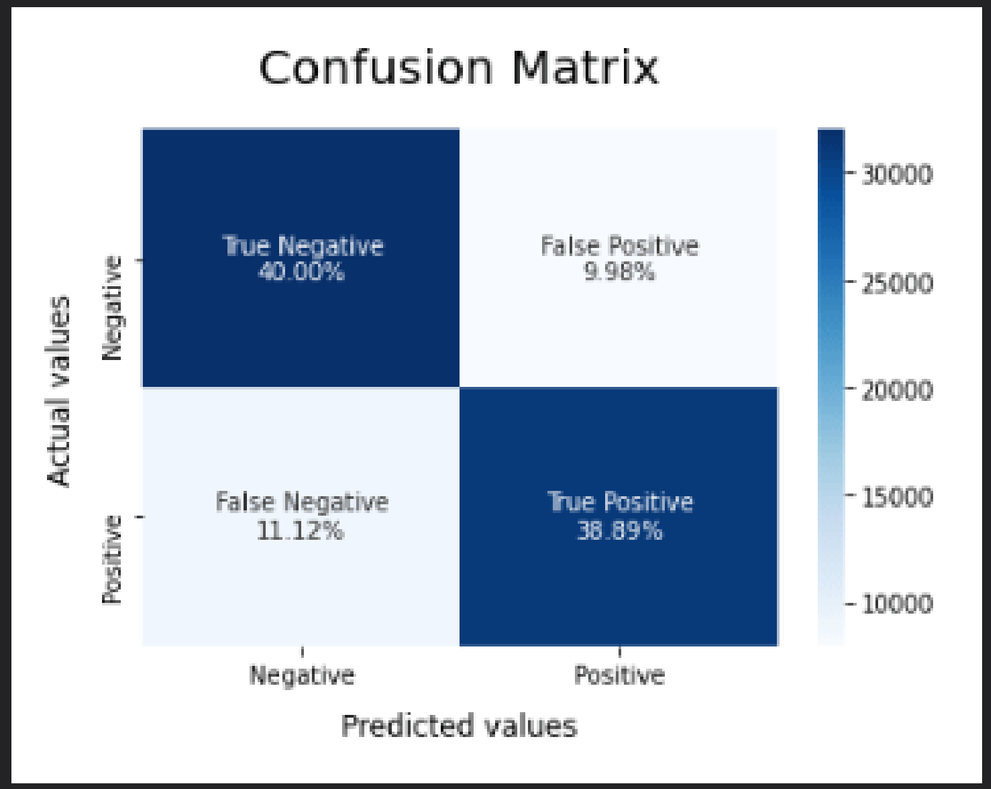
Distribution & Preprocessing of Data

There is a total of 1.6 million tweets which are equally divided into half positive tweets and half negative tweets. We have preprocessed the data i.e. remove unnecessary data from the tweets and plotted the **Word Clouds** for **Positive and Negative** tweets from our dataset. Here in the word cloud, word size shows the frequency of the word in data. Words that occur the most have the biggest font size while words that occur the least have the smallest font size.





Word Cloud of positive tweets



Result 2

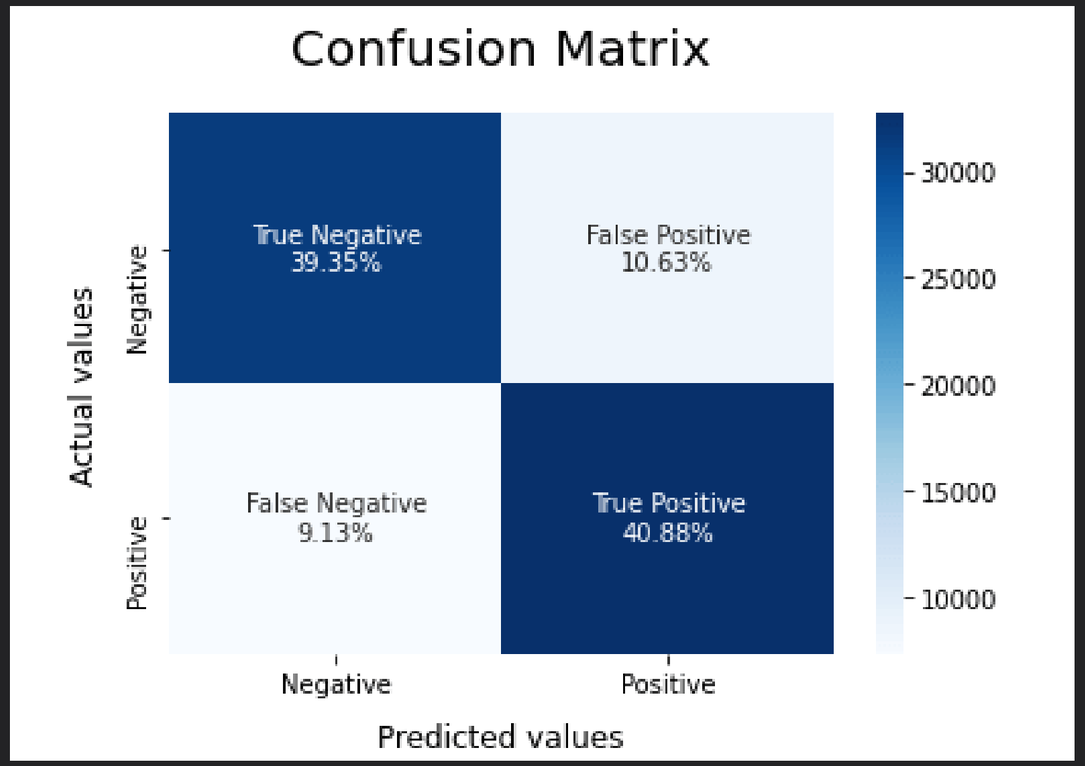
Multinomial Naive Bayes

Multinomial NB Classifier was applied to the training data and was tested on the training data. The multinomial NB model performs decently well giving us an overall accuracy of 79% while classifying the sentiment of a tweet. Training accuracy was 82.6% and testing accuracy was 78.9%. Both the accuracy are close to each other, which is a good sign indicating that the model is not overfitting. The confusion matrix of Multinomial NB is shown in the figure.

Result 3

Logistic Regression

The logistic regression model performs decently well giving us an overall accuracy of 80% while classifying the sentiment of a tweet. This model behaves similar to that of Naive Bayesian. The confusion matrix of Linear Regression is shown in the figure.



Future work

Linear SVC Model

It is mostly used in classification problems. Support Vectors are simply the coordinates of individual observation i.e. the weight of the tweets.

Decision Tree

This non-linear model is widely used in classification. The weights of the tweet can be used as a threshold at the nodes of the tree.

Random Forest

It comprises a set of decision trees and each of which is trained using random subsets of features.

KNN Algorithm

This supervised machine learning algorithm can be used for classification as well as regression based on the parameter K.

Role of each group member

Task	Hardi	Tejas	Khush	Jimil
Searching Dataset	✔			
Preprocessing data		✔		✔
TF-IDF Vectorizer			✔	✔
Naive Bayes		✔		✔
Logistic Regression	✔		✔	
Report Writing		✔		✔
Presentation	✔		✔	

References

- Depression detection from social network data using machine learning techniques - link
- Identifying Depression on Twitter - link
- A Literature Survey On Sentiment Analysis Techniques Involving Social Media And Online Platforms - link
- Twitter Sentimental Analysis Using Naive Bayes Classifier(Process Explanation) - link
- Speech and Language Processing. Daniel Jurafsky & James H. Martin - link

Thank you!