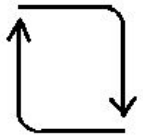
Twitter Sentiment Analysis
@tweetanalysis

Exploration in Data Science - Project Presentation by

@LarrisaCarvalho
@SiriKoduru
@RuchaKulkarni

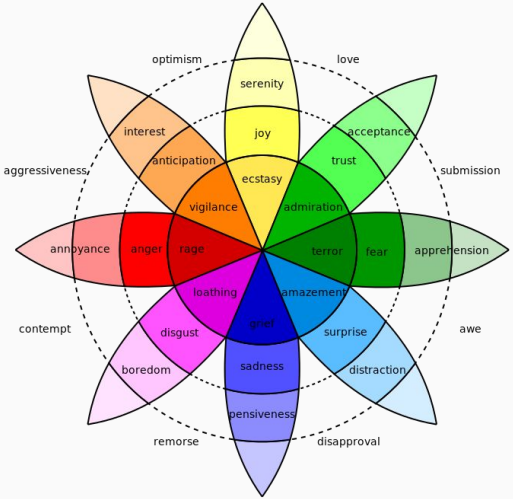
Objective

- Obtain high accuracy on classifying sentiment in Twitter messages using machine learning techniques.

- Compare all classifying algorithms based on the accuracy performed on real time tweets.

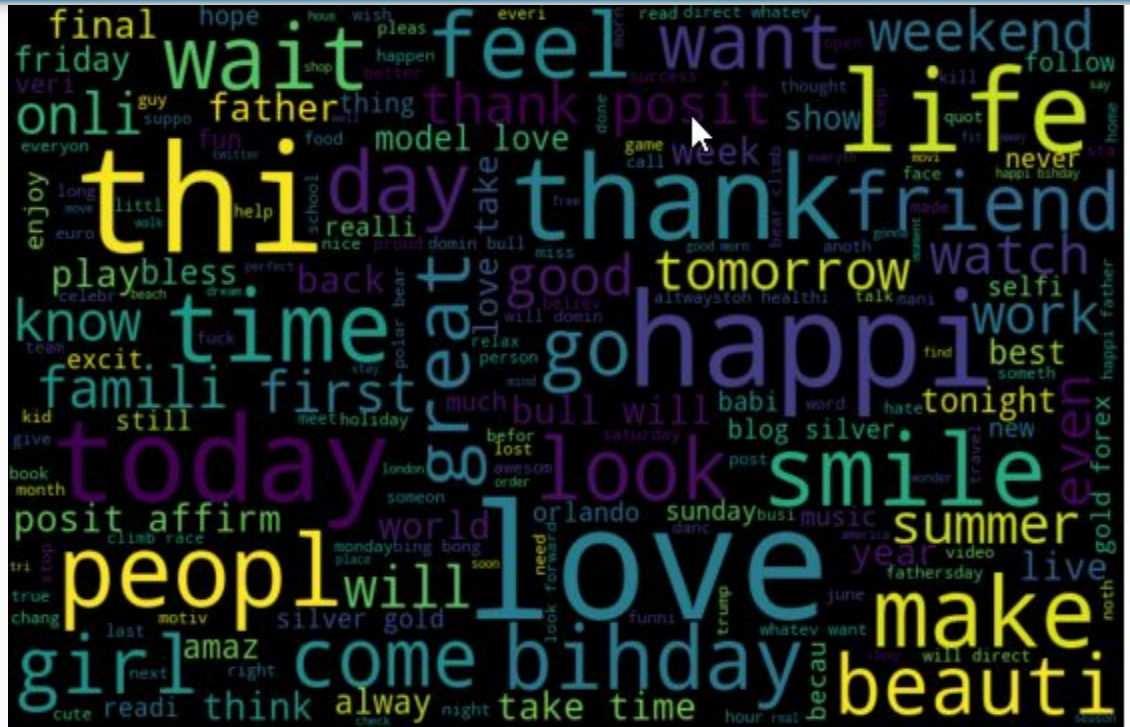
Why is Sentiment Analysis Important?

- Customers express their thoughts and feelings more openly
- Opinions in survey responses and social media conversations,
- It allows businesses to understand the sentiment of their customers
- Allows brands to learn what makes customers happy or frustrated
- Businesses can make better and more informed decisions.



Twitter Sentiment Analysis Use Cases

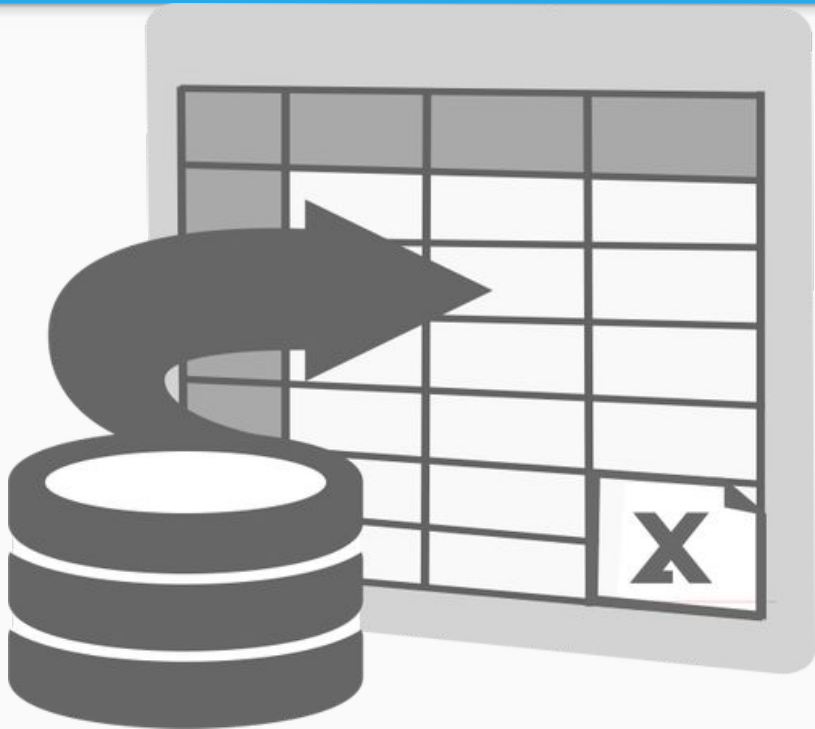
- Customer Service
- Market Research
- Brand Monitoring
- Social Media Monitoring
- Political Campaigns



Data Gathering

- Kaggle

- Consists of some tweets along with their sentiment - Tweepy



Data Cleaning

- Remove Stop Words

```
['Me', 'dragging', 'gym']
```

stop-words - Notepad

File Edit Format View Help

```
saw
say
says
second
seconds
see
seem
seemed
seeming
seems
sees
several
shall
she
should
show
showed
showing
shows
```

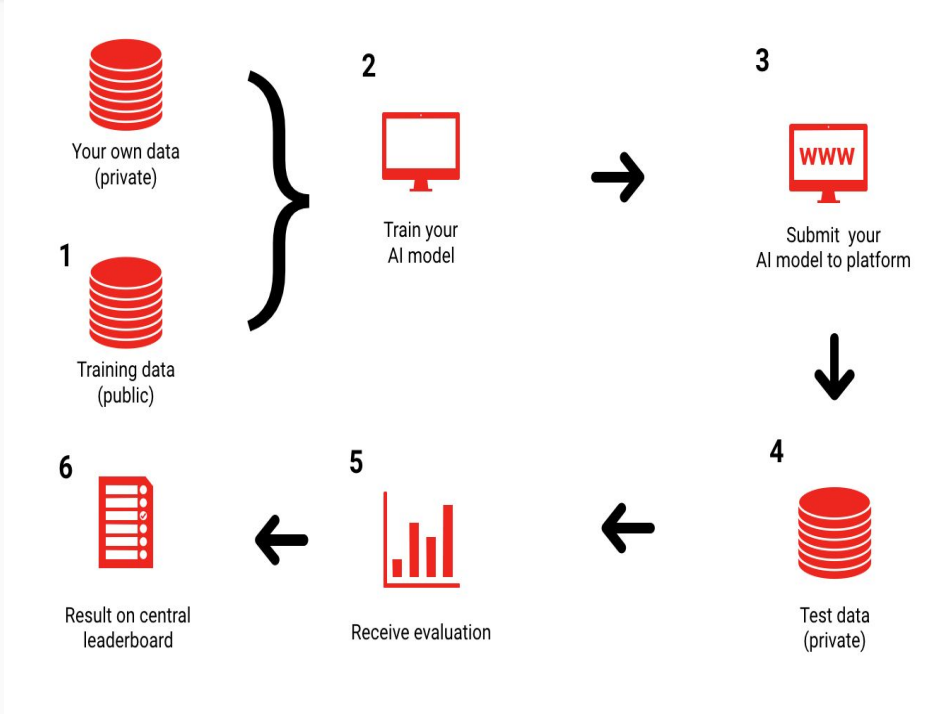
Train the Data

- We used tweets from data set to train our model.

- In order to train the model we need to represent each of the tweet in the form of a feature vector using Word2Vec.

Further Steps

- 80-20 train and test
- Represent the tweet as a feature vector
- Hypertune
- Extract eigen vectors
- Calculate the PCA features



How we are going to classify the data?

We are going to classify the Tweets from Twitter as positive, negative or neutral tweets.

Few Positive Tweets Examples:

1. @Msdebramaye I heard about that contest! Congrats girl!!
2. UNC!!! NCAA Champs!! Franklin St.: I WAS THERE!! WILD AND CRAZY!!!!!! Nothing like it...EVER http://tinyurl.com/49955t3

How we are going to classify the data?

Few Negative Tweets Examples:

1. no more taking Irish car bombs with strange Australian women who can drink like rockstars...my head hurts.
2. Just had some blood work done. My arm hurts

Why Word2Vec model?

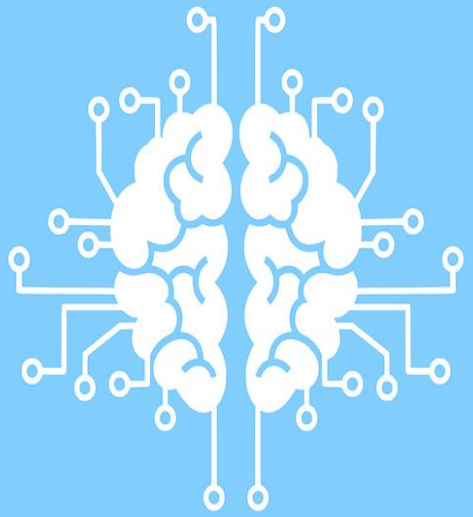
1. Deep learning model developed by Google.
2. Capturing the context of words..
3. Takes an input of a large corpus of documents like tweets or news articles and generates a vector space of typically several hundred dimensions.
4. Important concept is that word vectors close to each other in the vector space are of the same meaning and the same context.
5. delivers enhanced feature engineering for raw text data.

Training a classifier for classifying tweets

We used 3 machine learning algorithms in order to classify the feature vectors i.e. the tweets and then classify them accordingly.

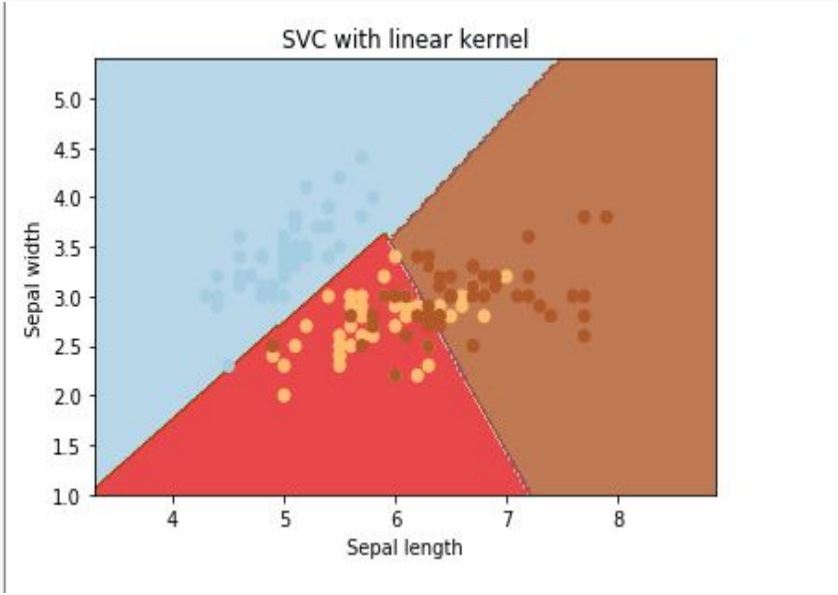
Algorithms:

1. SVM
2. Random Forest
3. Multi Layer Perceptron



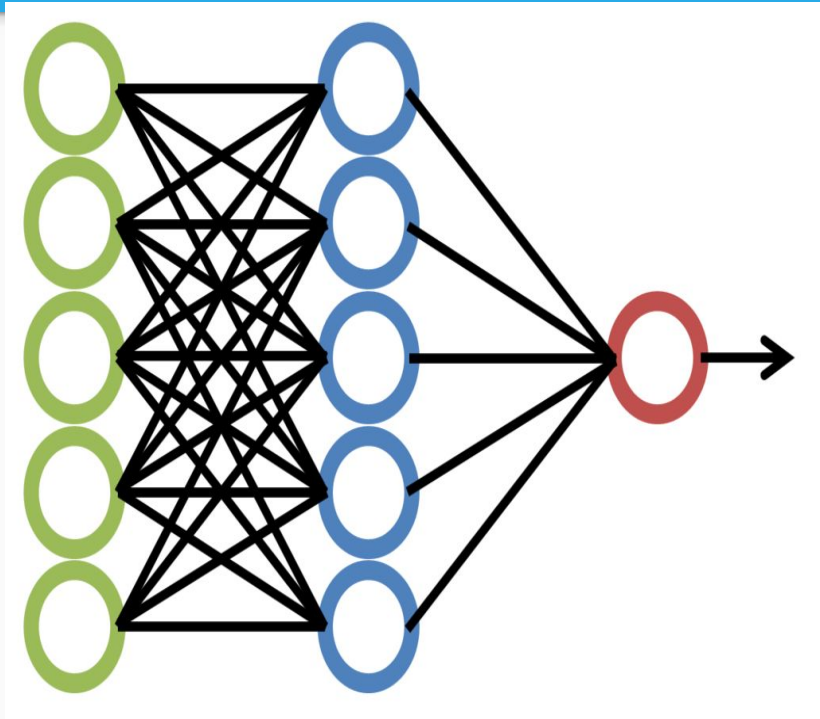
SVM

- SVM - a supervised learning algorithm.
- most robust prediction methods.
- Maps training examples to points in space to maximize width between categories.
- New examples are then mapped into the same space and classified accordingly.
- Non linear classification also can be done.



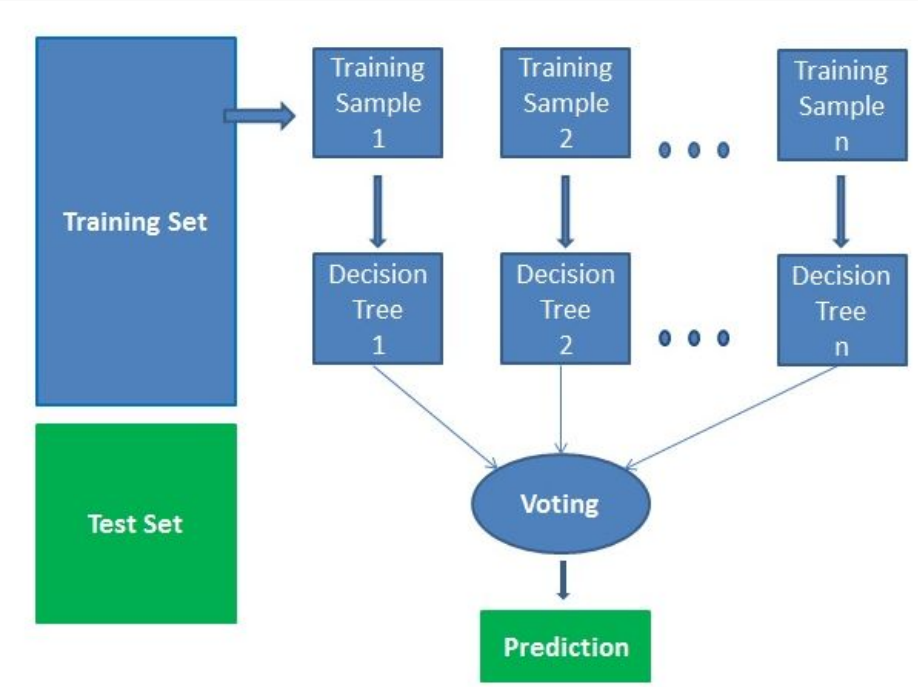
MLP

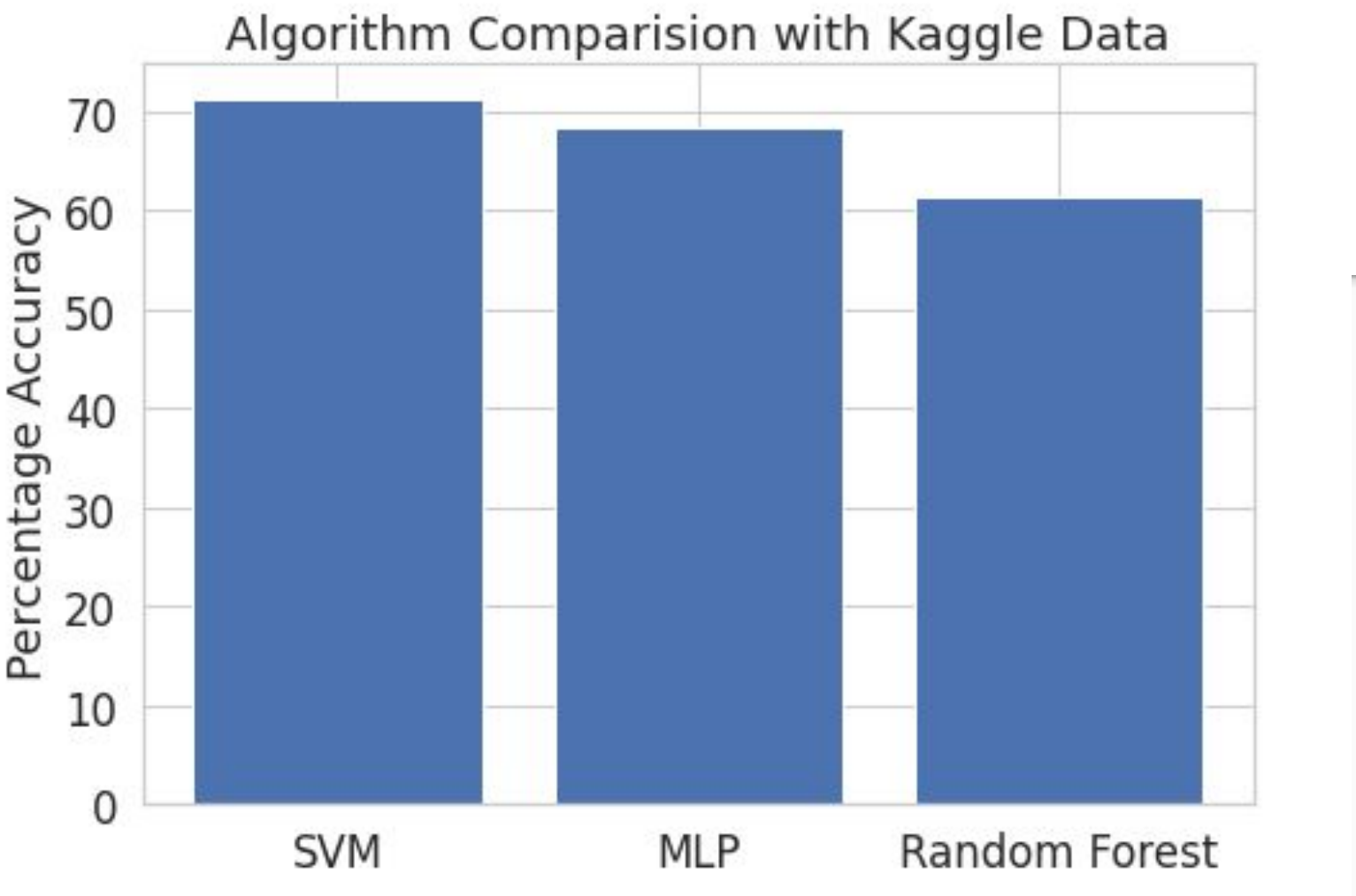
- A class of feed forward neural networks.
- 3 layers - input, hidden, output layers.
- Activation function
  for all nodes except input nodes.
- Backpropagation.
- MLP is distinguished from a linear perceptron by its numerous layers and non-linear activation.
- It can tell the difference between data that isn't linearly separable and data that is.



Random Forest

- Ensemble learning method for classification.
- The output of the random forest is the class selected by most trees.
- Random decision forests correct the decision trees' habit of overfitting to their training set.
- Performance is better than decision trees.



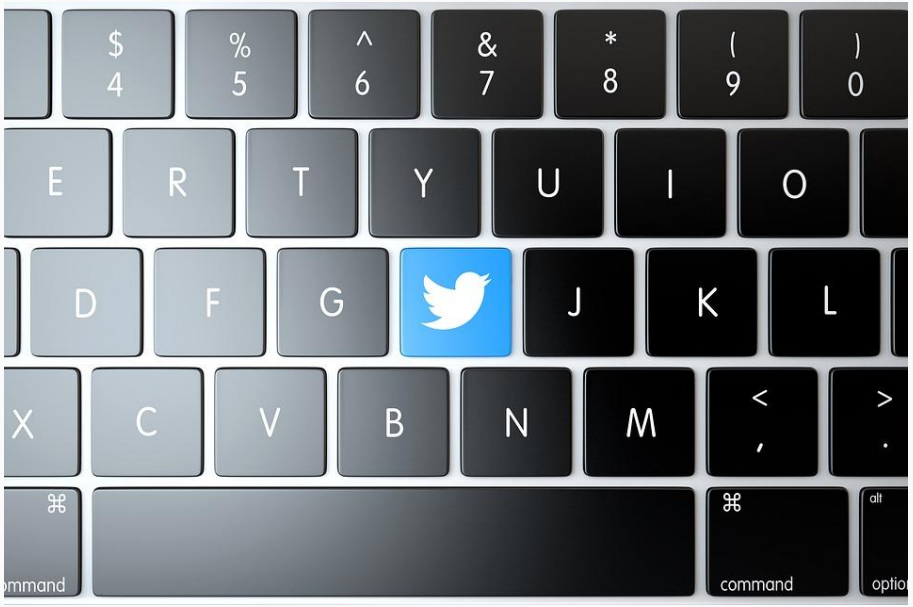

Algorithm Comparision with Kaggle Data
Percentage Accuracy
0
10
20
30
40
50
60
70
SVM
MLP
Random Forest

What to do after we train the classifier with train data?

After training we need to test the data to check the performance of each algorithm.

Now we test the data with the real time twitter data.

How do we do this?



Tweepy

Tweepy is the python client for the official Twitter API.

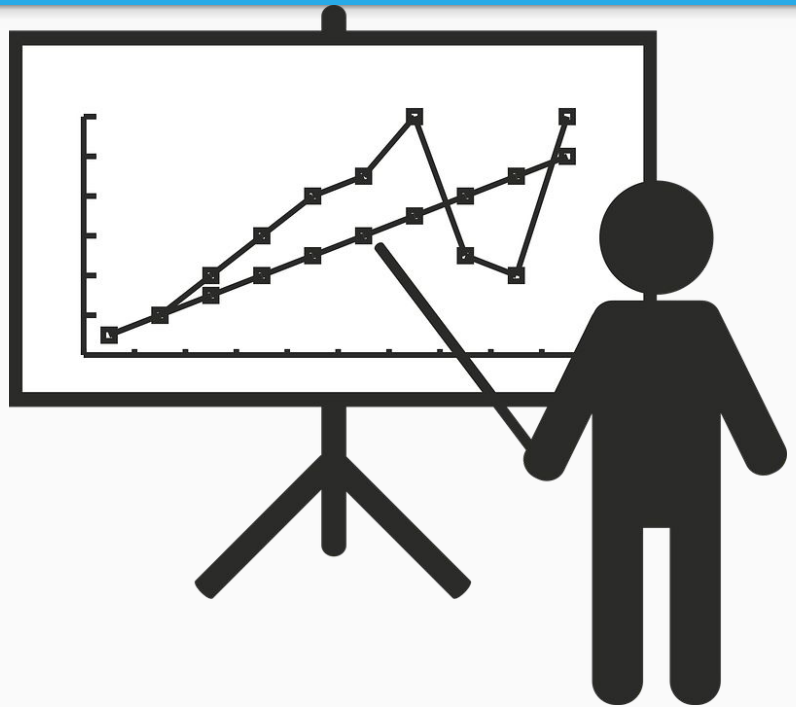
- The tweets need to be gathered so as to perform Sentiment analysis on those tweets. They can be fetched from Twitter using the Twitter API.
- In order to fetch tweets through Twitter API, one needs to register an App through their twitter account.
- Get the 'Consumer Key', 'Consumer Secret', 'Access token' and 'Access Token Secret'
- Establish connection with twitter get the tweets.

What next to do after getting data from twitter

We use Tweepy Cursor object in order to get the tweets from twitter based on a keyword.

For example we need to get the tweets for a movie to see whether the tweets have been positive, negative or neutral regarding the movie.

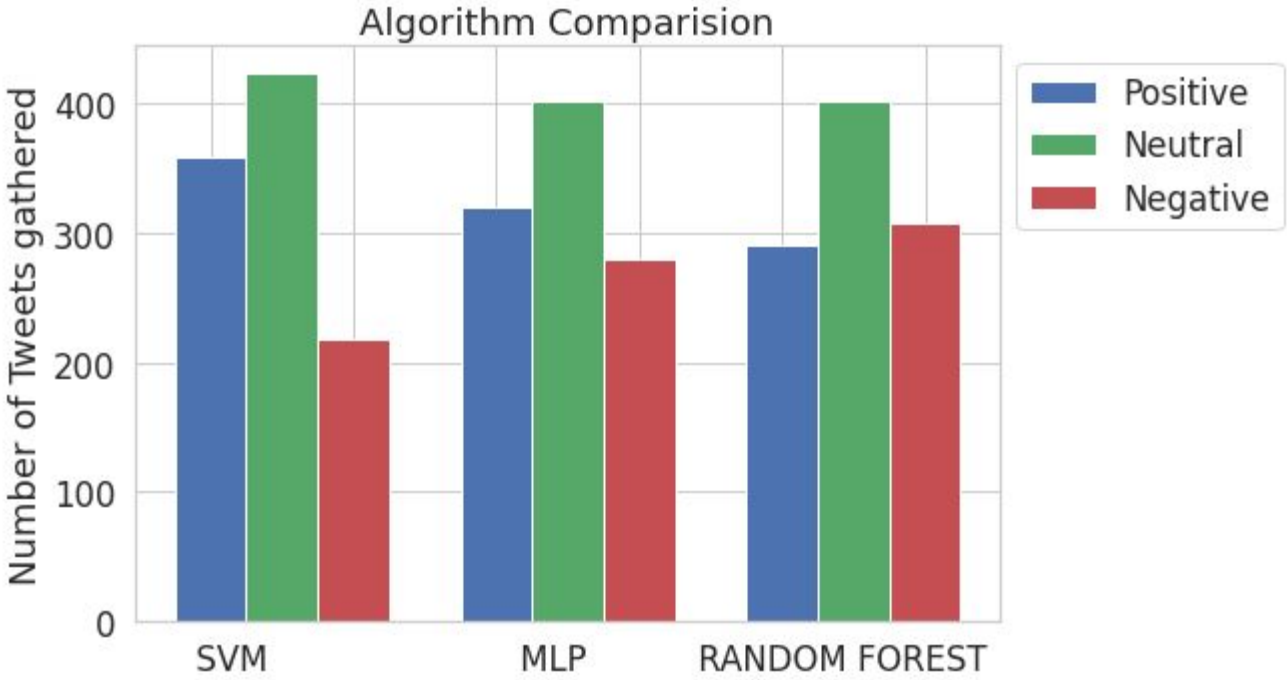
So the key word movie name “Fast and Furious” for example is passed as keyword for Tweepy cursor and then we predict from the tweets.



OLYMPICS

'Thank you for coming to Japan I supported them regardless of nationalityTOKYO2020 Olympics ARIGATO',
'Gold medalist weighs in on unprecedented Olympics as Games come to a closeOlympics TokyoOlympics',
'Delhi Gold medalist javelin thrower Neeraj Chopra and silver medalwinning wrestler Ravi Dahiya share a fist bump',
'Former Short Track Speed Skater Eddy Alvarez joined a highly selective group of summer and winter Olympic medal',
'India is the worlds largest importer of gold and holds the worlds largest private stock of yellow metal 25K ton',
'olympics why did tv coverage not show most of drone show for both ceremonies Amazing on YouTube but lousy on NBC',
'Sitting in a park on a sunny winters morningGlebe inner latte sipping SydneyIts good that Im healthy and do n',
'olympicchannel Olympics Cherry BombSong by The Runaways',
'So happy to see our cultureWearing your culture proudly playing it in a place where there is no special need',
'Wearing a Hijab can get you fired from your job in France but it wont stop you from winning a Gold Medal in the',
'There were so many athletes from around here that we got to cheer on in Tokyo Who knows maybe youll skate swim',
'what began as a marketing gimmick has evolved into a badge of honor thats legitimately respected by athletes',
'Were counting down to the 2024 Games Olympics Tokyo2020',
'What should be an Olympics Event but isnt',

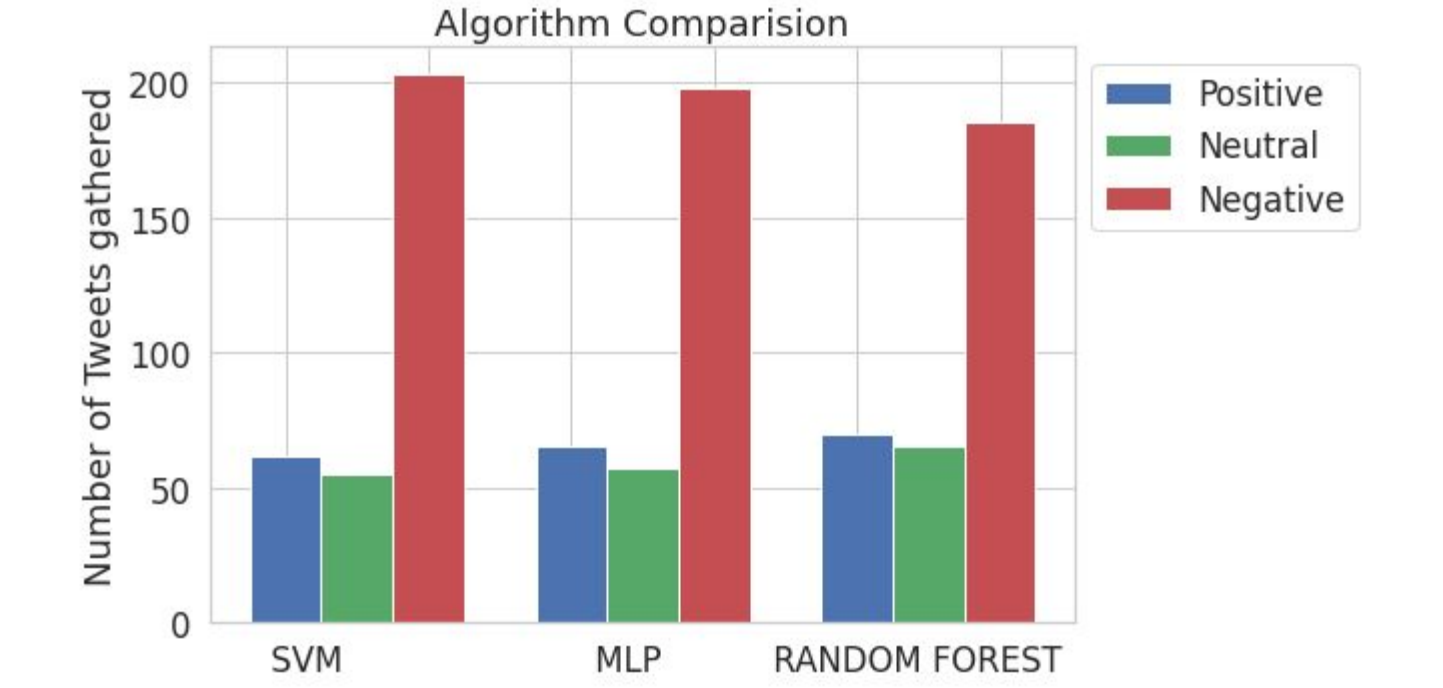
OLYMPICS


Algorithm Comparision
Number of Tweets gathered
0
100
200
300
400
SVM
MLP
RANDOM FOREST
Positive
Neutral
Negative

ANGRY

'Just trying to take a pic with Wesley Doesnt Dax look angry CatsOfTwitter',
'Check the link in our bio to listen to bryanvillain latest playlist',
'Wtf am I doingsketchbook sketch angry portrait',
'WERE GOING TO A DIFFERENT BUBBLE TEA PLACE angry',
'You do want your smartphone to look smart dont youThis Ragin iPhonecase is a shock absorbent TPU case with a',
'My Oh My Oh Mayadog showing competition westminster dogshow canine winner BichonFrise breeds furbaby',
'Hadith HadithOfTheDay ProphetMuhammad PBUH said Fatima is a part of me and he who makes her angry',
'EW I JUST SAW MY REFLECTION THIS IS WHY I HATE DARK MODE ANGRY DISGUSTED',
'Why is everyone on Twitter so angry and aggressive Twitter angry angrypeople',

ANGRY


Algorithm Comparision
Number of Tweets gathered
0
50
100
150
200
SVM
MLP
RANDOM FOREST
Positive
Neutral
Negative

Learnings

- Identify Anomalies
- Missing Data
- Mob Programming

Appendix

- Our project Twitter sentiment analysis achieved the goals and objectives we set for.
- After our project was completed we reflected on it and for better future development we would like to work on sarcasm sentiment analysis apart from positive, negative and neutral.

Thank you