CAPSTONE PROJECT

PROJECT TITLES

Coronavirus Tweet Sentiment Analysis

by

Sudhanshu Chouhan

Kapil Narayan Singh

Problem Statement

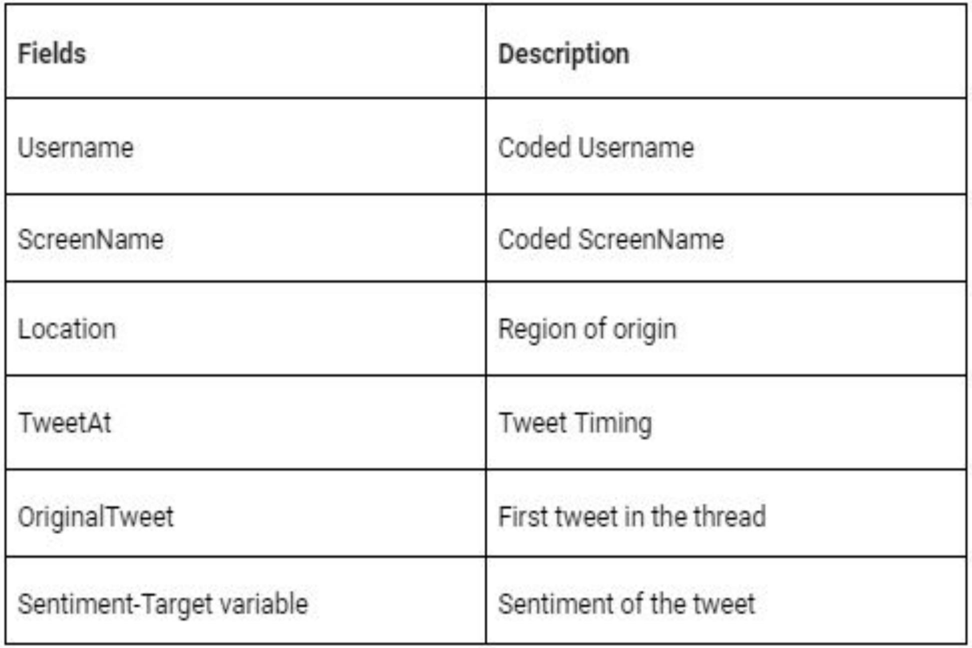
In this Project we have given the data set of tweets done by the people during the coronavirus pandemic from various countries. We have to see in this data what type of comment that has been done during the pandemic.
We get to know from our data that there are five types of sentiment that are extremely positive, extremely negative , positive, negative and neutral.
From various classification algorithm we have to find out that whether the tweet that has been made belongs to which type of sentiment.

Data Summary

- The original dataset has 6 columns and 41157 rows.

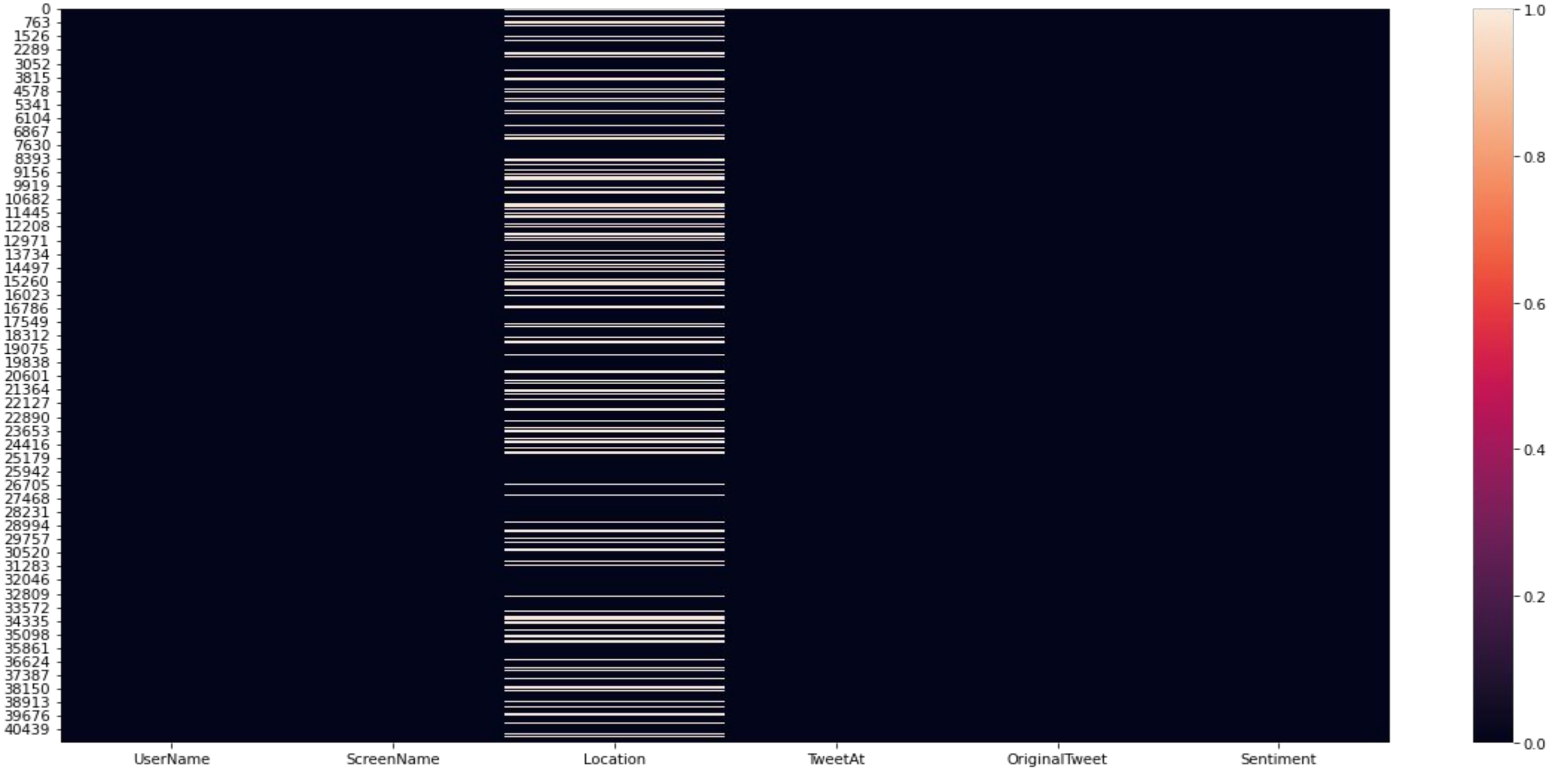
- In order to analyse various sentiments, We require just two columns named Original Tweet and Sentiment.
- There are four types of sentiments- Extremely Negative, Negative, Neutral, Positive and Extremely Positive.

Dataset Description



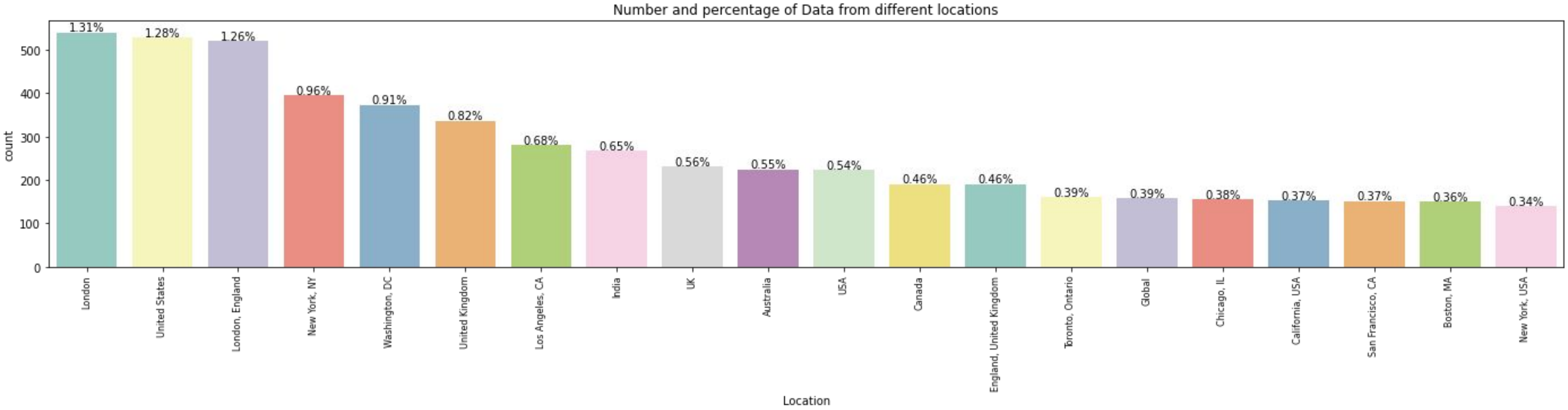
Fields	Description
Username	Coded Username
ScreenName	Coded ScreenName
Location	Region of origin
TweetAt	Tweet Timing
OriginalTweet	First tweet in the thread
Sentiment-Target variable	Sentiment of the tweet

Heatmap to check the Nan values



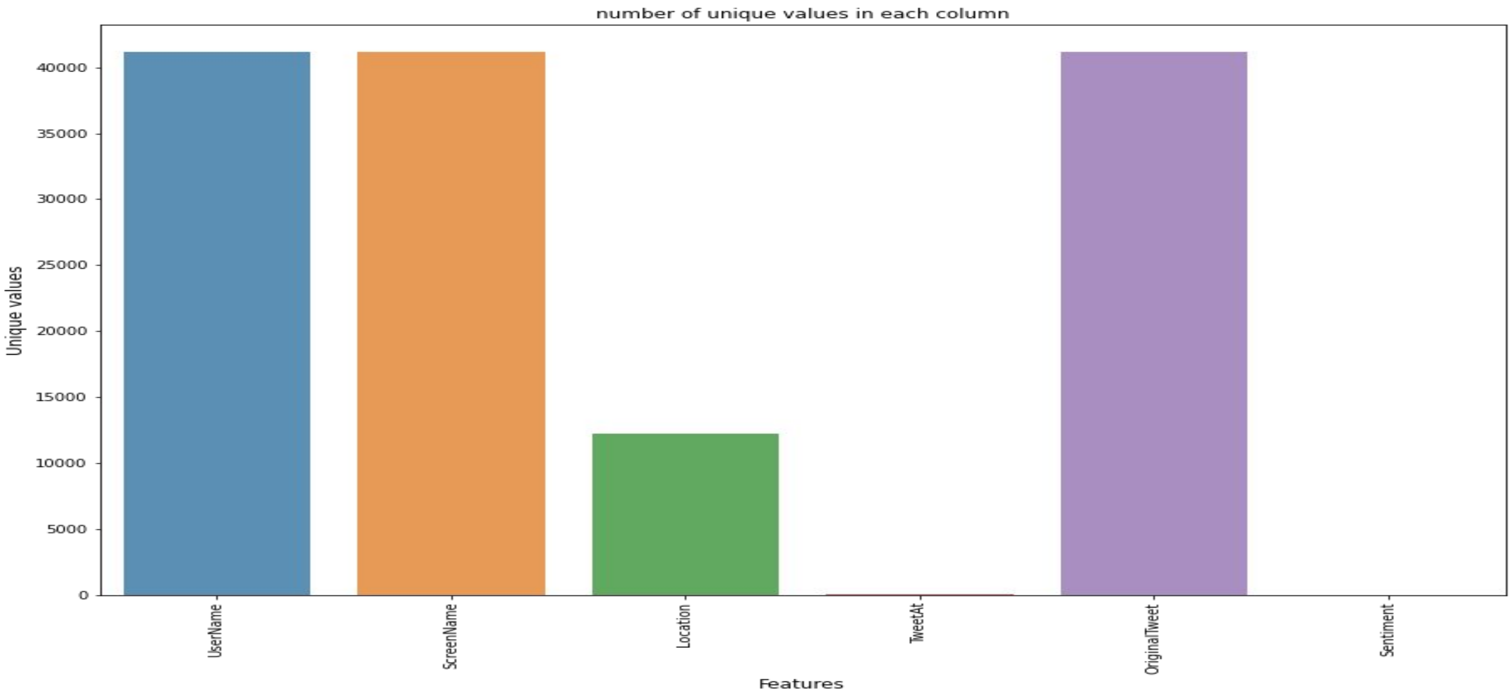
There are 20.87%(8567) null values in various places of location column.

Number and percentage of data from different locations

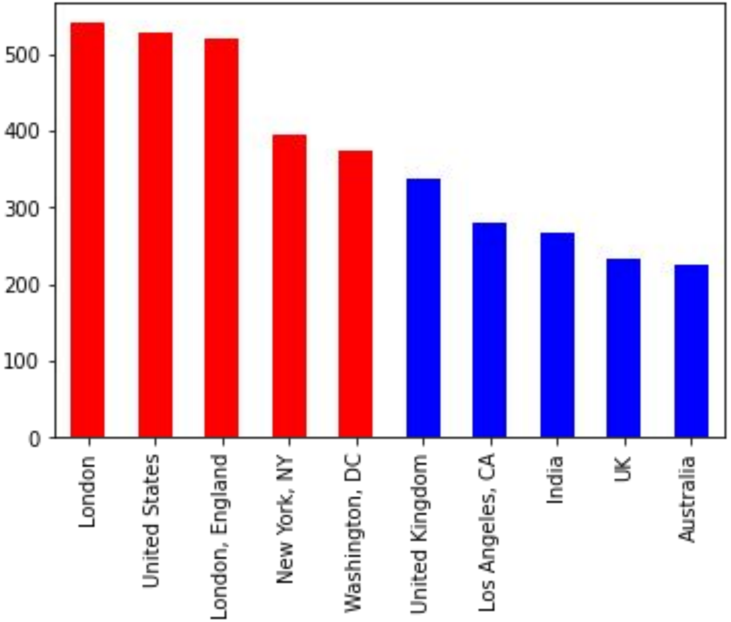


The above observation shows that the number of data percentage wise from each country that of total data present. We can see that mainly United kingdom,United states of america , Indian and Canada are the countries from where tweet has been done.

Number of unique values in each columns



The number of unique value in different feature are as followed Username , ScreenName , OriginalTweet has unique values of around 40000 then followed by Location that is between 10000-15000.

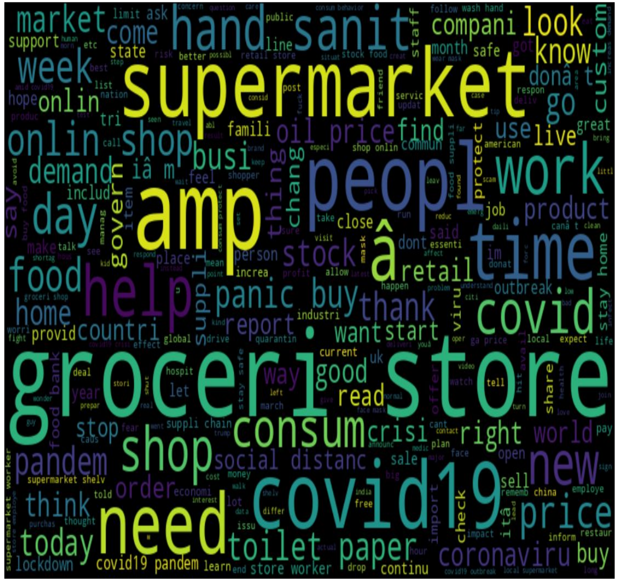


London has most number of unique tweet in developed cities and India has most number of tweet in list of developing nation.

Wordplay

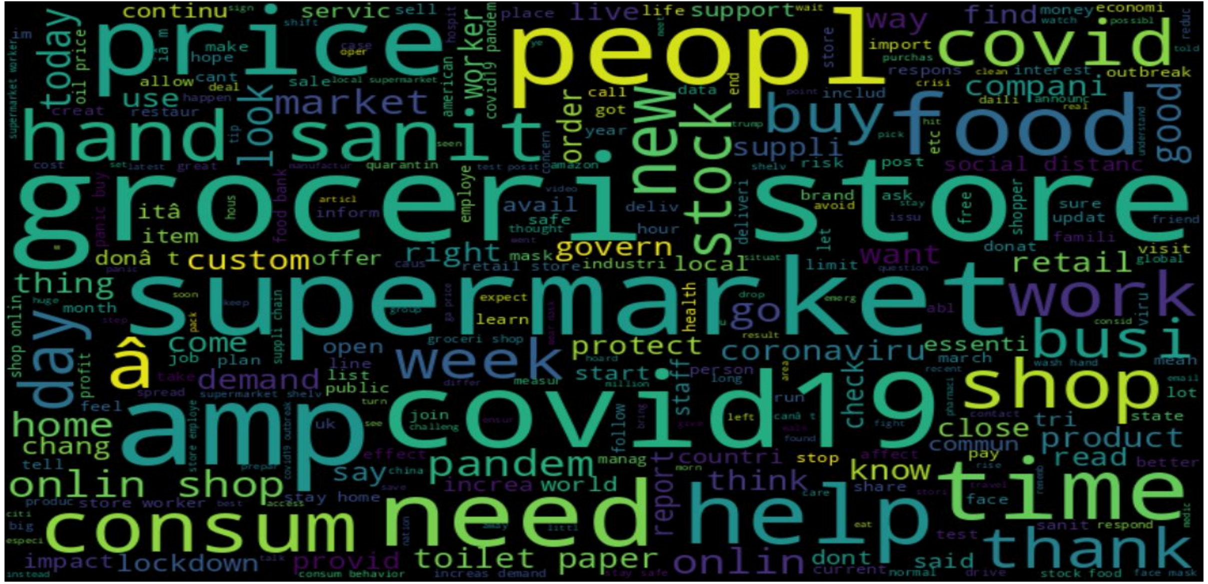
Wordcloud

All words wordcloud!



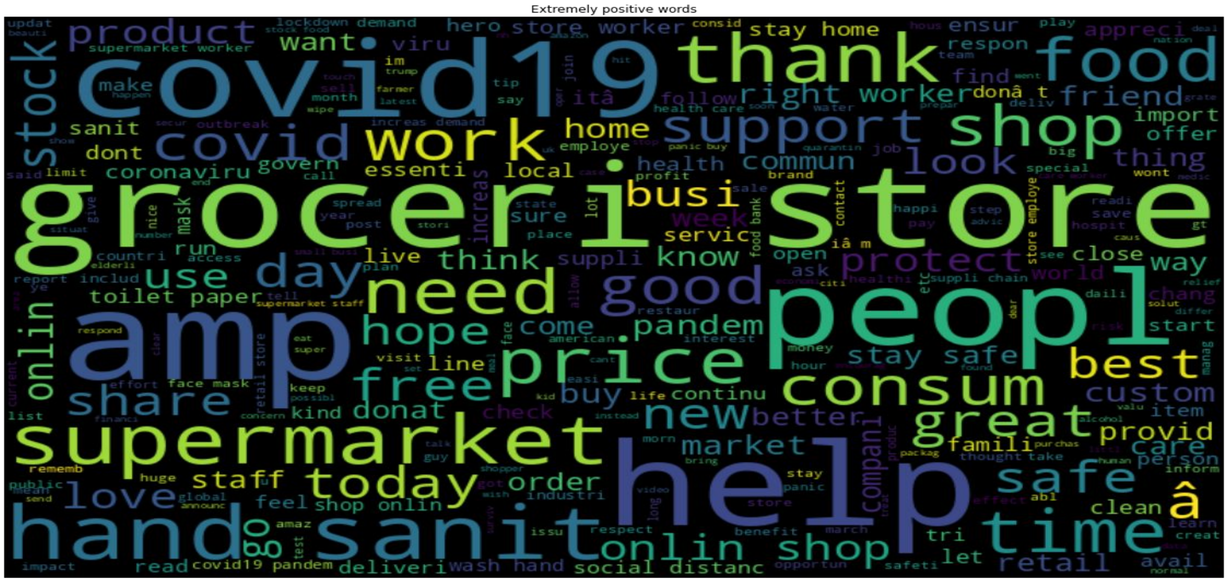
In this analysis we get to know which type of word is most frequent and how many times that they have been used during the pandemic while tweeting.

Positive words.


positive words

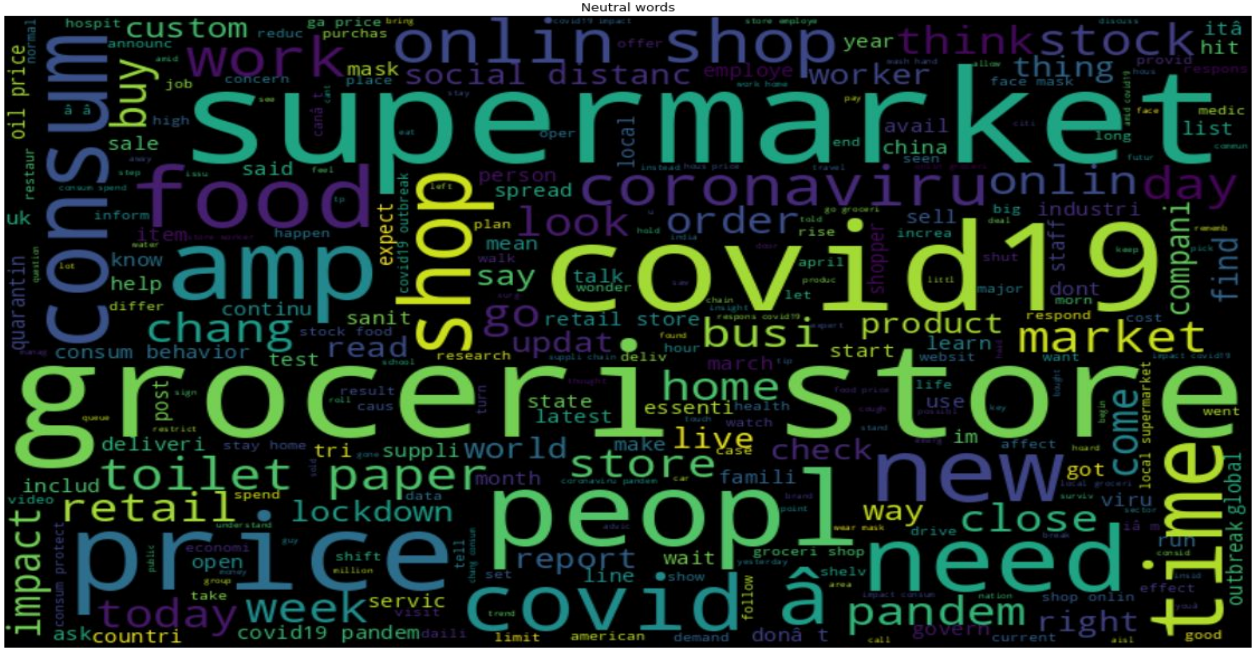
We can see that there is not much difference between the Positive words and extremely positive words.

Extremely Positive words



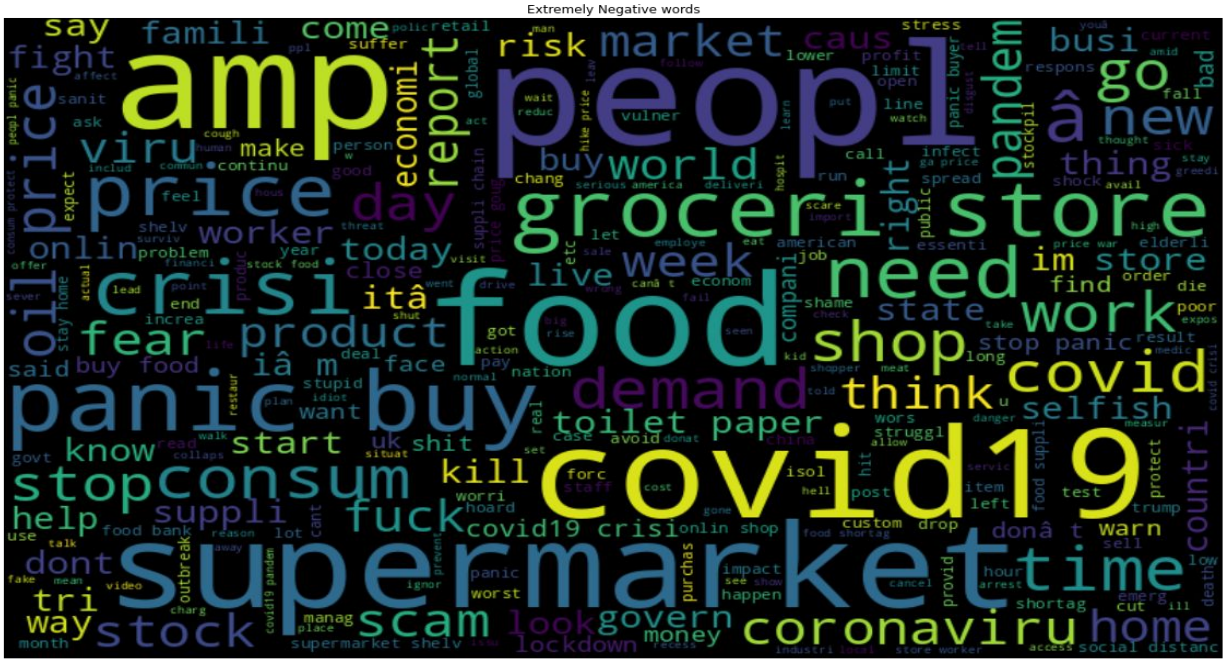
Here, like extremely positive and positive tweets an analogy has been developed the same analogy has been seen in negative types of tweets.

Neutral words



Neutral words follows as positive type of word composition and there is no such difference.

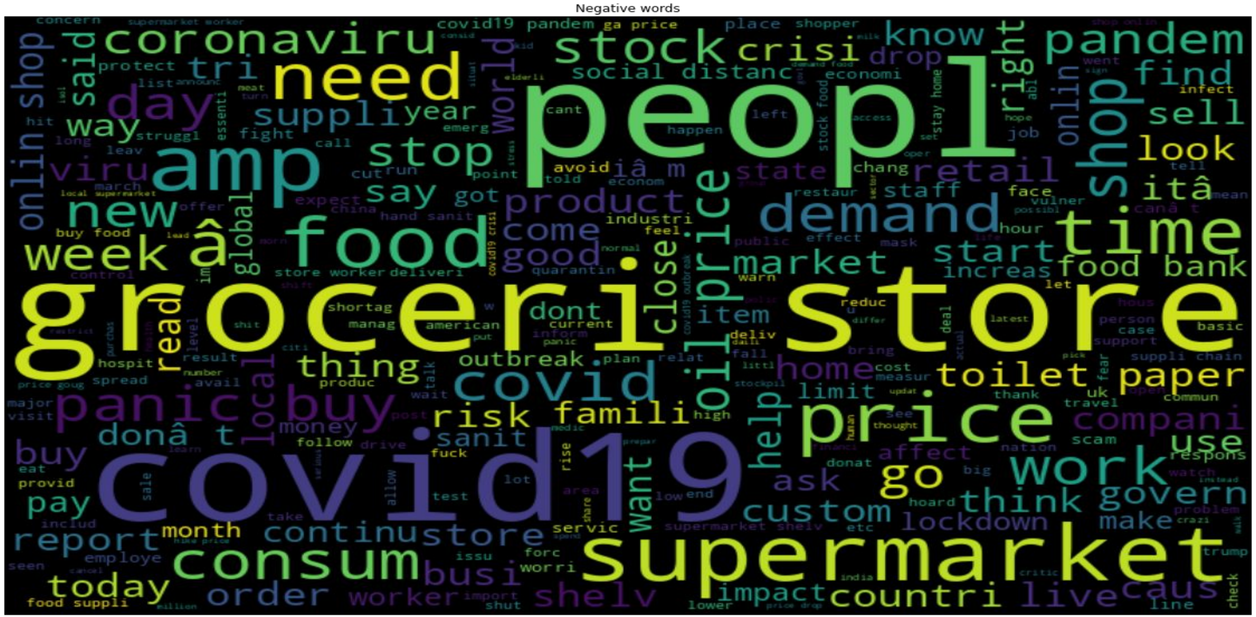
Extremely negative word



Extremely Negative words

From this analysis also we get to know that the same type of word has been used. There is no such difference that can be seen but the usage of negative word have increased like price, demand, crisis, oil prices and panic.

Negative words



Here, like extremely positive and positive tweets an analogy has been developed the same analogy has been seen in negative types of tweets.

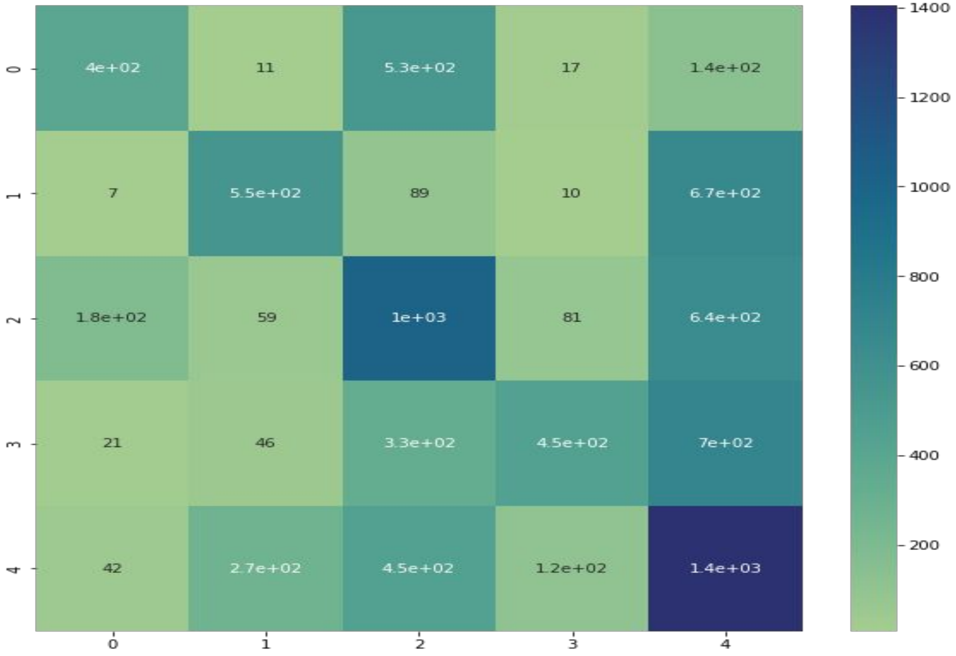
Classification

Models Used:

1. Naive Bayes
2. Logistic Regression
3. Random Forest
4. XGBoost
5. Support Vector Machines
6. Stochastic Gradient Descent

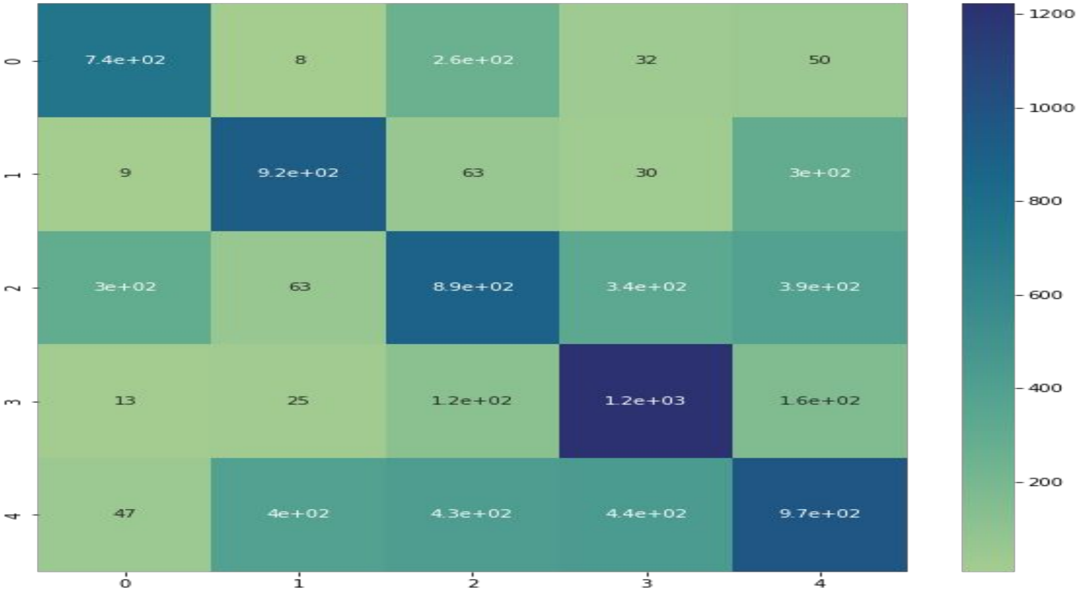
7.K-Neighbors Classifier

Naive Bayes



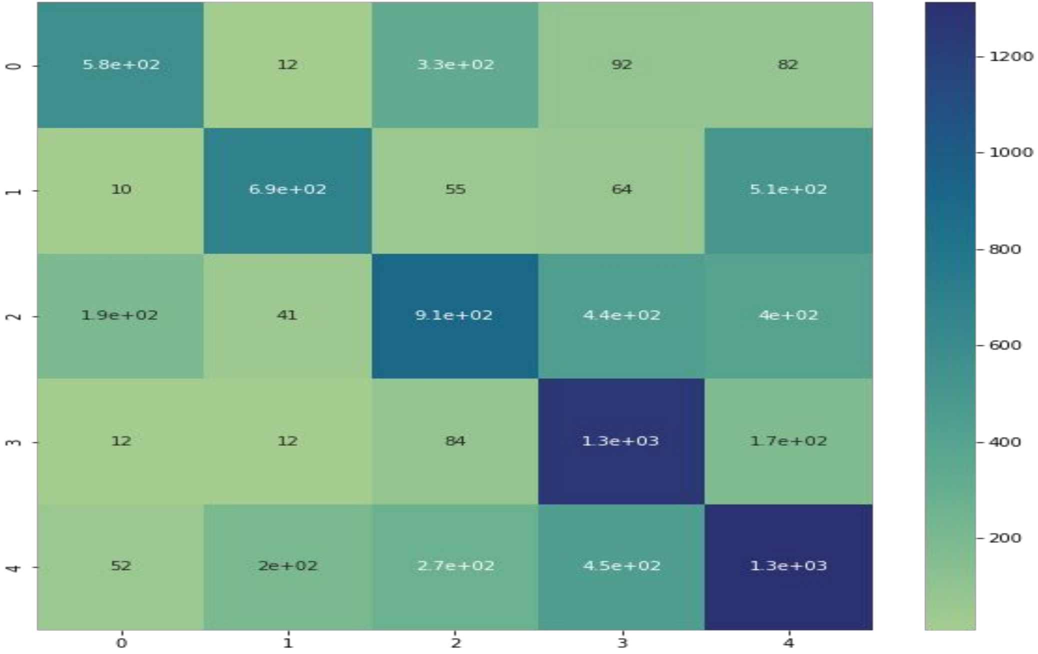
4e+02
11
5.3e+02
17
1.4e+02
7
5.5e+02
89
10
6.7e+02
1.8e+02
59
1e+03
81
6.4e+02
21
46
3.3e+02
4.5e+02
7e+02
42
2.7e+02
4.5e+02
1.2e+02
1.4e+03
0
1
2
3
4
0
1
2
3
4
1400
1200
1000
800
600
400
200

K-Neighbors Classifier



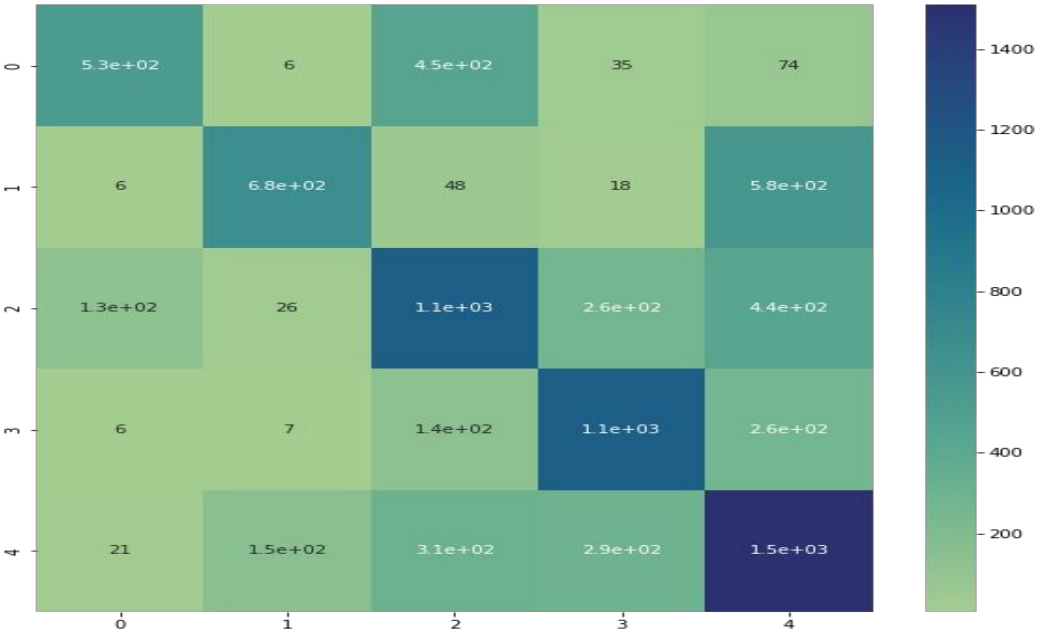
	0	1	2	3	4
0	7.4e+02	8	2.6e+02	32	50
1	9	9.2e+02	63	30	3e+02
2	3e+02	63	8.9e+02	3.4e+02	3.9e+02
3	13	25	1.2e+02	1.2e+03	1.6e+02
4	47	4e+02	4.3e+02	4.4e+02	9.7e+02

Stochastic Gradient Descent



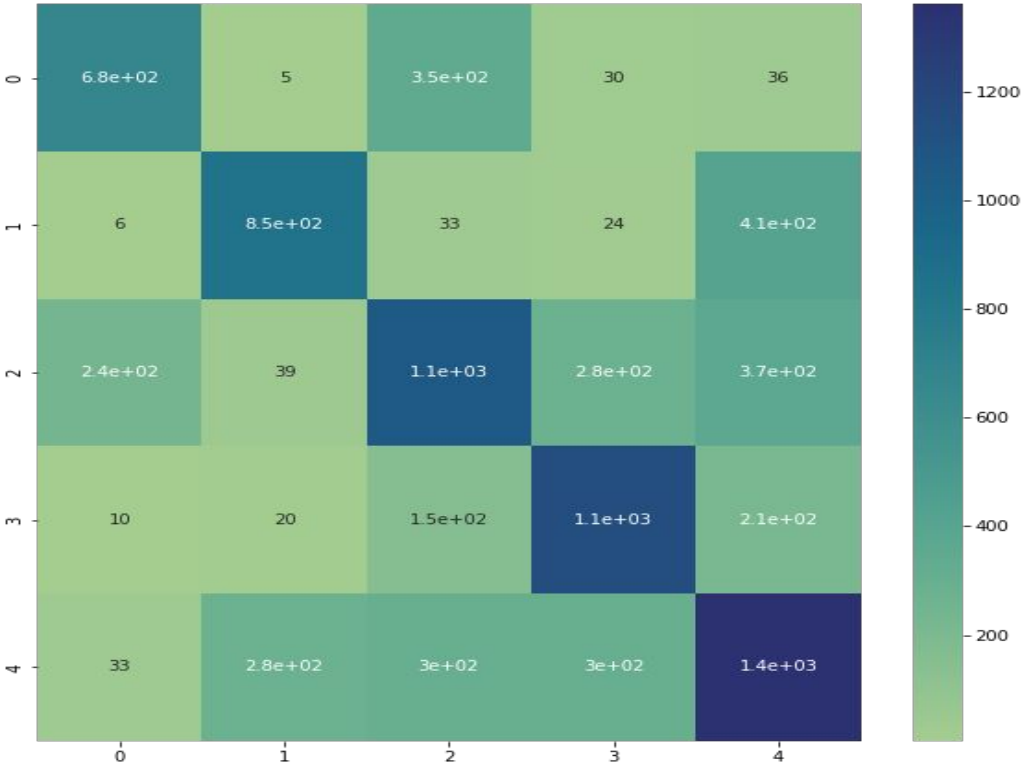
	0	1	2	3	4
0	5.8e+02	12	3.3e+02	92	82
1	10	6.9e+02	55	64	5.1e+02
2	1.9e+02	41	9.1e+02	4.4e+02	4e+02
3	12	12	84	1.3e+03	1.7e+02
4	52	2e+02	2.7e+02	4.5e+02	1.3e+03

Extreme Gradient Boosting



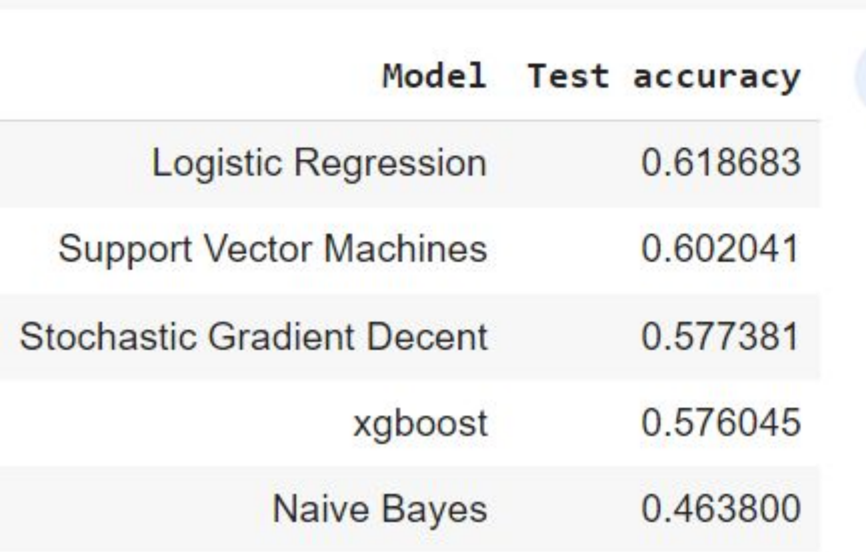
	0	1	2	3	4
0	5.3e+02	6	4.5e+02	35	74
1	6	6.8e+02	48	18	5.8e+02
2	1.3e+02	26	1.1e+03	2.6e+02	4.4e+02
3	6	7	1.4e+02	1.1e+03	2.6e+02
4	21	1.5e+02	3.1e+02	2.9e+02	1.5e+03

Logistic Regression



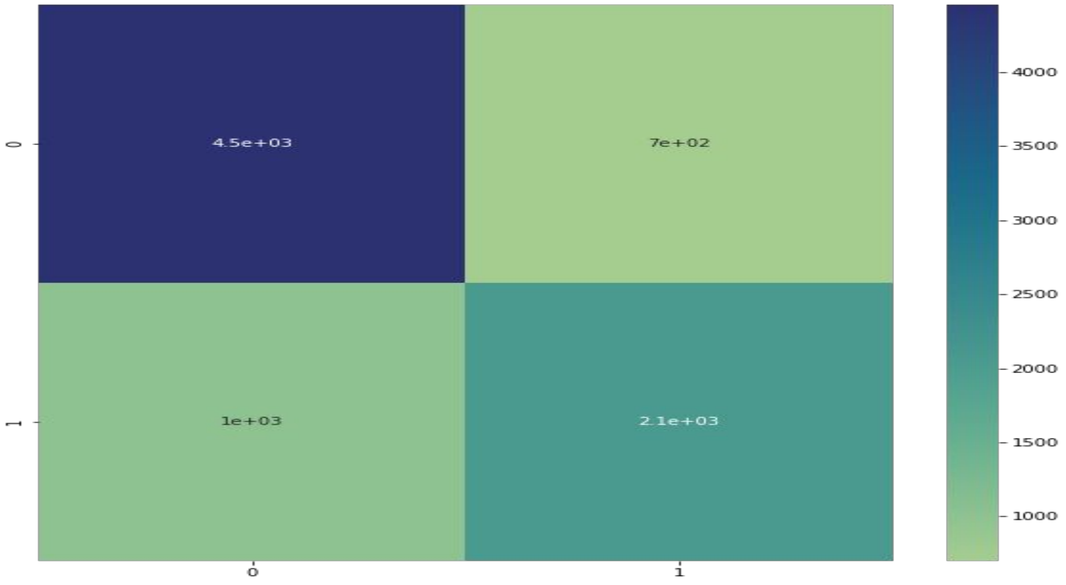
	0	1	2	3	4
0	6.8e+02	5	3.5e+02	30	36
1	6	8.5e+02	33	24	4.1e+02
2	2.4e+02	39	1.1e+03	2.8e+02	3.7e+02
3	10	20	1.5e+02	1.1e+03	2.1e+02
4	33	2.8e+02	3e+02	3e+02	1.4e+03

Results Of Different Algorithm On Multi Classification

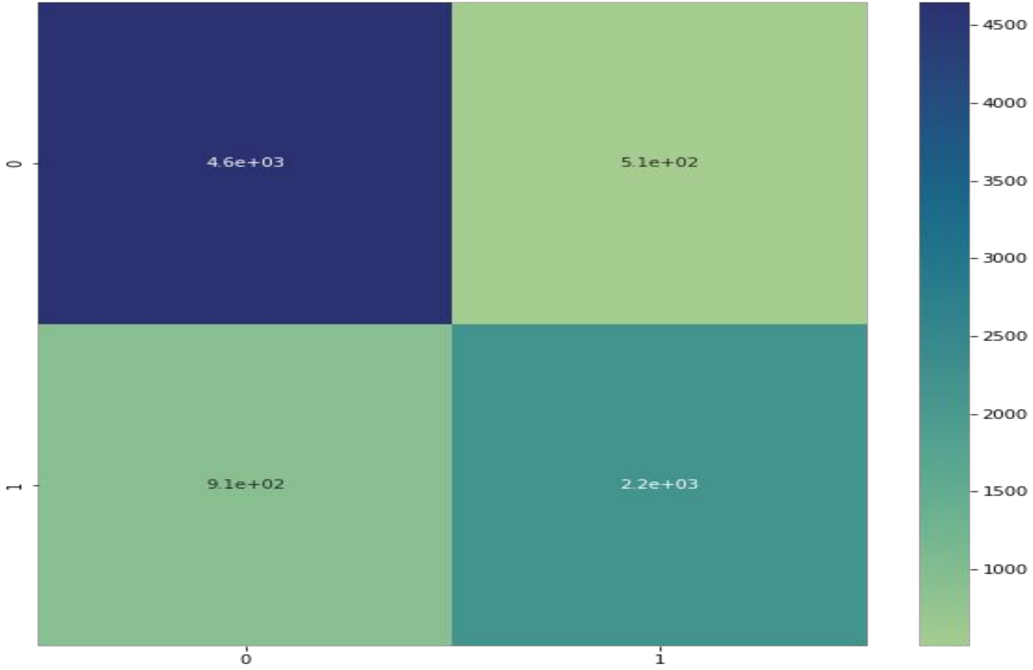


Model	Test accuracy
Logistic Regression	0.618683
Support Vector Machines	0.602041
Stochastic Gradient Decent	0.577381
xgboost	0.576045
Naive Bayes	0.463800

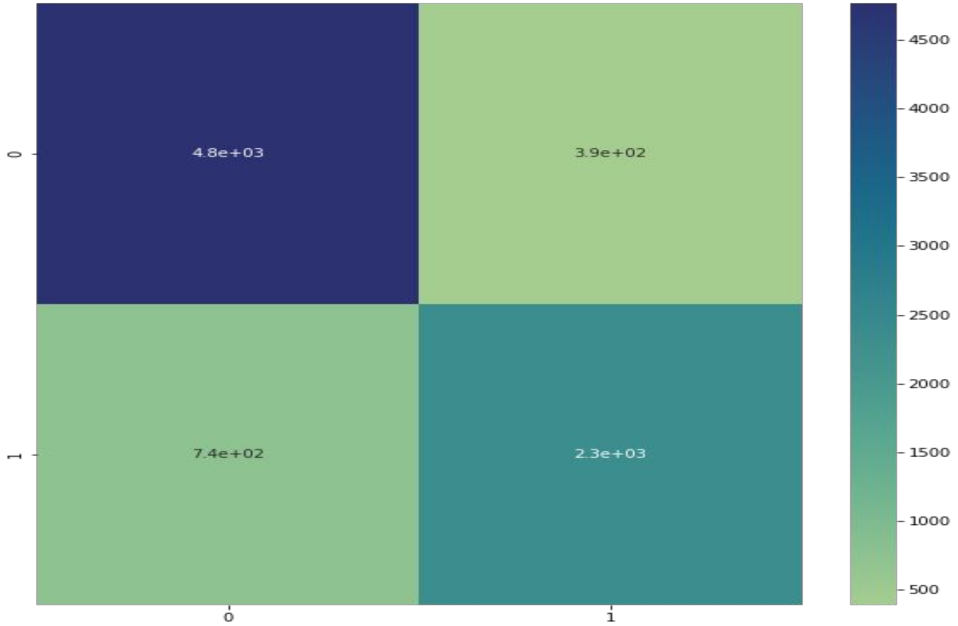
Naive Bayes


4.5e+03
7e+02
1e+03
2.1e+03
0
1
0
1
4000
3500
3000
2500
2000
1500
1000

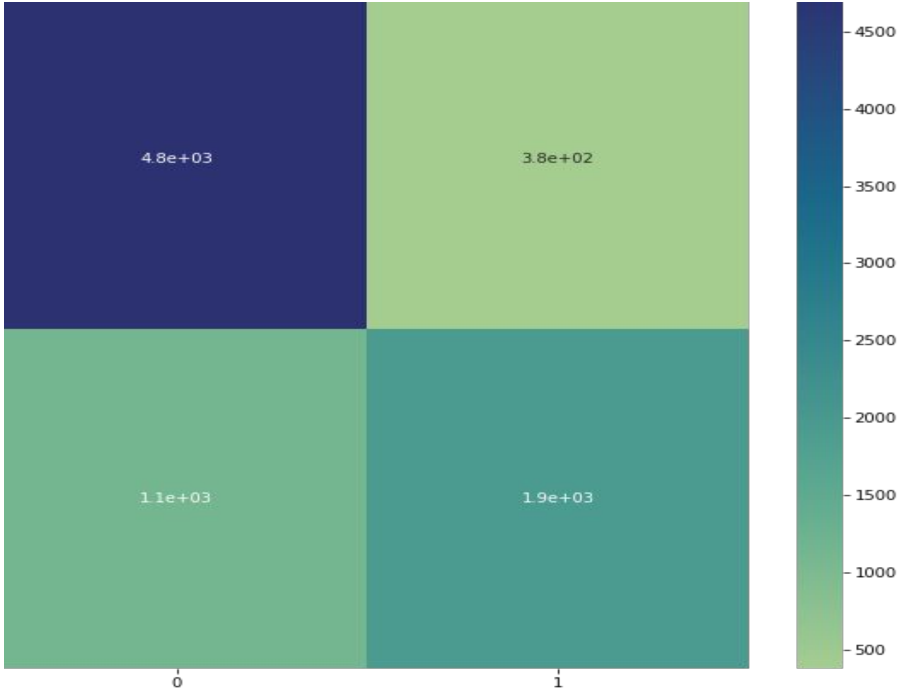
Random Forest Classifier



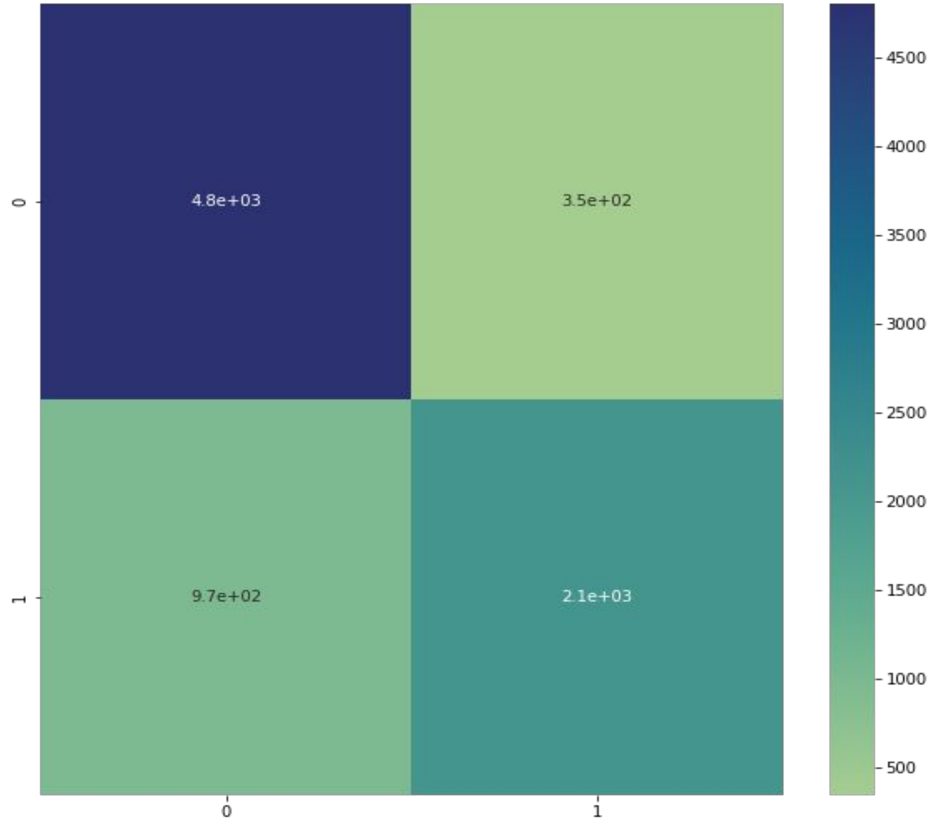
Logistic Regression



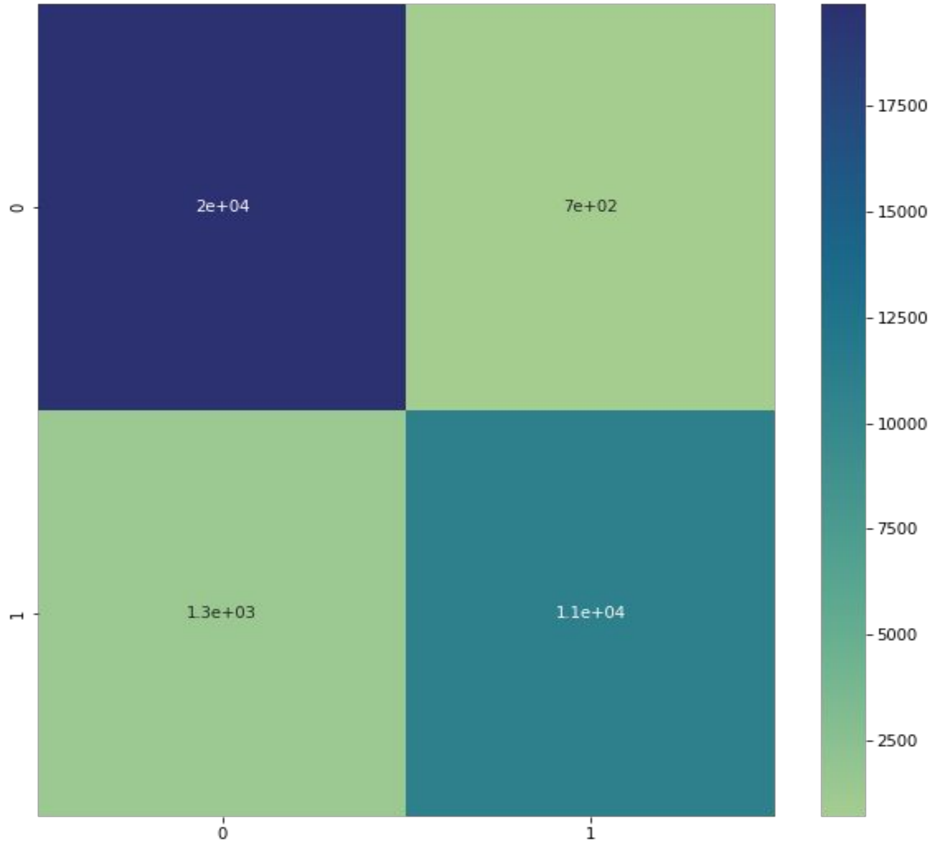
Extreme Gradient Boosting



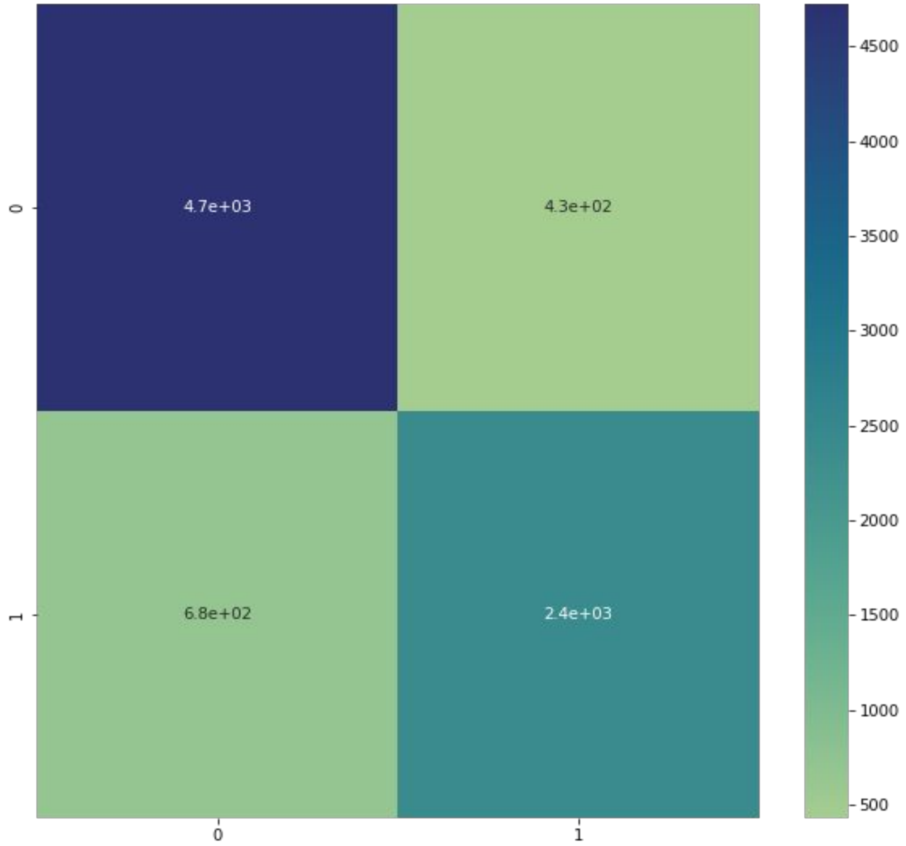
Support Vector Machines



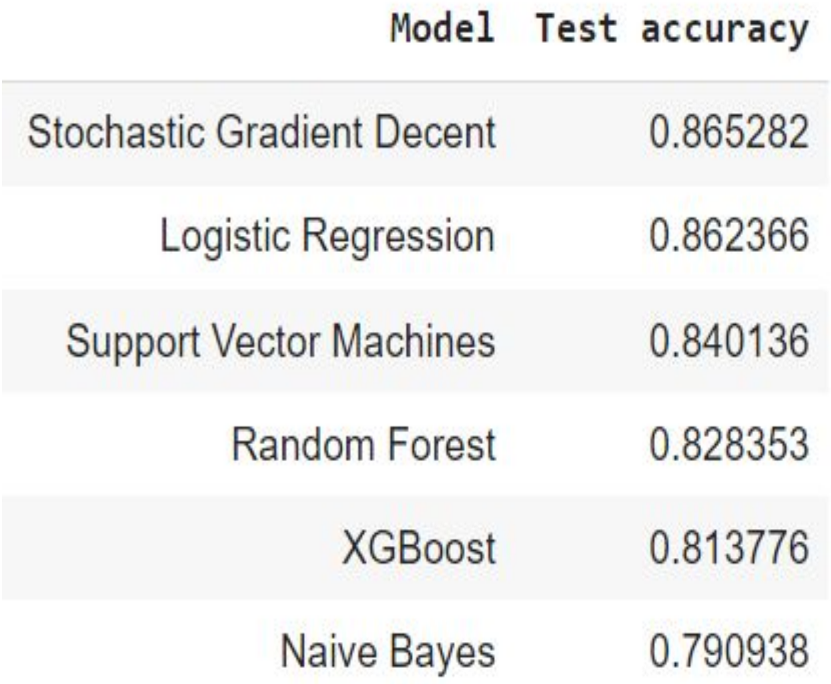
Stochastic Gradient Descent



Stochastic Gradient Descent



Results of binary classification of different algorithm



Model	Test accuracy
Stochastic Gradient Decent	0.865282
Logistic Regression	0.862366
Support Vector Machines	0.840136
Random Forest	0.828353
XGBoost	0.813776
Naive Bayes	0.790938

** Conclusion**

From above discussion we first evaluated the the model on the basis of multipleclassification where we have taken multiple sentiments into consideration which includes positive ,negative,neutral, extremely positive and extremely negative sentiment.

- After applying various classification algorithm we come to a conclusion that the best is logistic classification with a score of 61 percent.
- After multiple classification we divided our data into binary classification, in which extremely positive, positive and netral are taken as 0 and extremely negative and negative are taken as 1 for classification.
- The output after Deploying various Model into the system the best algorithm comes out to be Stochastic Gradient Descent.
- We have seen confusion matrix for individual cases prediction and seen that lighter part that is i has less frequency and darker part has high frequency.