Project Report

Twitter Sentiment Analysis

By:

- Rohit Dhuratkar
- Abhijeet Patil
- Abhinav Rahate
- Akshay Mukundwar
- Soham Wani

INTRODUCTION

CONTEXT

- Numerous outlets are available for individuals to express opinions and emotions… positive, negative, and neutral.
- Need to promote positive news, react to the negative, and move the needle favorably on neutral news....as near real-time as possible
- Mining high volume, high-velocity data for meaningful insights is not easy!...too much, too fast
- Similar challenges exist across all industries/verticals

WHY ANALYTICS?

- What is trending positively/negatively over a period of time and why?
- Who is being talked about, where, and why?
- What college is being talked about?
- What topics are being discussed the most?
- Who is being talked about most positively?
- What are the best sources for positive exposure?
- What is the geographic location of the comments?

WHY VISUALIZATION?

- Data visualization is the presentation of data in a pictorial or graphical format.
- It enables decision-makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns.
- With interactive visualization, you can take the concept a step further by using technology to drill down into charts and graphs for more detail, interactively changing what data you see and how it’s processed. Tables, bar plots, timelines, word clouds, histograms, and pie chart can be used for visualization.

WHAT IS TWITTER SENTIMENT ANALYSIS?

- Twitter is an online news and social networking service that enables users to send and read short 140-character messages called "tweets". Registered users can read and post tweets, but those who are unregistered can only read them.

- Hence Twitter is a public platform with a mine of public opinion of people all over the world and of all age categories.

- As of October 2016, Twitter has more than 315 million monthly active users.

- Twitter Sentiment Analysis is the process of determining the emotional tone behind a series of words, used to gain an understanding of the attitudes, opinions, and emotions expressed within an online mention.

- Sentiment Analysis is the process of ‘computationally’ determining whether a piece of writing is positive, negative or neutral. It’s also known as opinion mining, deriving the opinion or attitude of a speaker.

Why sentiment analysis?

Business: In marketing field companies use it to develop their strategies, to understand customers’ feelings towards products or brand, how people respond to their campaigns or product launches and why consumers don’t buy some products.

Politics: In the political field, it is used to keep track of political view, to detect consistency and inconsistency between statements and actions at the government level. It can be used to predict election results as well!

Public Actions: Sentiment analysis also is used to monitor and analyze social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.

Authentication:

In order to fetch tweets through Twitter API, one needs to register an App through their twitter account. Follow these steps for the same:

- Open this link and click the button: 'Create New App'
- Fill the application details. You can leave the callback URL field empty.
- Once the app is created, you will be redirected to the app page.
- Open the 'Keys and Access Tokens' tab.
- Copy 'Consumer Key', 'Consumer Secret', 'Access token' and 'Access Token Secret'.

We follow these 3 major steps in our program:

- Authorize twitter API client.
- Make a GET request to Twitter API to fetch tweets for a particular query.
- Parse the tweets. Classify each tweet as positive, negative or neutral.

Problem Statement :

Here, we are trying to classify sentiment of the tweets i.e. whether the given tweet is Positive, Negative and Neutral. We are trying to check the sentiments of public towards Narendra Modi and Rahul Gandhi.

We have scrape data from Twitter with hashtag #narendramodi and #rahulgandhi. Our dataset contains 29302 tweets

OVERVIEW

Tweets are imported using tweepy library in python from Twitter API and the data is cleaned by removing emoticons and URLs. TextBlob and VaderSentiment are used for assigning the polarity to tweets. The opinion are expressed graphically through Pie Chart and Word Cloud. Then we build Naive Bayes, Random Forest and SVM models to predict the sentiment of tweets.

SYSTEM REQUIREMENTS

- Python
- Twitter Authentication to access API

FEATURES

1. Create a Twitter application
2. Extraction of Tweets:

tweepy - Provides an interface to the Twitter web API

OAuthHandler - Python Interface For authentication

Create twitter authenticated credential object. It is done using consumer key, consumer secret, access token, access secret.

Following models are applied:

I. Naive Bayes
II. Random Forest
III. Support Vector Machine

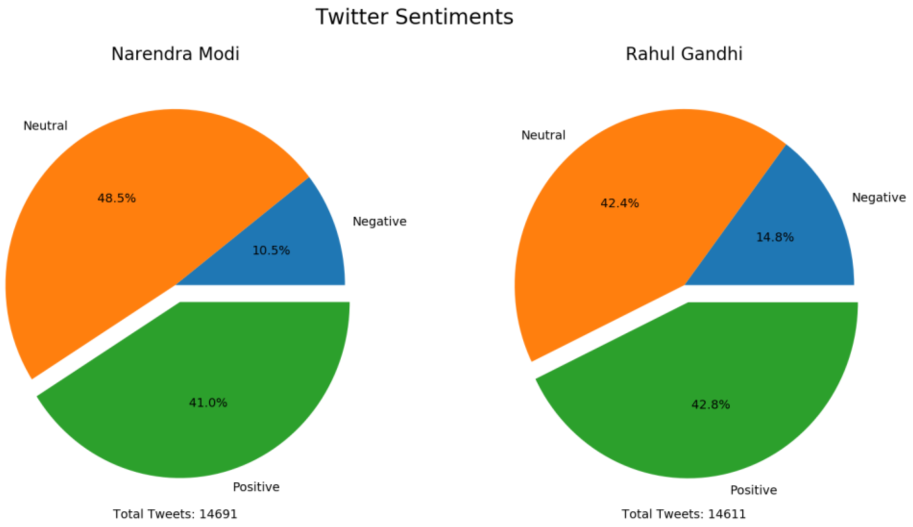
Walkthrough to Code:

- First of all, we have scrape tweets from twitter using twitter API, importing the tweepy library in python to call the Twitter API for fetching tweets.
- For cleaning the tweets we have created a function tidy which will take string as an input and return a string with all the handlers, links, special characters, and hashtags removed.
- We are then storing this clean tweets in a new column named tidy_tweet for future use.
- We have also removed the stop words using porter stemmer function which is also used to remove ly,s, ion suffix from the word so that one word will not represent in a different format.
- Here we have used two libraries to get the polarity of tweets i.e TextBlob and VaderSentiment.
- TextBlob is actually a high-level library built over top of NLTK library.
- Then we have created two polarity column one by TextBlob and the other by VaderSentiment.
- After getting the polarity we labeled it into Positive, Negative and Neutral i.e if the polarity of the tweet is below -0.05 then we have labeled it as Negative and if greater than 0.05 Positive else Neutral.
- For visualization, we have created a pie chart as well as a word cloud. A pie chart shows the percentage of Positive, Negative and Neutral tweets for Narendra Modi and Rahul Gandhi. And word cloud shows the trending hashtags.
- For building the model we have divided data into train and test keeping random_state = 42.

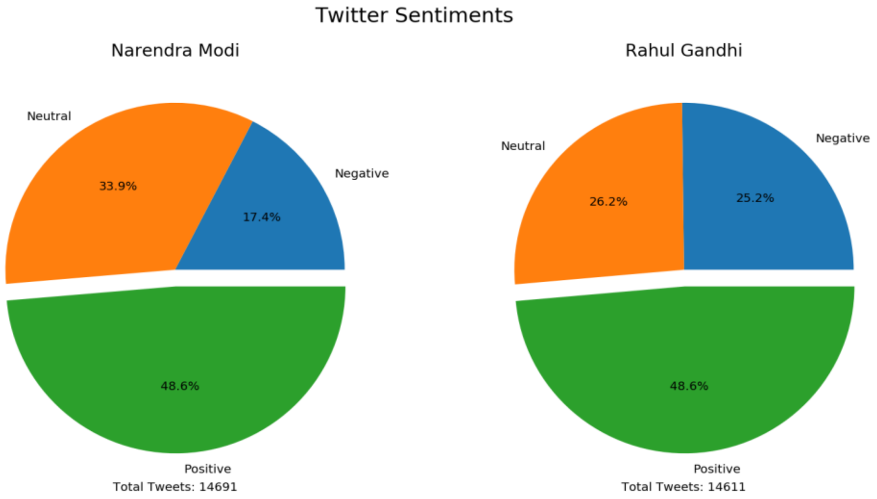
- For building the model we have vectorized the data using tfidf Vectorizer function as the data was in text format.
- Then by using the sklearn library we have build three models i.e Naive Bayes, Random Forest and SVM.
- On our dataset, Random Forest has outperformed than other two models.

Visual Output

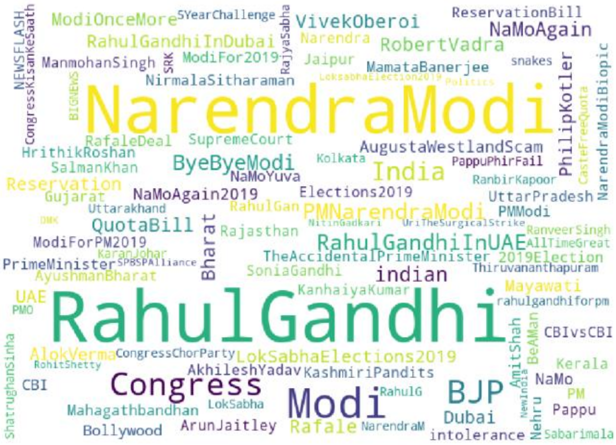
TextBlob

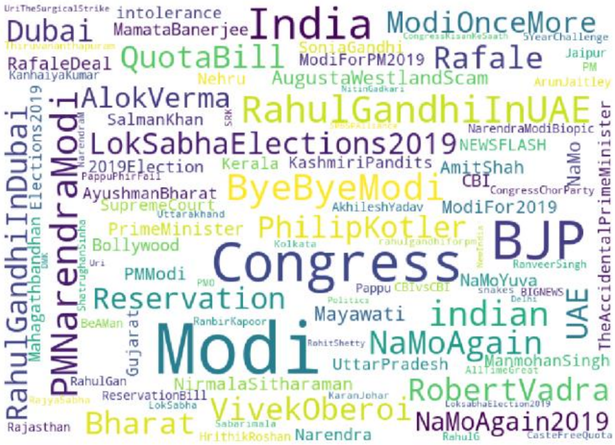


VaderSentiment



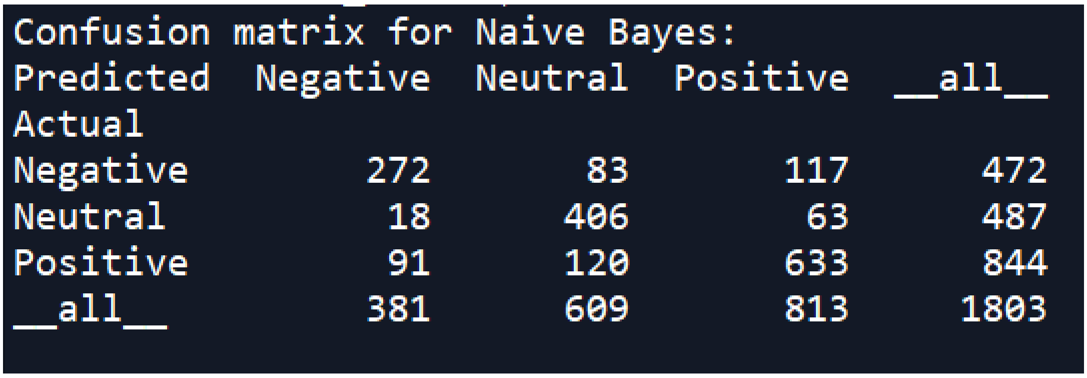
Word Cloud


NarendraModi
RahulGandhi
Modi
Congress
BJP
India
ModiOnceMore
5YearChallenge
VivekOberoi
ReservationBill
NaMoAgain
RahulGandhiInDubai
Narendra
RobertVadra
ManmohanSingh
ModiFor2019
Jaipur
MamataBanerjee
snakes
NirmalaSitharaman
NEWSFLASH
CongressKisanKeSaath
BIGNEWS
SRK
RajyaSabha
LoksabhaElection2019
Politics
PhilipKotler
NarendraModiBiopic
CasteFreeQuota
RafaleDeal
SupremeCourt
AugustaWestlandScam
HrithikRoshan
ByeByeModi
Kolkata
PappuPhirFail
SalmanKhan
NaMoYuva
RanbirKapoor
Reservation
Elections2019
UttarPradesh
Gujarat
NaMoAgain2019
PMNarendraModi
PMModi
Uttarakhand
RahulGan
DMK
QuotaBill
Bharat
NitinGadkari
UriTheSurgicalStrike
RanveerSingh
Rajasthan
ModiForPM2019
RahulGandhiInUAE
AllTimeGreat
KaranJohar
TheAccidentalPrimeMinister
2019Election
PrimeMinister
SPBSPAlliance
SoniaGandhi
Thiruvananthapuram
AyushmanBharat
indian
UAE
KanhaiyaKumar
Mayawati
PMO
rahulgandhiforpm
AmitShah
BeAMan
CBIvsCBI
AlokVerma
CongressChorParty
LokSabhaElections2019
RohitShetty
Kerala
AkhileshYadav
KashmiriPandits
NaMo
CBI
Shatrughansinha
RahulG
NewIndia
PM
Mahagathbandhan
LokSabha
Nehru
Pappu
Dubai
Bollywood
ArunJaitley
Rafale
NarendraM
intolerance
Sabarimala

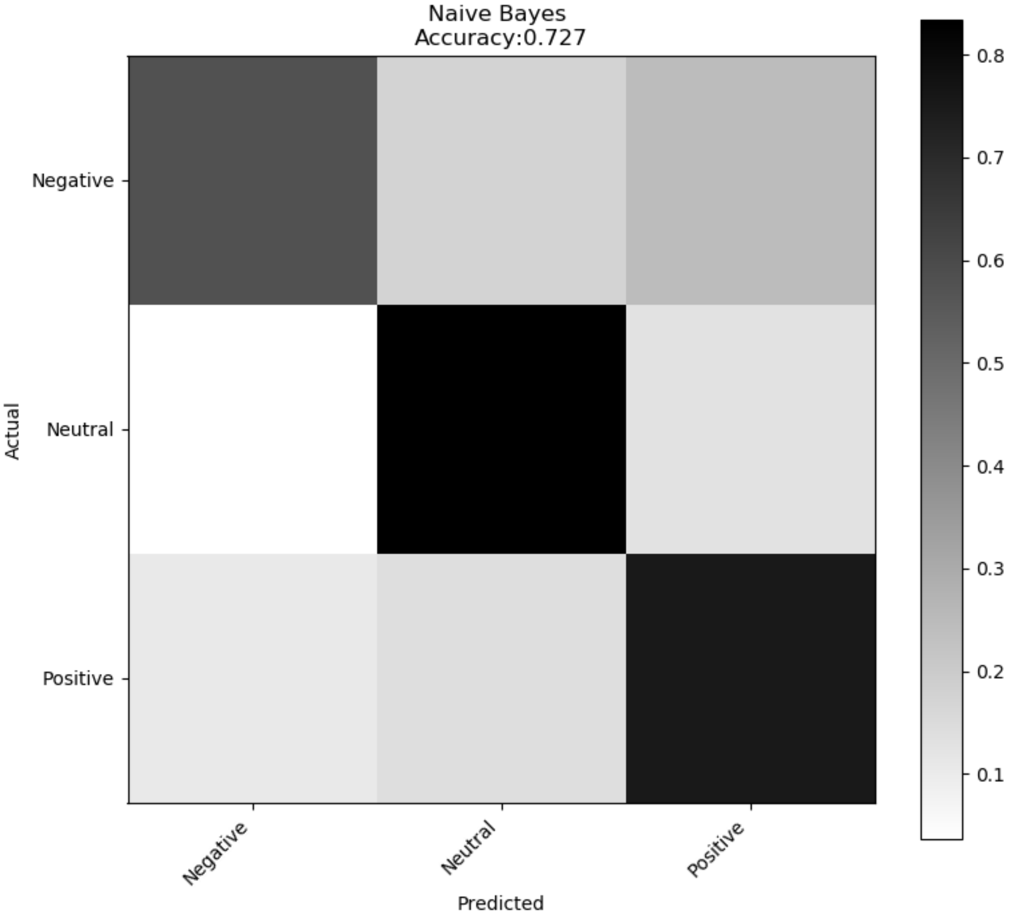

UriTheSurgicalStrike
intolerance
India
ModiOnceMore
Dubai
MamataBanerjee
CongressKisanKeSaath
5YearChallenge
Thiruvananthapuram
SoniaGandhi
RafaleDeal
QuotaBill
ModiForPM2019
Rafale
Jaipur
PM
KanhaiyaKumar
Nehru
AugustaWestlandScam
ArunJaitley
NitinGadkari
AlokVerma
RahulGandhiInUAE
SalmanKhan
SRK
NarendraModiBiopic
RahulGandhiInDubai
PMNarendraModi
Elections2019
LokSabhaElections2019
NEWSFLASH
NaMo
TheAccidentalPrimeMinister
2019Election
Kerala
KashmiriPandits
AmitShah
PappuPhirFail
CBI
CongressChorParty
AyushmanBharat
ByeByeModi
SupremeCourt
AkhileshYadav
ModiFor2019
Uttarakhand
PrimeMinister
PhilipKotler
BJP
Bollywood
Kolkata
rahulgandhiforpm
Mahagathbandhan
DMK
Uri
Congress
Shatrughansinha
PMModi
PMO
NaMoYuva
RanveerSingh
Pappu
CBIvsCBI
snakes
BIGNEWS
Delhi
Reservation
Politics
indian
UAE
Mayawati
BeAMan
RanbirKapoor
Gujarat
Modi
RohitShetty
NaMoAgain
UttarPradesh
ManmohanSingh
AllTimeGreat
RahulGan
NirmalaSitharaman
RobertVadra
RajyaSabha
ReservationBill
KaranJohar
LokSabha
LoksabhaElection2019
VivekOberoi
Bharat
Sabarimala
NaMoAgain2019
Rajasthan
HrithikRoshan
Narendra
RahulG
CasteFreeQuota

Naive Bayes:

- The simplest solutions are usually the most powerful ones, and Naive Bayes is a good proof of that. In spite of the great advances of the Machine Learning in the last years, it has proven to not only be simple but also fast, accurate and reliable. It has been successfully used for many purposes, but it works particularly well with natural language processing (NLP) problems.
- Naive Bayes is based on the Bayes theorem of probability, It is a supervised machine learning algorithm.
- It assumes independence among all predictors
- It is Extremely fast compared to other classification algorithms. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.
- Following is the confusion matrix for naïve bayes



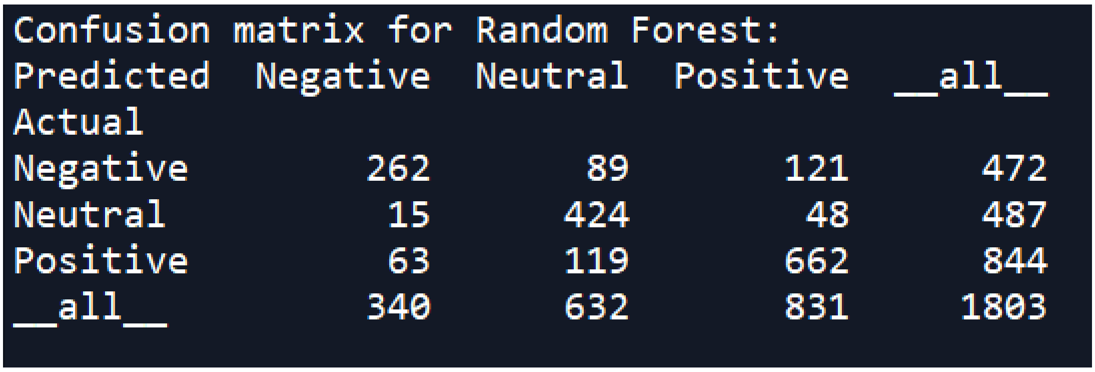
```
Confusion matrix for Naive Bayes:
Predicted  Negative  Neutral  Positive  __all__
Actual
Negative        272       83       117      472
Neutral          18      406        63      487
Positive         91      120       633      844
__all__         381      609       813     1803
```



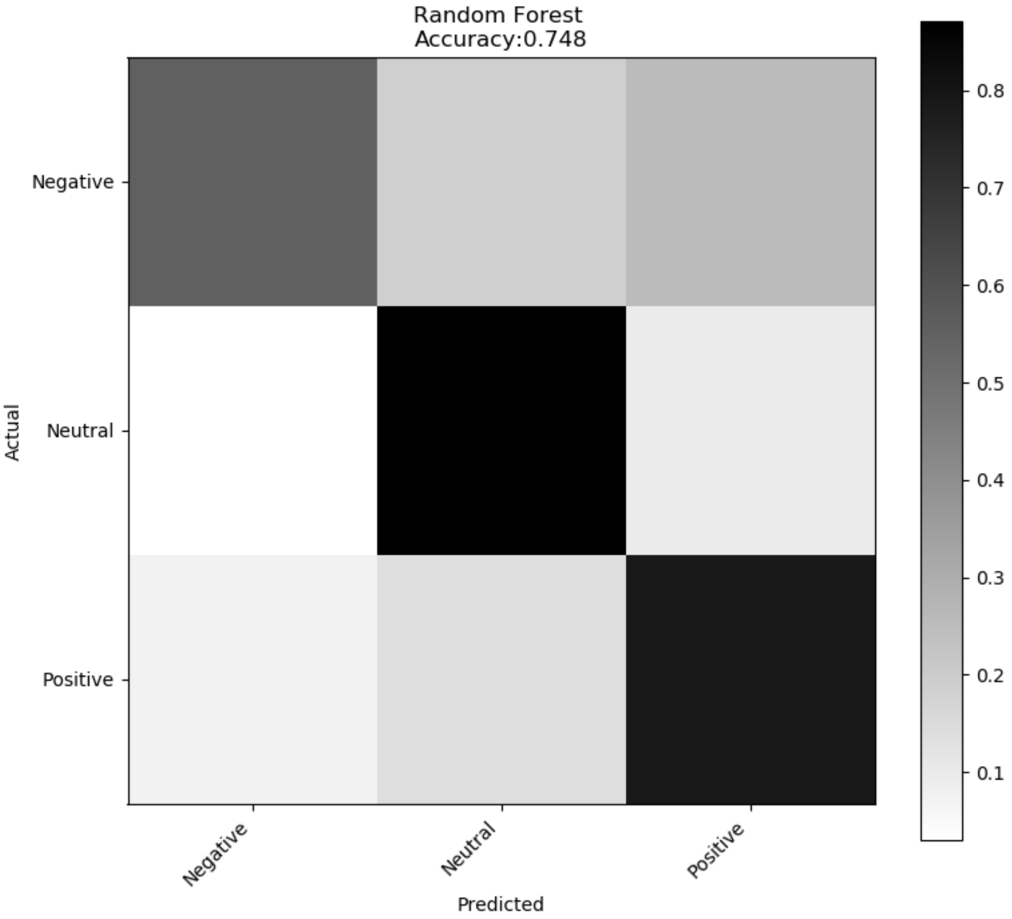
RandomForest

Random Forest is an ensemble learning algorithm for classification and regression. Random Forest generates a multitude of decision trees classifies based on the aggregated decision of those trees. For a set of tweets x 1 , x 2 , . . . x n and their respective sentiment label y 1, y 2,. . . n bagging repeatedly selects a random sample (X b, Y b) with replacement. Each classification tree f b is trained using a different random sample (X b, Y b) where b ranges from 1 . . . B. Finally, a majority vote is taken of predictions of these B trees.

Following is the confusion matrix for random forest.



```
Confusion matrix for Random Forest:
Predicted  Negative  Neutral  Positive  __all__
Actual
Negative        262       89       121      472
Neutral          15      424        48      487
Positive         63      119       662      844
__all__         340      632       831     1803
```



Support Vector Machine

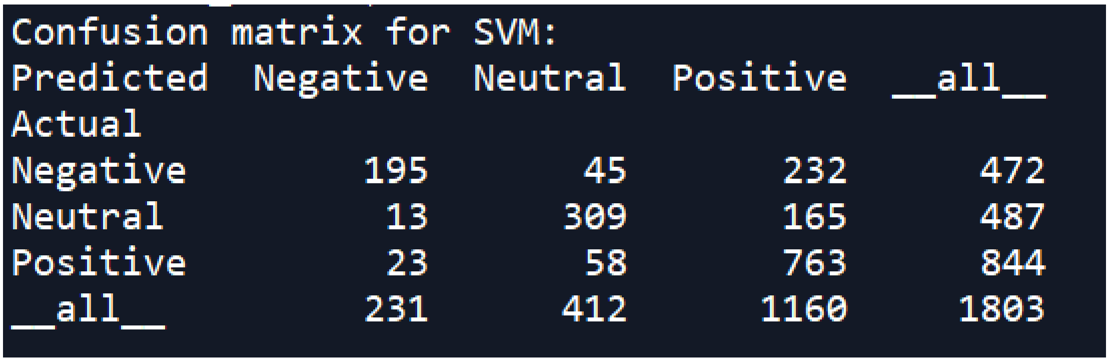
SVM, also known as support vector machines, is a non-probabilistic binary linear classifier. For a training set of points (xi, y i) where x is the feature vector and y is the class, we want to find the maximum-margin hyperplane that divides the points with y i = 1 and y i = −1.

The equation of the hyperplane is as follow

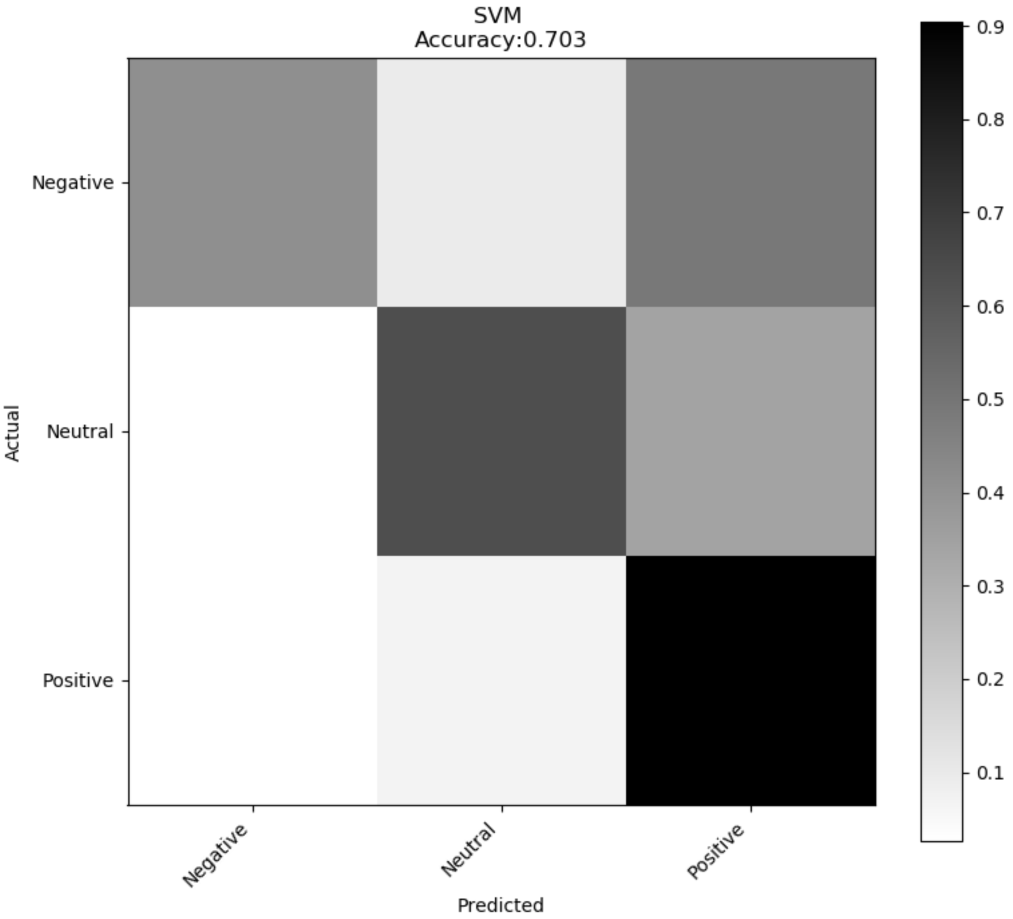
$$w \, . \, x - b = 0$$

We want to maximize the margin.

Following is the confusion matrix for SVM :-



```
Confusion matrix for SVM:
Predicted  Negative  Neutral  Positive  __all__
Actual
Negative        195       45       232      472
Neutral          13      309       165      487
Positive         23       58       763      844
__all__         231      412      1160     1803
```



Result from models

Results of different classification Algorithms:
For polarity assigned using TextBlob

	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.696	0.661	0.623	0.641
SVM	0.677	0.732	0.541	0.622
RandomForest	0.755	0.726	0.691	0.708

Results of different classification Algorithms:
For polarity assigned using VaderSentiment

	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.727	0.720	0.720	0.720
SVM	0.703	0.751	0.651	0.697
RandomForest	0.748	0.746	0.737	0.741

LIMITATIONS

- The Twitter Search API can get tweets upto a maximum of 7 days old.
- Not effective in detecting sarcasm.
- Cannot get 100% efficiency in analyzing the sentiment of tweets.

FUTURE WORK

- Detect sarcasm in tweets
- Analyse images for emotions
- Add Hindi words to the dataset.
- Find no. of mentions of n particular organizations (And analyze sentiment)

Conclusion

The random forest classifier algorithm gave us an accuracy of about 75%. We presume that some of the reasons for not getting a higher accuracy while predicting the sentiments of the review could be because of the following type of misclassifications: when the person writing the tweet talks more about the plot/characters than about his opinion about the topic. This can potentially be misclassified. Another such case is when a review contains a lot of quotes. In such a case when the quotations are stripped off in the data preprocessing stage, the quotes are considered as part of the reviewer's opinion. This could result in a lot of false positives and false negatives leading to lower prediction accuracy.

References

- Alexander Pak, Patrick Paroubek. 2010, Twitter as a Corpus for Sentiment Analysis and Opinion Mining.
- Alec Go, Richa Bhayani, Lei Huang. Twitter Sentiment Classification using Distant Supervision.
- Jin Bai, Jian-Yun Nie. Using Language Models for Text Classification.
- Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau. Sentiment Analysis of Twitter Data.
- Fuchun Peng. 2003, Augmenting Naive Bayes Classifiers with Statistical Language Models
- .Twitter Sentiment, an online application performing sentiment classification of Twitter. <http://twittersentiment.appspot.com/
- Ben Parr. Twitter Has 100 Million Monthly Active Users; 50% Log In Everyday.