Capstone Project - 3

Project Title – Coronavirus Tweet Sentiment Analysis

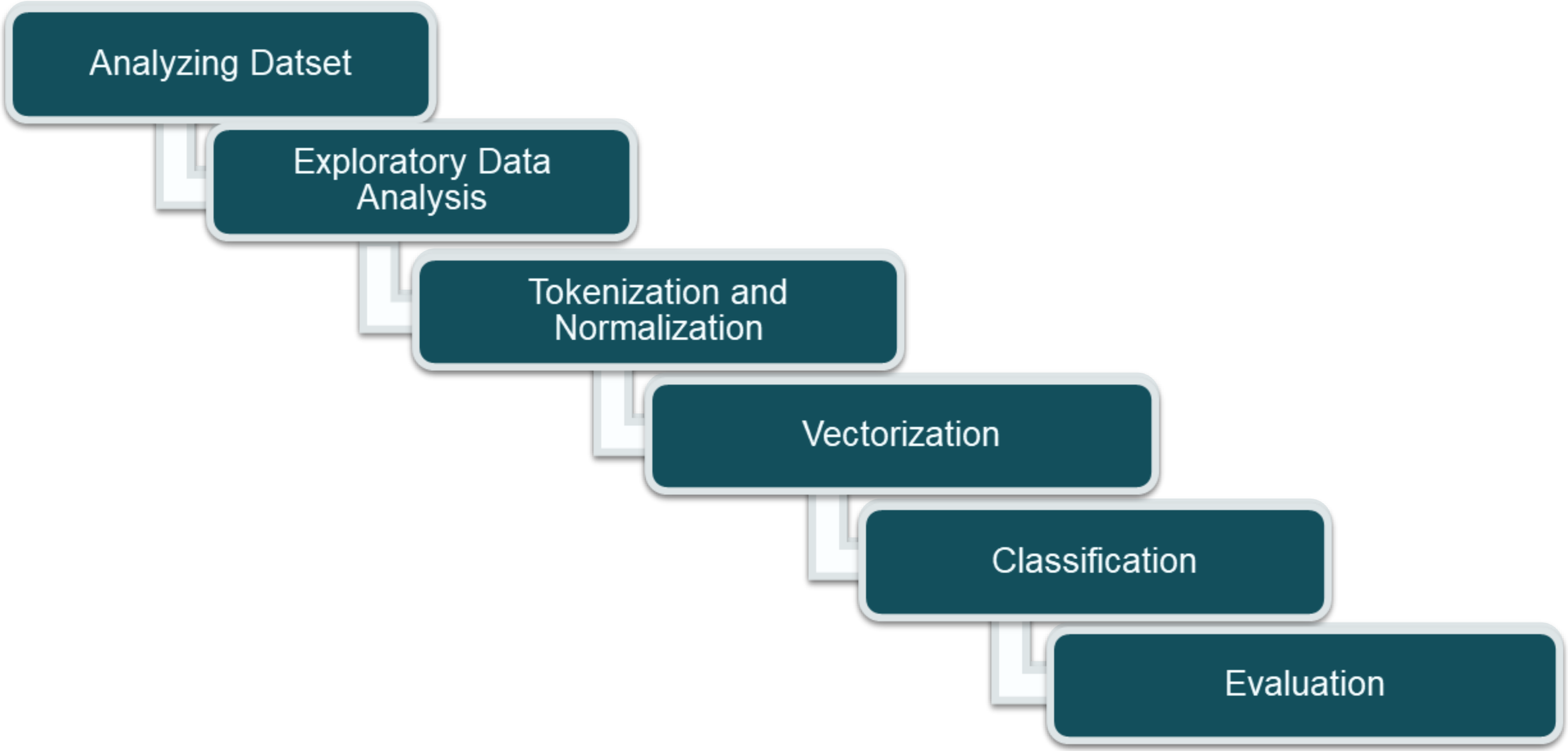
Team Members

Nivya T

Manjushree K C

Shaurabh Pandey

STEPS INVOLVED



Analyzing Datset

Exploratory Data Analysis

Tokenization and Normalization

Vectorization

Classification

Evaluation

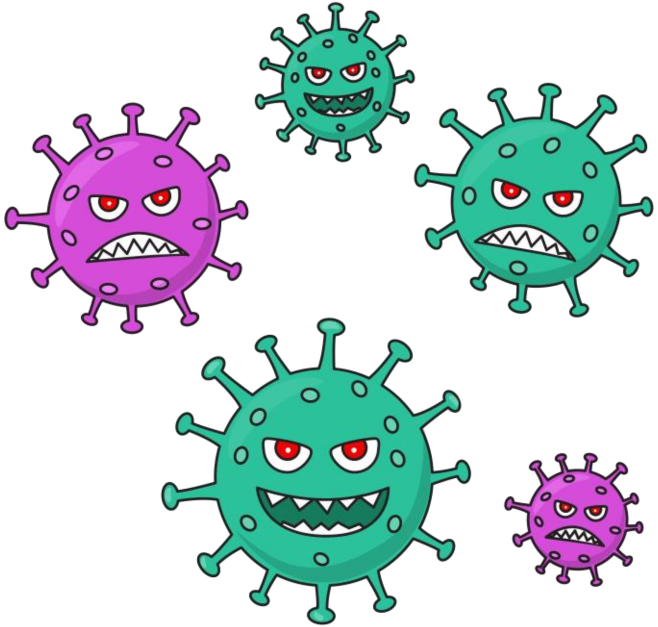
COVID 19

- Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. It was discovered in December 2019, it is very contagious and has quickly spread around the world.

- The virus can cause mild to severe respiratory illness, including death. The best preventive measures include getting vaccinated, wearing a mask, staying 6 feet apart, washing hands often and avoiding sick people

- COVID-19 was declared a global pandemic on March 11, 2020. As of January 23, 2022, over 346 million cases including over 5.5 million deaths have been reported worldwide.

.



WHAT IS SENTIMENT ANALYSIS

Sentiment Analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic is Positive, Negative, or Neutral.

Sentiment analysis provides answers into what the most important issues are. Because sentiment analysis can be automated, decisions can be made based on a significant amount of data rather than plain intuition that isn't always right.

TWEET SENTIMENT ANALYSIS

Twitter is one of the most powerful social media platform in the world right now, which is used every day by people to express opinions about different topics, such as products, movies, music, politicians, events, social events, among others.

Tweets are often useful in generating a vast amount of sentiment data upon analysis. These data are useful in understanding the opinion of the people about a variety of topics.

PROBLEM STATEMENT

COVID 19 is a global pandemic that is still infecting millions of people around the world.

For this project we are provided with a coronavirus tweets csv file which contains more than 40000 tweets from people around the world on covid 19 and our aim is to analyze these tweets made on Covid-19 from around the world and predict the sentiment of each of the tweet by classifying them into three categories positive, negative and neutral.

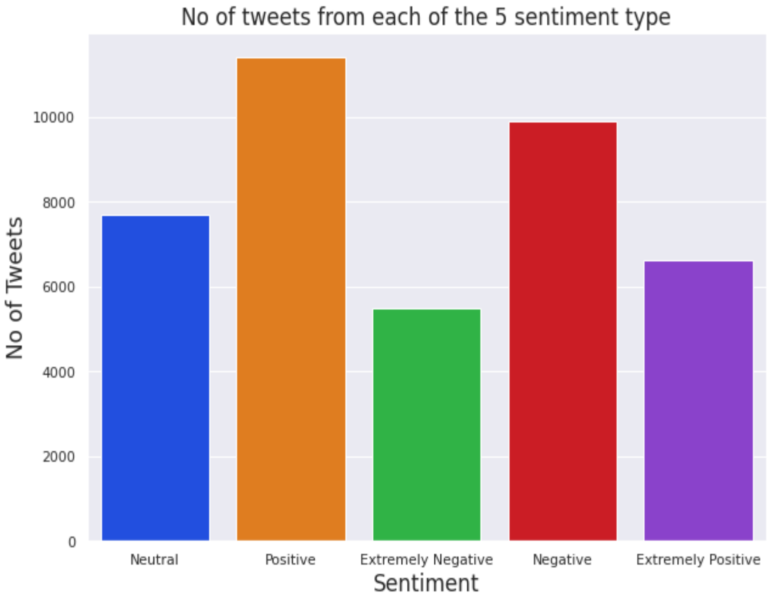
OBJECTIVE

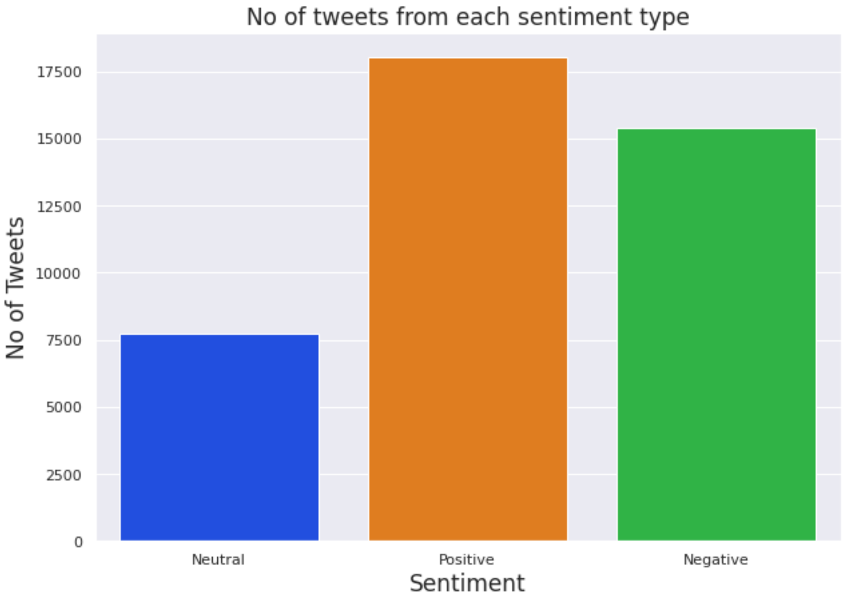
Analyze the tweets regarding COVID 19 and get insights regarding people’s sentiment.
To build a classification model to predict the sentiment of COVID-19 tweets which have been pulled from Twitter.

Defining the variables

- Username : The username of the person on on twitter
- Screenname : The screenname of the person on twitter
- Location : The location from where the tweet was tweeted
- TweetAt : The date of the tweet
- OriginalTweet : The tweet itself unfiltered
- Sentiment : The sentiment of the tweet our target variable

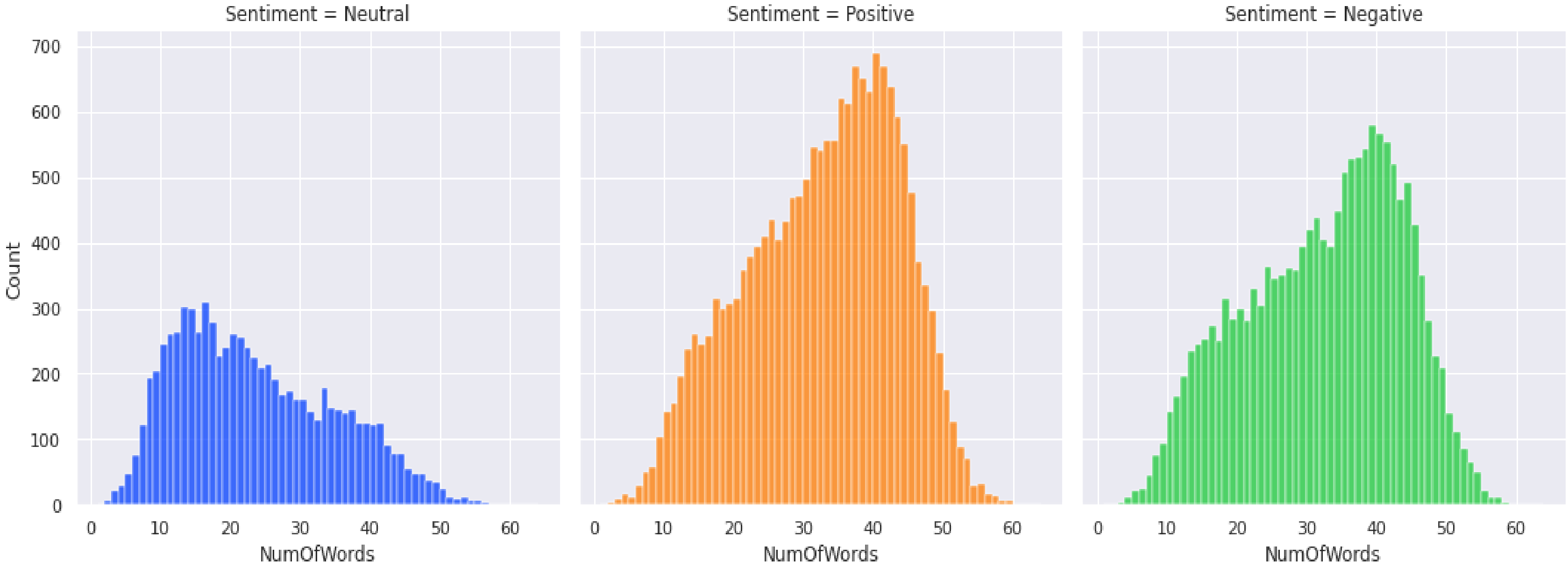
Target Variable Study





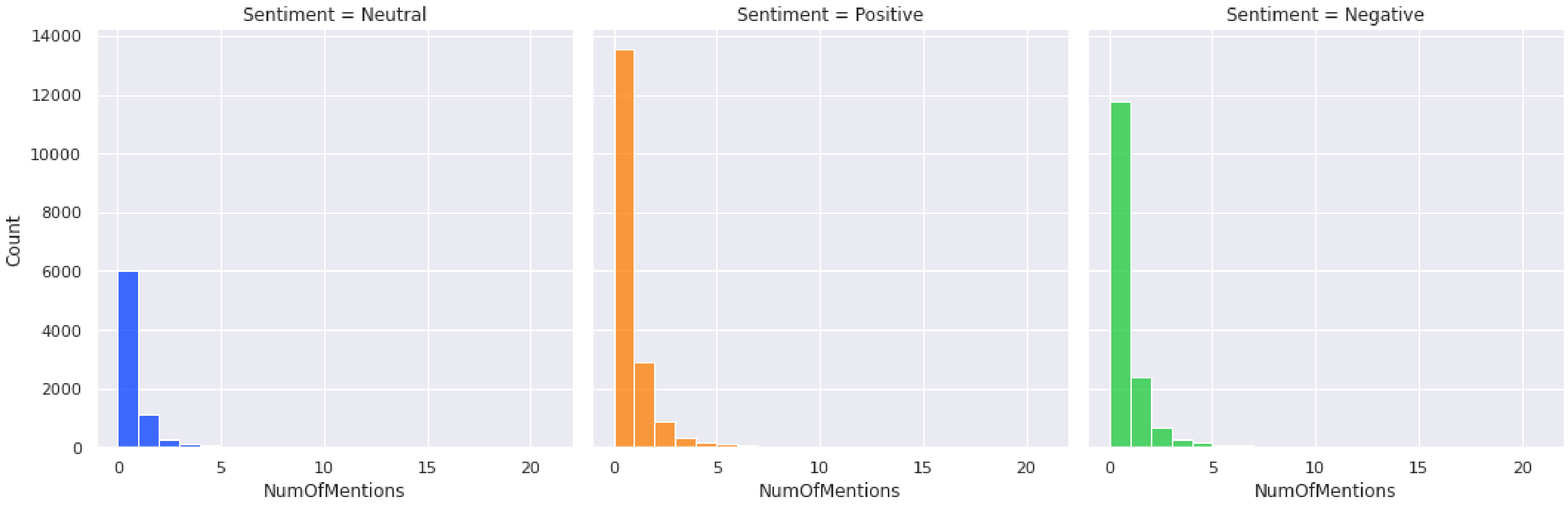
Majority of the tweets in our dataset are of positive sentiment , also the number of neutral sentiment tweets are quite less in number as compared to other two sentiments

Tweets Analysis (Words)



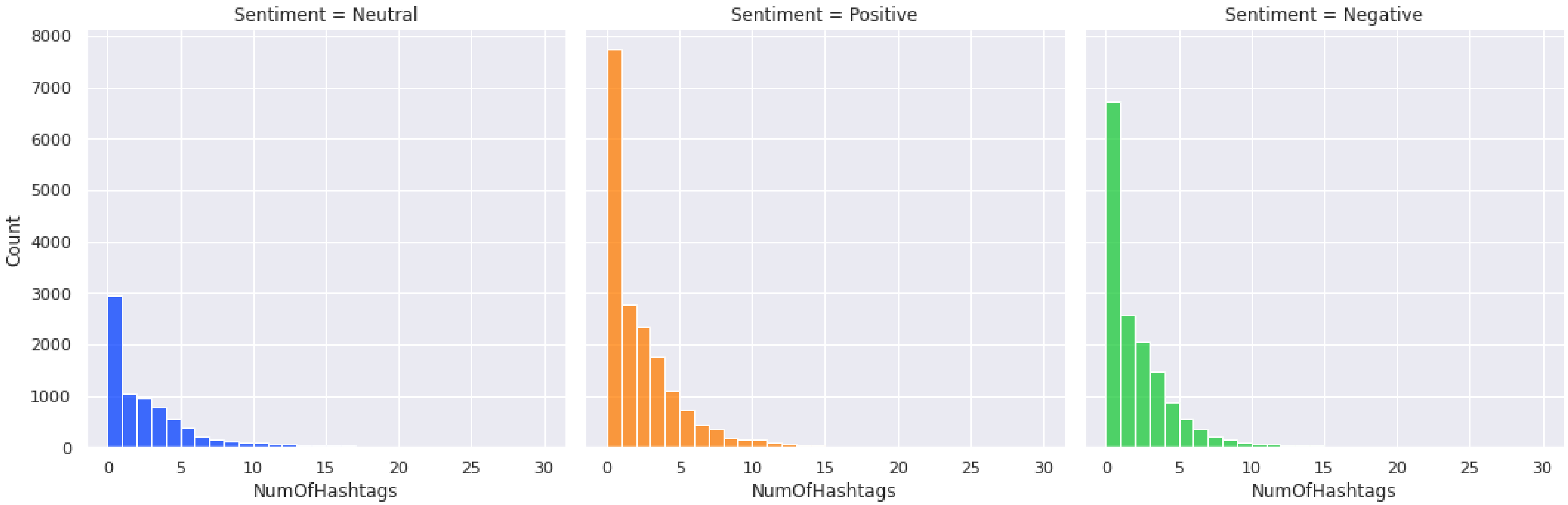
On an average tweets of positive sentiment have a higher words count if compared with tweets of negative sentiment . Neutral Sentiment tweets have a much lesser word count in comparison to positive and negative sentiment tweets

Tweets Analysis (Mentions)



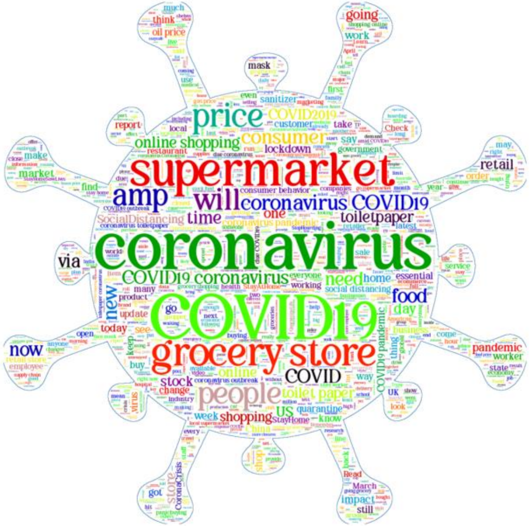
Very few of the tweets be of any sentiment type had mentions in them , also there is no difference in number of mentions with regard to the sentiment

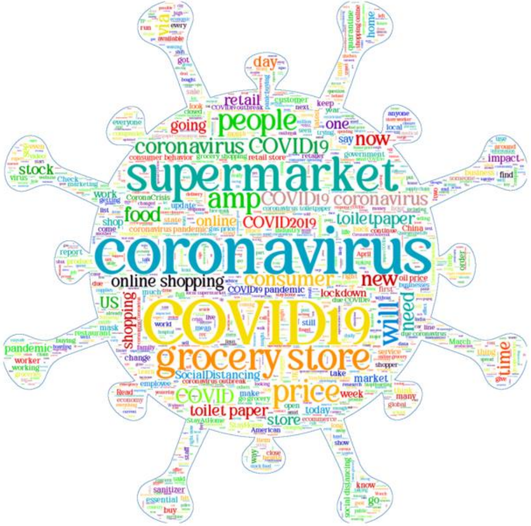
Tweets Analysis (Hashtags)

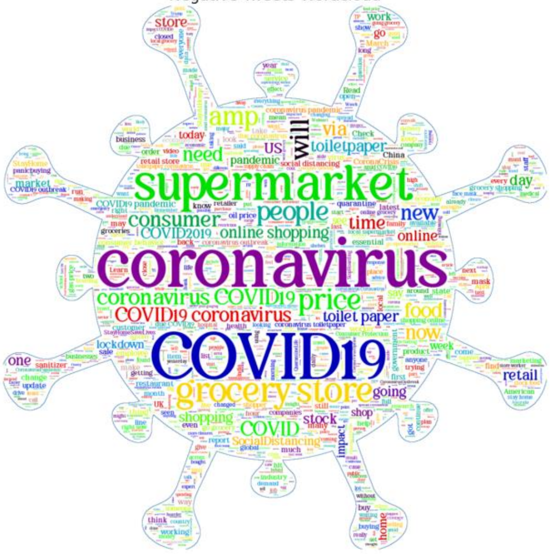


Slightly more than half the number of tweets in our dataset had some kind of hashtags in them and the average number is almost constant for tweets of all sentiment types

Tweet Wordcloud

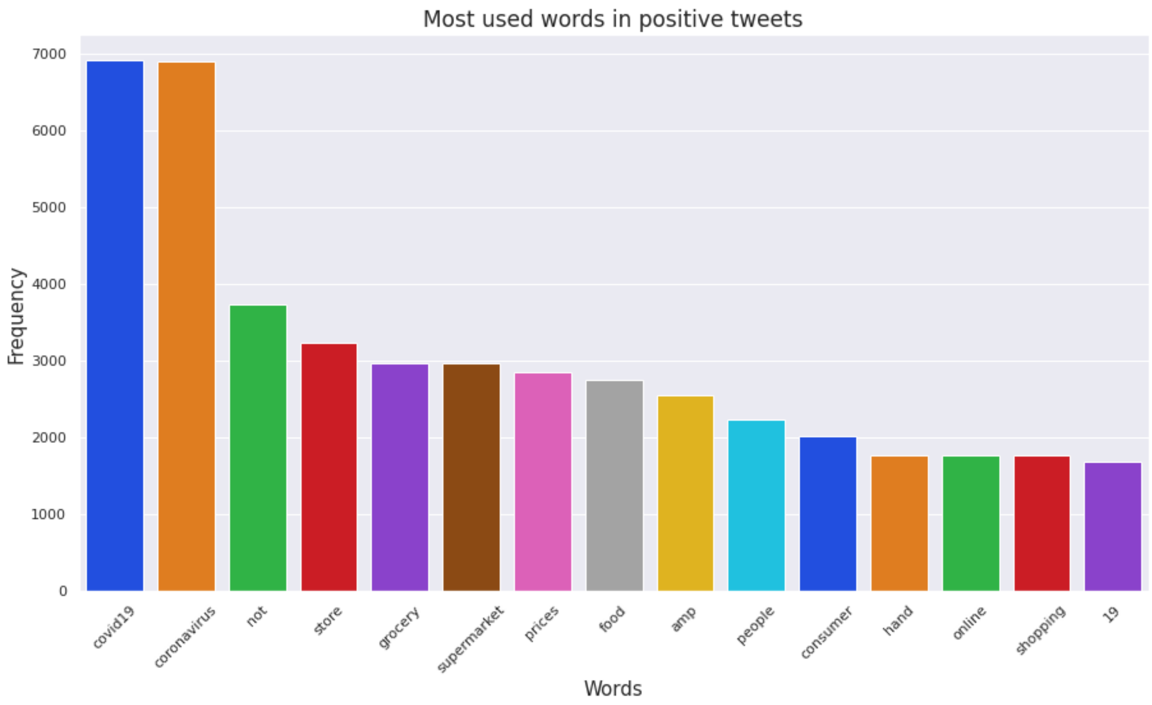

Positive Tweets Wordcloud


Neutral Tweets WordCloud


Negative Tweets Wordcloud

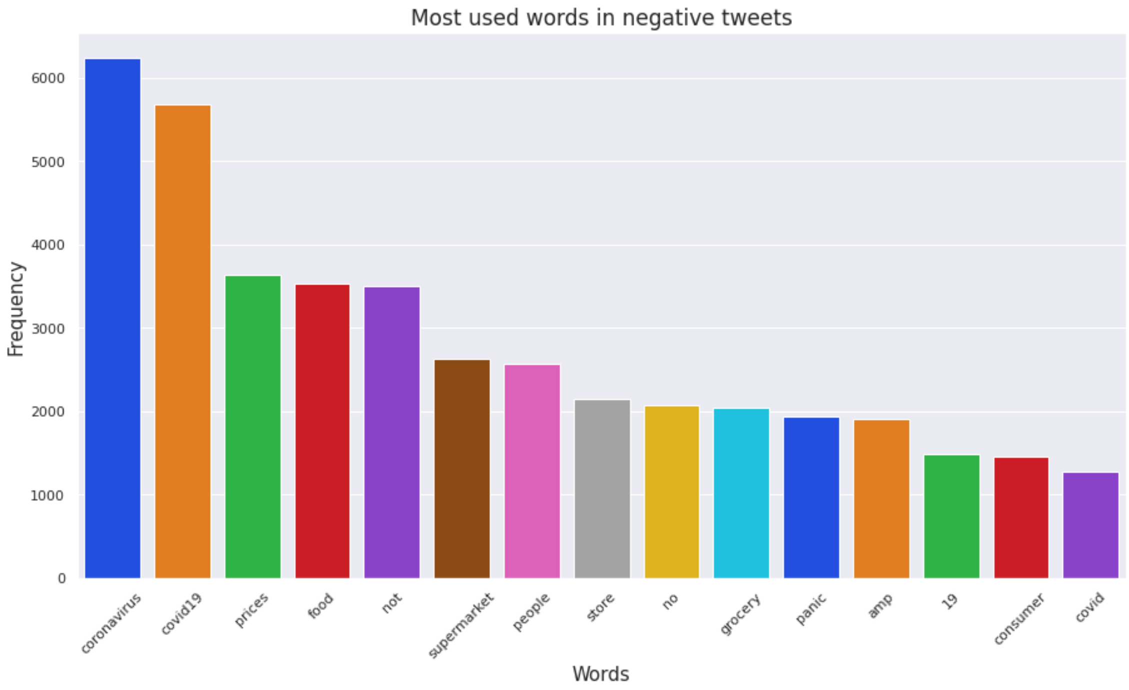
In all the tweets be of any sentiment the most frequently used words apart from the name of the pandemic and disease are supermarket, grocery store, online shopping, toilet paper , food and stock which signals how much of an important cause of concern it was for the people to even get basic day to day items during the pandemic

Most Used Words(Positive)



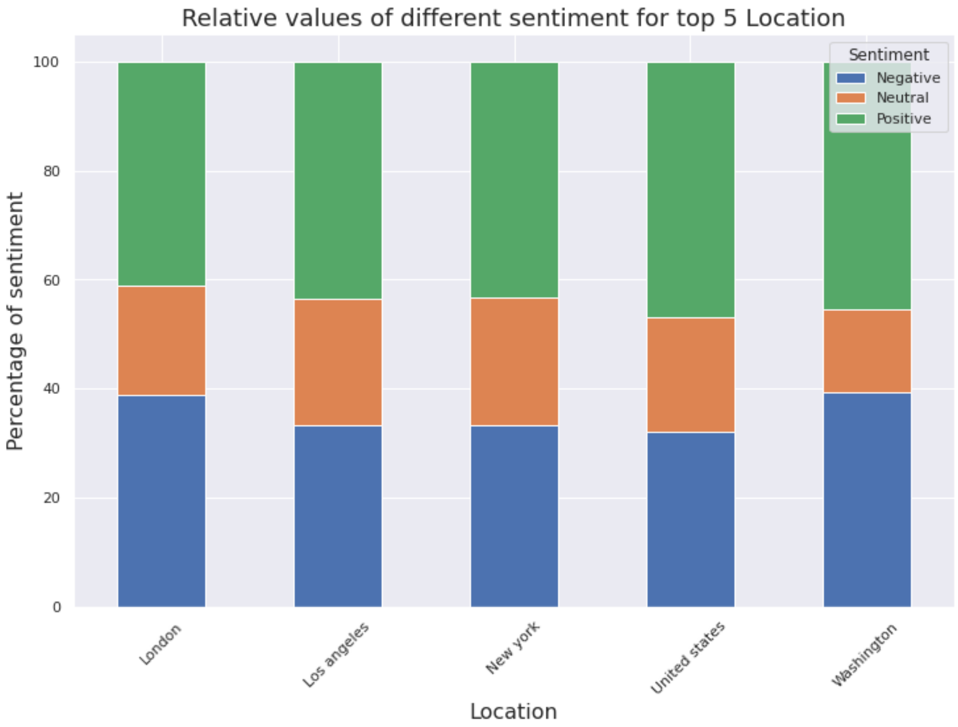
These are the top 15 most used words in positive tweets . We can notice that store, grocery , supermarket , food and shopping all of these are making it into the list and this indicates how much of a challenge it was for the people to buy groceries, food and other daily need items during the pandemic

Most Used Words(Negative)



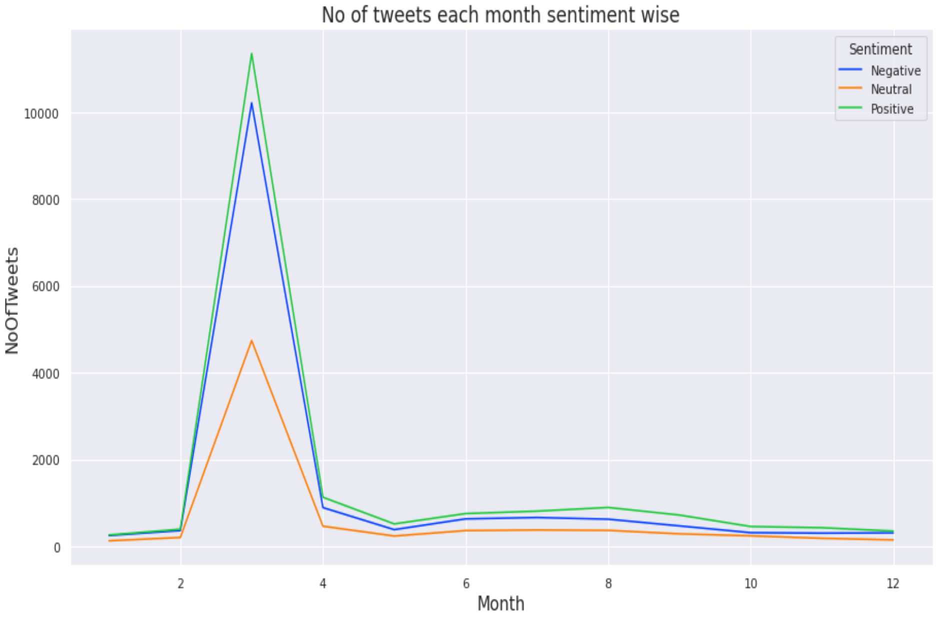
These are the top 15 most used words in negative tweets . Like the graph for positive tweets in this graph as well we can see words like food, grocery , supermarket etc . Panic is one important word which is present here but was missing from positive tweets graph

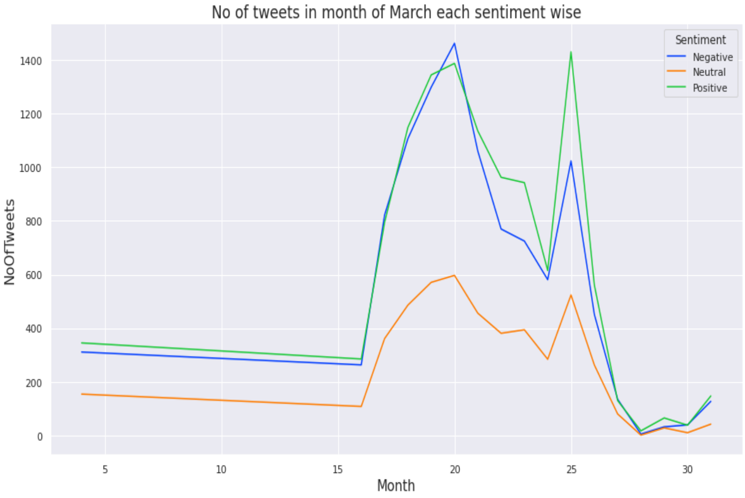
Relation of tweet sentiment and location



Looking at the bar graph we can say that among the top 5 location, tweets from United states are relatively more positive and less negative as compared to all other locations . In London the sentiment of people is almost equally divided with positive tweets being as much in number as negatives tweets

Relation of time of tweet with sentiment



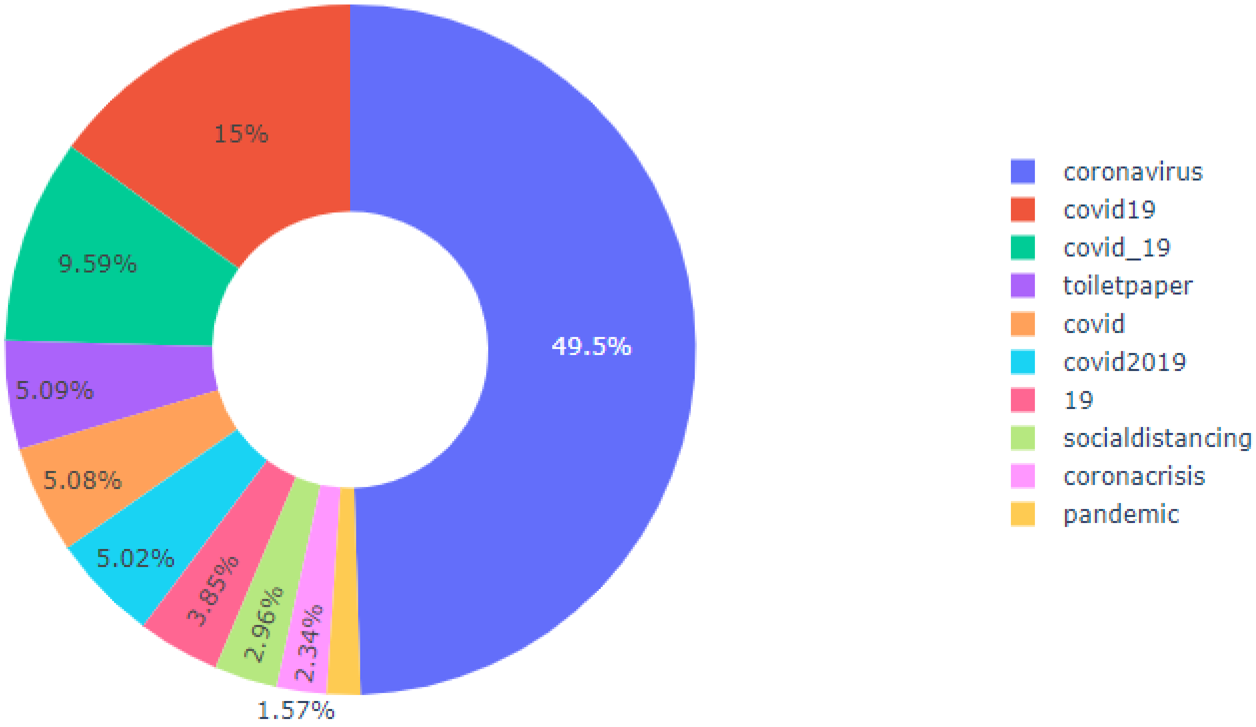


A very high majority of tweets in our record are from the month of march only whereas for all other months the number of tweets are more or less constant

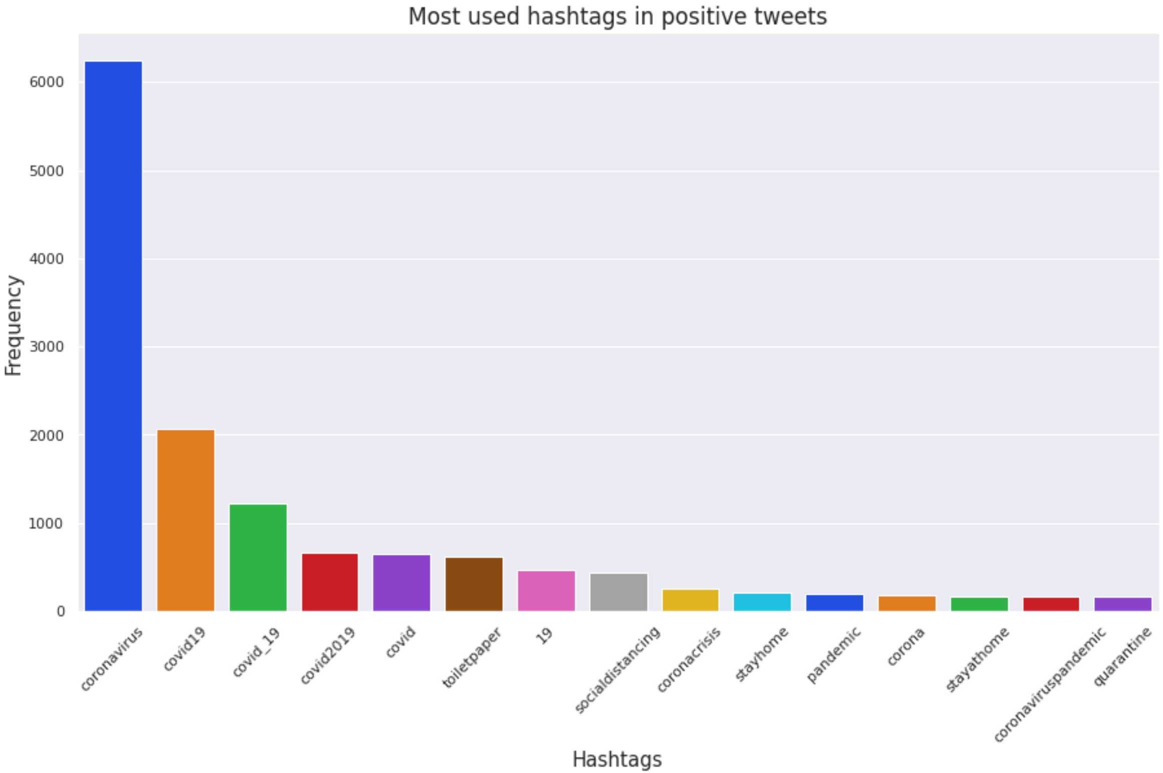
In the month of march from around the start of second week we can see a rise in number of both positive and negative tweets, this continues till around 26th of march

Tweets Hashtag Analysis

Relative Percentage of top 10 Hashtag

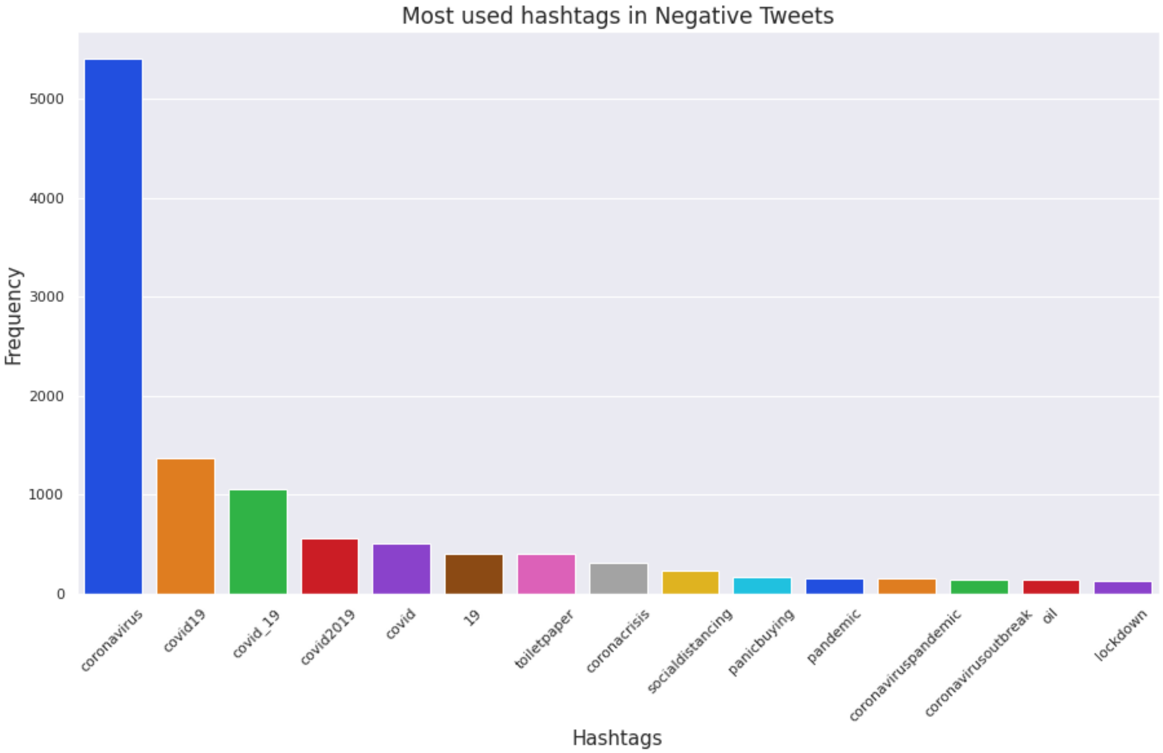


Most Used Hashtag(Positive)



Among the different version of hashtag on coronavirus name itself , we also have hashtags like socialdistancing , quarantine, stayathome all of which are healthy practices one can follow to avoid coronavirus , so these hashtags makes sense to be the most used ones in positive tweets

Most Used Hashtag(Negative)



Again like the previous bar graph even in this one the most used hashtags are just different version of the coronavirus name itself but also have hashtags like panicbuying and lockdown , the kind of hashtags which are more related to the negative outcomes of covid

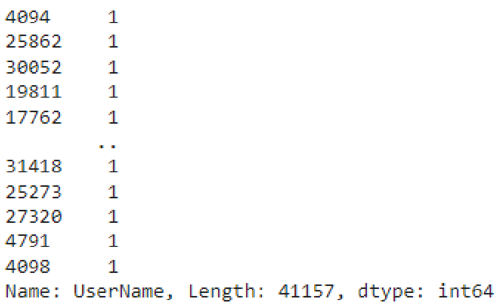
Text Pre Processing

To prepare the text data for the model building we performed text preprocessing. For text pre processing we performed a number of steps :

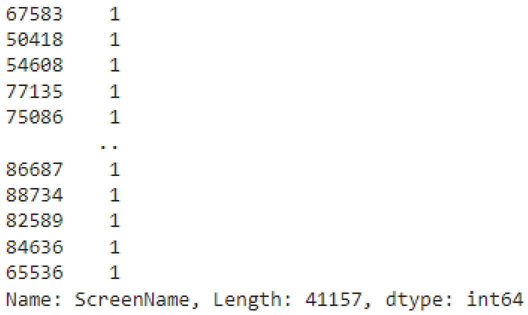
- Removed hashtags, URLs and mentions as all these characters gave no useful information about the sentiment of the tweet and just adds noise into algorithms if left untouched
- Expanded Contractions Converting each contraction to its expanded, original form helps with text standardization.
- Removed Punctuation and stop words as these are often added to sentence to make them grammatically correct , they do not add any values to our analysis as they carry less or no meaning as far as churning our information about the sentiment is concerned
- At last we performed lemmatization which is a popular text pre processing technique. It will help us to group different inflected form of words into the root form called lemma which will carry the same meaning. This helped us to diminish the number of token required to transfer the information

Feature Engineering and Selection

- To turn Location column into a form we can feed to our ml models , we created a new column LocationCat from it , depending on the percentage of positive, negative and neutral tweets from an area we assigned different values for LocationCat ranging from 2 to -2 where 2 means the percentage of positive tweets from that location is very high and -2 means the percentage of negative tweets is very , and so we dropped Location column from list of training variables
- We dropped UserName and ScreenName columns as they all contained unique values for each of the row because in our dataset we had records of only one tweet from a specific UserName .



```
4094     1
25862    1
30052    1
19811    1
17762    1
        ..
31418    1
25273    1
27320    1
4791     1
4098     1
Name: UserName, Length: 41157, dtype: int64
```

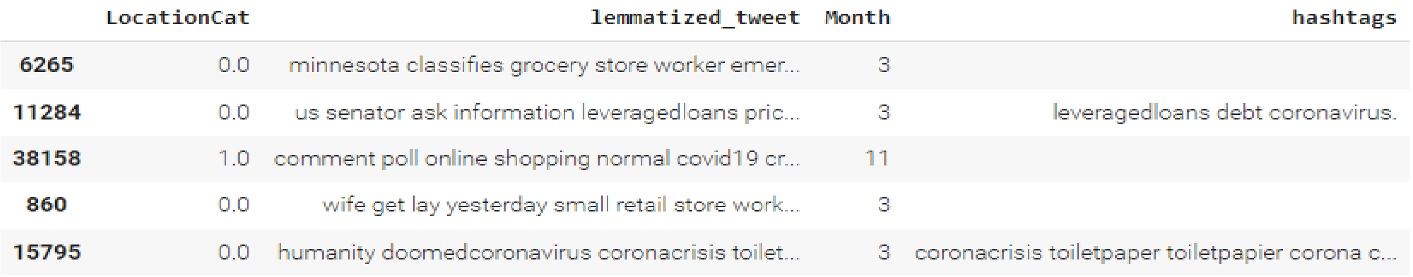


```
67583    1
50418    1
54608    1
77135    1
75086    1
        ..
86687    1
88734    1
82589    1
84636    1
65536    1
Name: ScreenName, Length: 41157, dtype: int64
```

Preparing Dataset for modelling

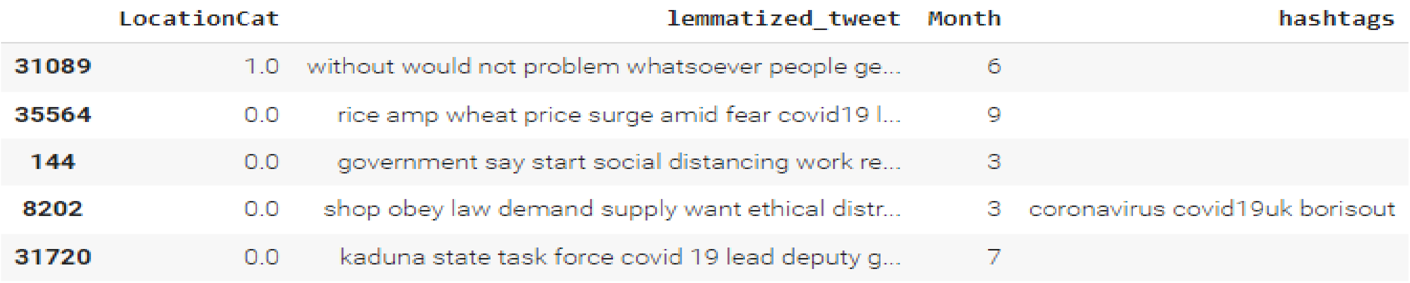
Task : Classification

Train Set:- (32925,4)



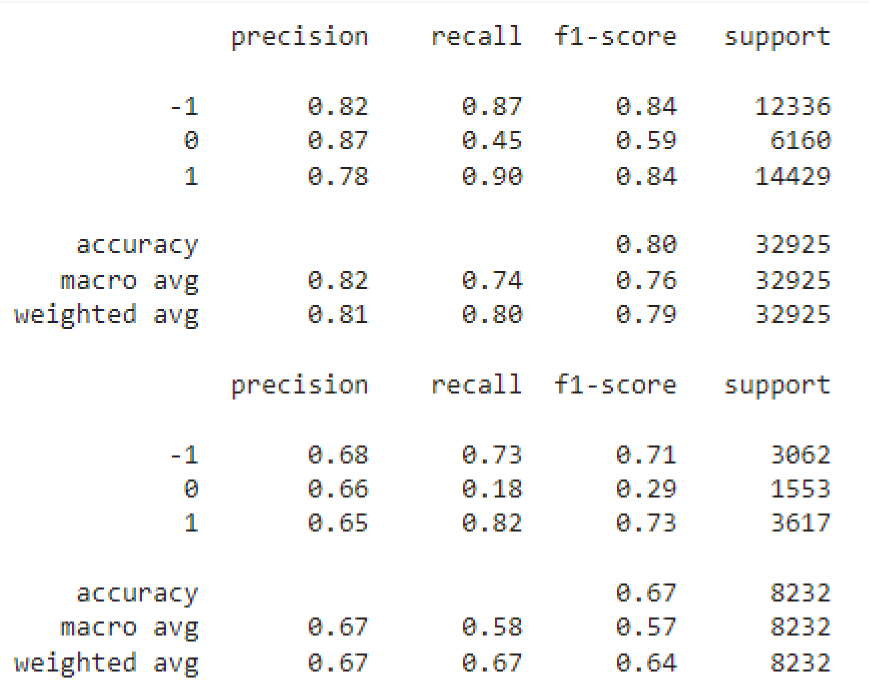
	LocationCat	lemmatized_tweet	Month	hashtags
6265	0.0	minnesota classifies grocery store worker emer...	3	
11284	0.0	us senator ask information leveragedloans pric...	3	leveragedloans debt coronavirus.
38158	1.0	comment poll online shopping normal covid19 cr...	11	
860	0.0	wife get lay yesterday small retail store work...	3	
15795	0.0	humanity doomedcoronavirus coronacrisis toilet...	3	coronacrisis toiletpaper toiletpapier corona c...

Test Set:- (8232,4)

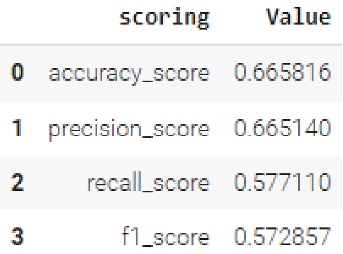


	LocationCat	lemmatized_tweet	Month	hashtags
31089	1.0	without would not problem whatsoever people ge...	6	
35564	0.0	rice amp wheat price surge amid fear covid19 l...	9	
144	0.0	government say start social distancing work re...	3	
8202	0.0	shop obey law demand supply want ethical distr...	3	coronavirus covid19uk borisout
31720	0.0	kaduna state task force covid 19 lead deputy g...	7	

Applying Model (Baseline Model)



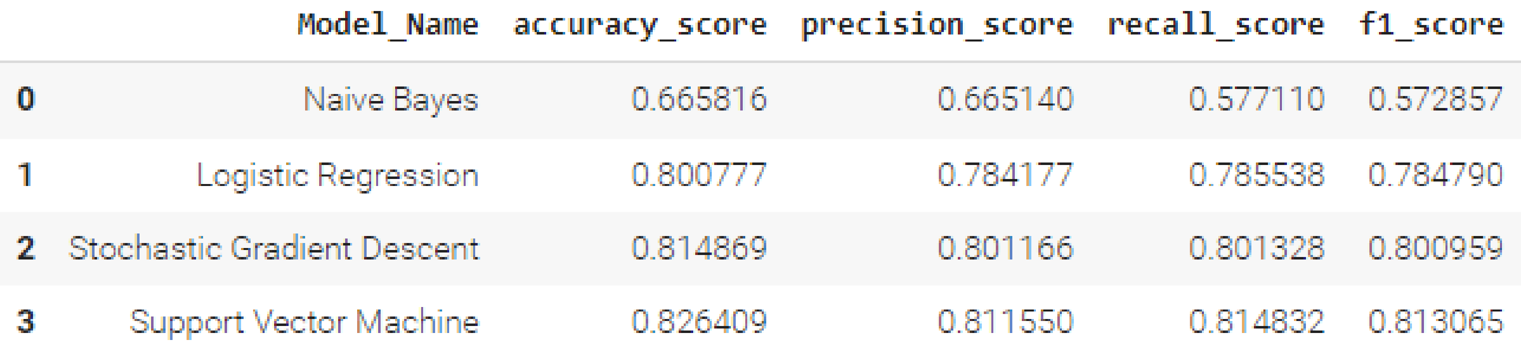
```
              precision    recall  f1-score   support

          -1       0.82      0.87      0.84     12336
           0       0.87      0.45      0.59      6160
           1       0.78      0.90      0.84     14429

    accuracy                           0.80     32925
   macro avg       0.82      0.74      0.76     32925
weighted avg       0.81      0.80      0.79     32925

              precision    recall  f1-score   support

          -1       0.68      0.73      0.71      3062
           0       0.66      0.18      0.29      1553
           1       0.65      0.82      0.73      3617

    accuracy                           0.67      8232
   macro avg       0.67      0.58      0.57      8232
weighted avg       0.67      0.67      0.64      8232
```



	scoring	Value
0	accuracy_score	0.665816
1	precision_score	0.665140
2	recall_score	0.577110
3	f1_score	0.572857

With an accuracy score of 0.66 ,our Baseline Model is Multinomial naive bayes . The recall value for neutral labels were pretty low of this model

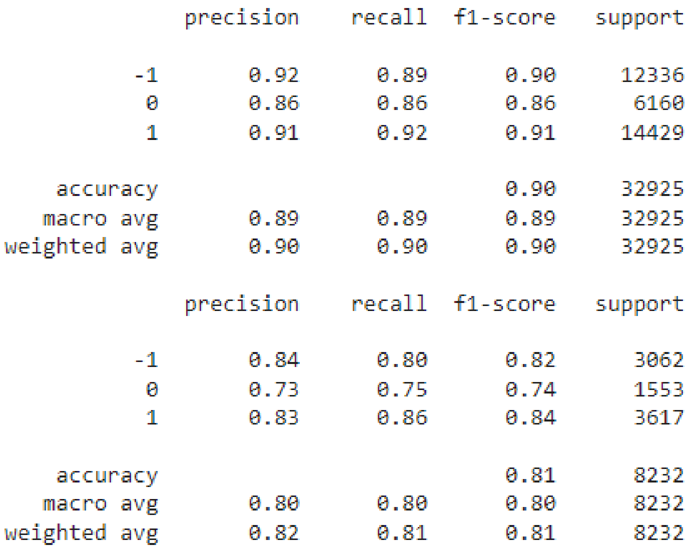
Model Validation and Selection (Multinomial)



	Model_Name	accuracy_score	precision_score	recall_score	f1_score
0	Naive Bayes	0.665816	0.665140	0.577110	0.572857
1	Logistic Regression	0.800777	0.784177	0.785538	0.784790
2	Stochastic Gradient Descent	0.814869	0.801166	0.801328	0.800959
3	Support Vector Machine	0.826409	0.811550	0.814832	0.813065

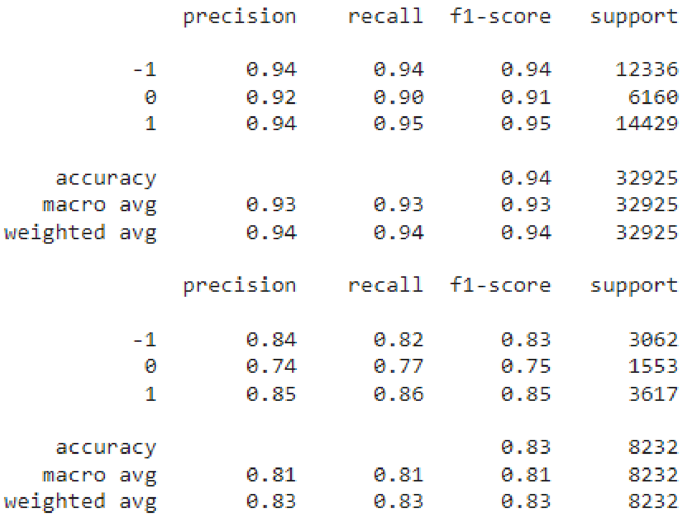
1. The scores of naïve bayes are very ordinary as compared to all the other three models
2. Support Vector Machine (LinearSVC) is our best performing model with highest scores across all kind of evaluation metrics
3. So based on the above observation we have chosen Support Vector machine as our model for multinomial classification

Model validation & Selection (Multinomial)



```
              precision    recall  f1-score   support

          -1       0.92      0.89      0.90     12336
           0       0.86      0.86      0.86      6160
           1       0.91      0.92      0.91     14429

    accuracy                           0.90     32925
   macro avg       0.89      0.89      0.89     32925
weighted avg       0.90      0.90      0.90     32925
              precision    recall  f1-score   support

          -1       0.84      0.80      0.82      3062
           0       0.73      0.75      0.74      1553
           1       0.83      0.86      0.84      3617

    accuracy                           0.81      8232
   macro avg       0.80      0.80      0.80      8232
weighted avg       0.82      0.81      0.81      8232
```

Stochastic Gradient Descent



```
              precision    recall  f1-score   support

          -1       0.94      0.94      0.94     12336
           0       0.92      0.90      0.91      6160
           1       0.94      0.95      0.95     14429

    accuracy                           0.94     32925
   macro avg       0.93      0.93      0.93     32925
weighted avg       0.94      0.94      0.94     32925
              precision    recall  f1-score   support

          -1       0.84      0.82      0.83      3062
           0       0.74      0.77      0.75      1553
           1       0.85      0.86      0.85      3617

    accuracy                           0.83      8232
   macro avg       0.81      0.81      0.81      8232
weighted avg       0.83      0.83      0.83      8232
```

Linear Support Vector Classifier

Model validation and Selection (Multinomial)

We have chosen Linear Support Vector classifier for our predictions and the best hyperparameters obtained are as below.

dual = False

penalty = ‘l1’

C = 0.4

loss = “squared_hinge”

tol = 1e-4

multi_class = ‘ovr’

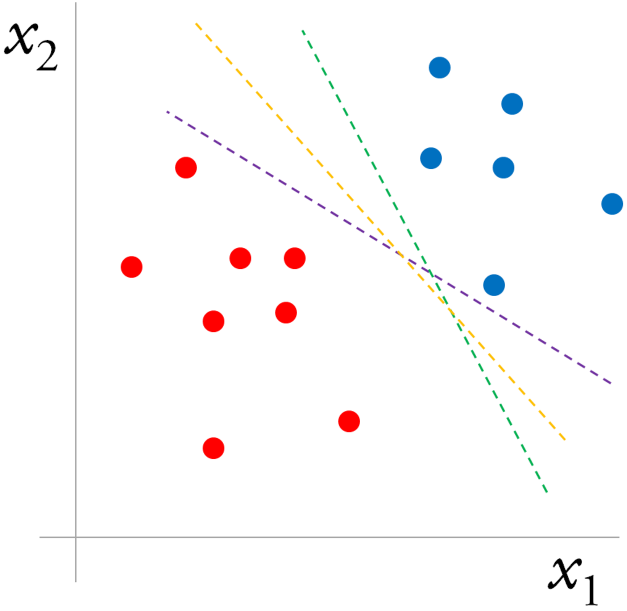
fit_intercept = True

intercept_scaling = 1

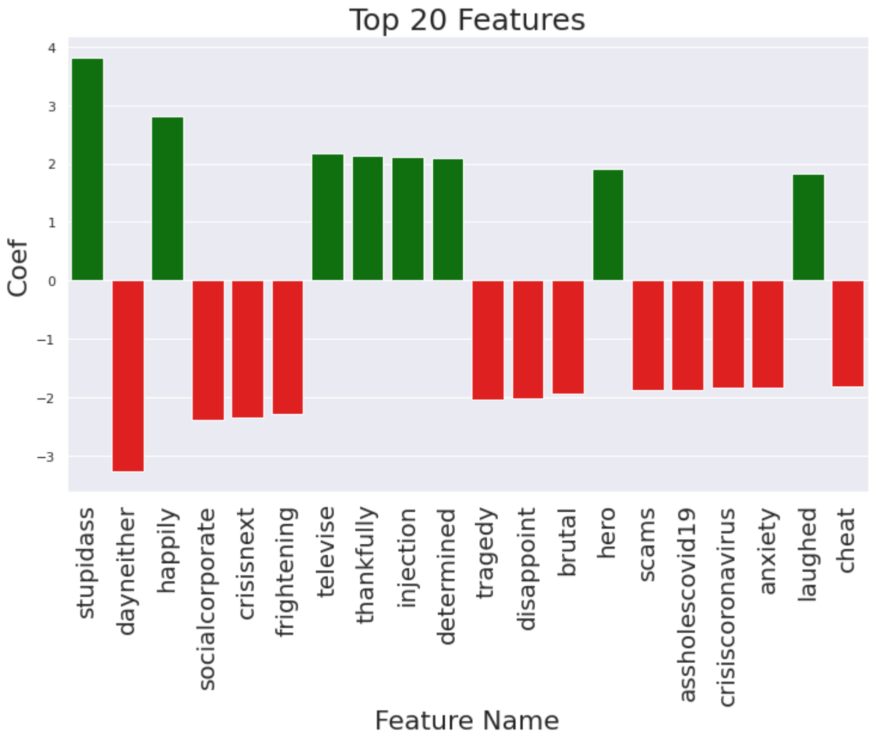
class_weight = None

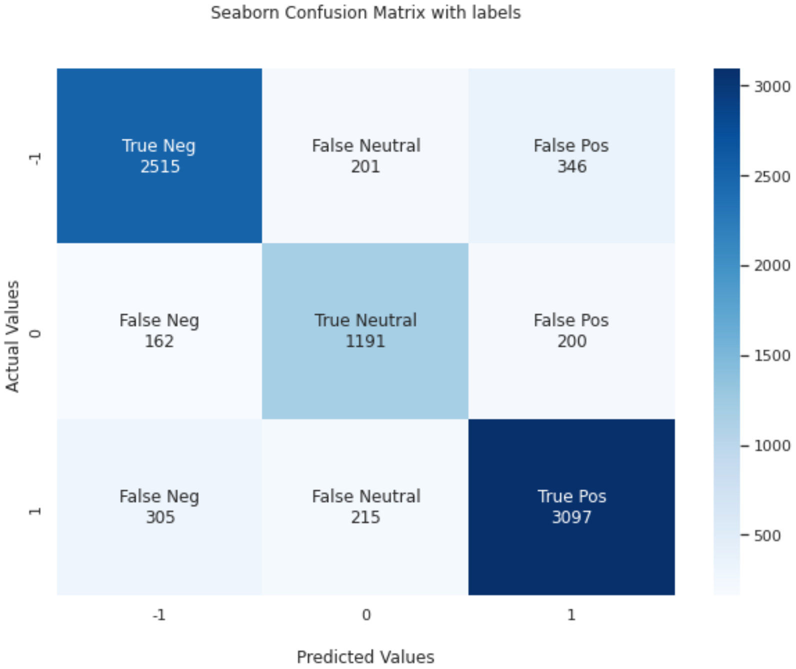
random_state = None,

max_iter = 1000



Model validation & Selection (Multinomial)

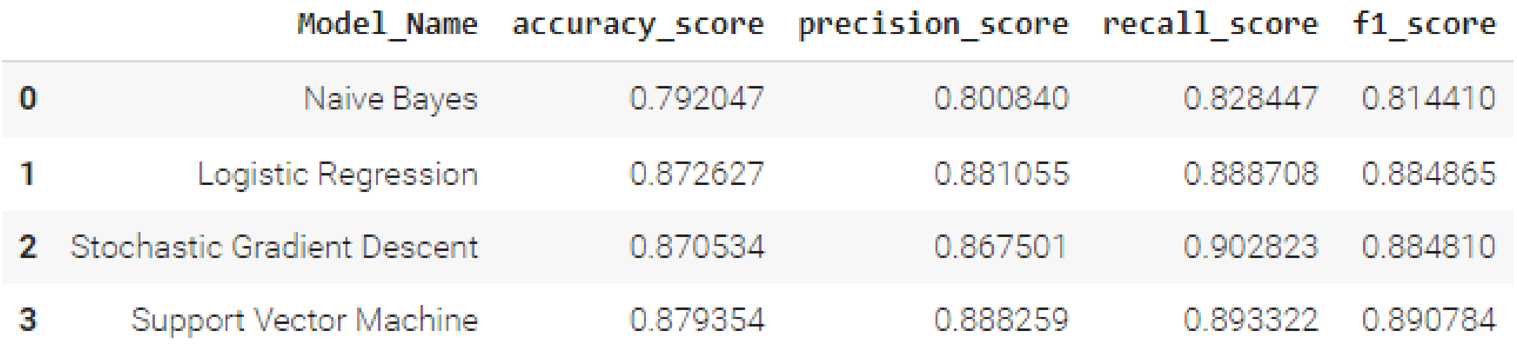




	-1	0	1
-1	True Neg 2515	False Neutral 201	False Pos 346
0	False Neg 162	True Neutral 1191	False Pos 200
1	False Neg 305	False Neutral 215	True Pos 3097

Linear Support Vector Classifier

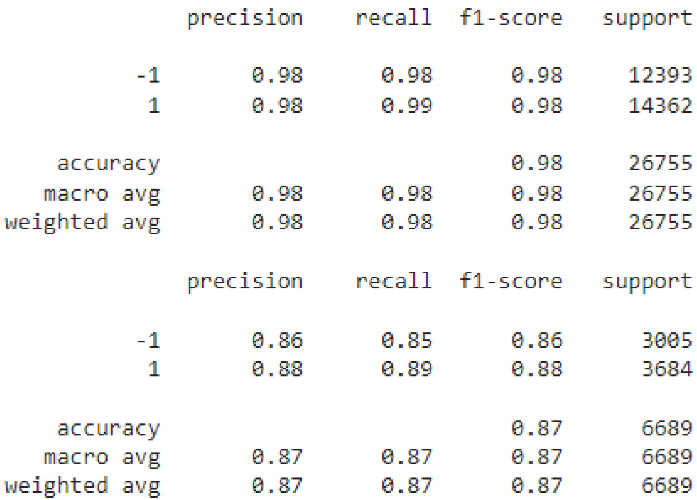
Model Validation and Selection (Binary)



	Model_Name	accuracy_score	precision_score	recall_score	f1_score
0	Naive Bayes	0.792047	0.800840	0.828447	0.814410
1	Logistic Regression	0.872627	0.881055	0.888708	0.884865
2	Stochastic Gradient Descent	0.870534	0.867501	0.902823	0.884810
3	Support Vector Machine	0.879354	0.888259	0.893322	0.890784

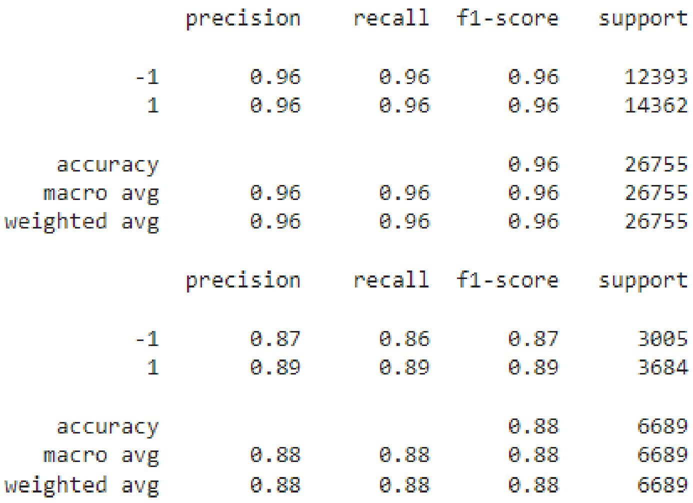
1. The scores of Bernoulli naive bayes are the lowest across all the metrics
2. Support Vector Machine (LinearSVC) is our best performing model with good overall scores , recall score of Stochastic gradient descent model is highest among all
3. So based on the above observation we have chosen Linear support vector classifier as our model for Binary classification

Model validation & Selection (binary)



```
              precision    recall  f1-score   support

          -1       0.98      0.98      0.98     12393
           1       0.98      0.99      0.98     14362

    accuracy                           0.98     26755
   macro avg       0.98      0.98      0.98     26755
weighted avg       0.98      0.98      0.98     26755
              precision    recall  f1-score   support

          -1       0.86      0.85      0.86      3005
           1       0.88      0.89      0.88      3684

    accuracy                           0.87      6689
   macro avg       0.87      0.87      0.87      6689
weighted avg       0.87      0.87      0.87      6689
```

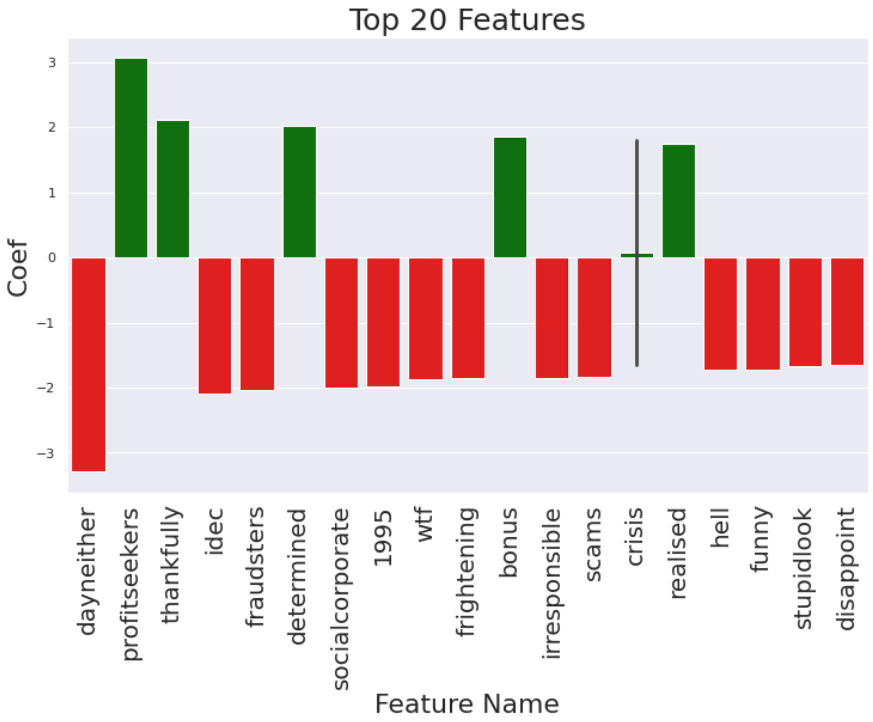
Logistic Regression

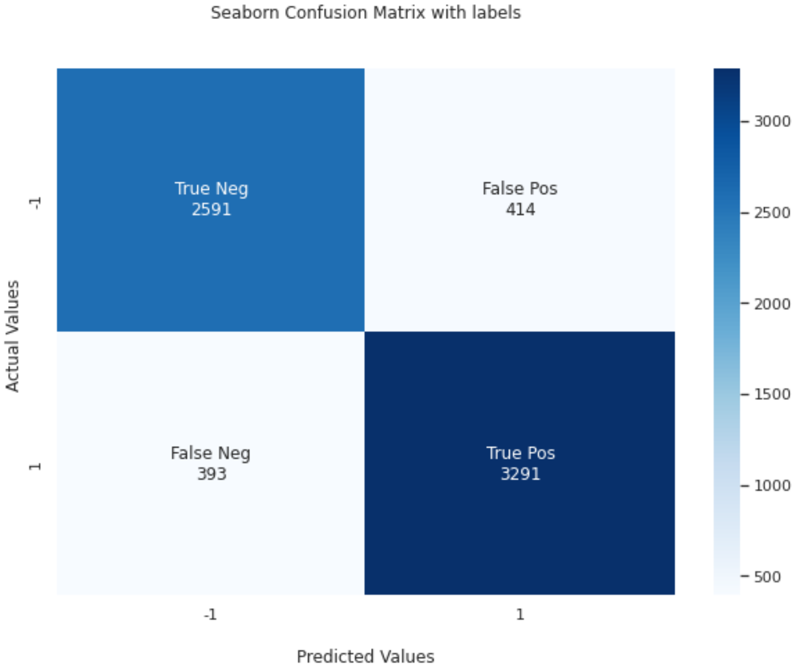


```
              precision    recall  f1-score   support

          -1       0.96      0.96      0.96     12393
           1       0.96      0.96      0.96     14362

    accuracy                           0.96     26755
   macro avg       0.96      0.96      0.96     26755
weighted avg       0.96      0.96      0.96     26755
              precision    recall  f1-score   support

          -1       0.87      0.86      0.87      3005
           1       0.89      0.89      0.89      3684

    accuracy                           0.88      6689
   macro avg       0.88      0.88      0.88      6689
weighted avg       0.88      0.88      0.88      6689
```

Linear Support Vector Classifier

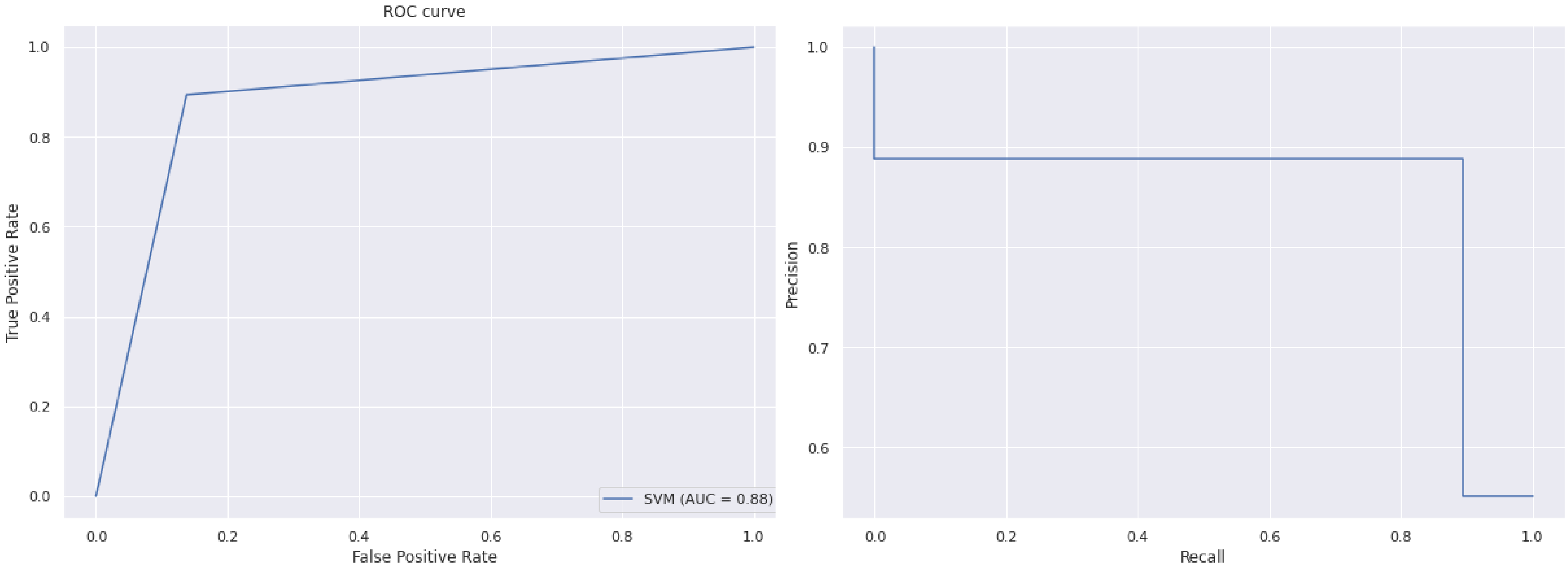
Model validation & Selection (Binary)





Linear Support Vector Classifier

Model validation & Selection (Binary)



Linear Support Vector Classifier

Conclusion

- Majority of the tweets in our database had a positive sentiment
- Average number of words in neutral sentiment tweets were far less as compared to positive sentiment tweets
- Grocery store , food, supermarket and online shopping these were some of the most used words in all of the tweets apart from the name of the disease itself
- Most of the tweets in our dataset be it positive, negative or neutral were from the month of march
- #quarantine , #stayathome these were some of the unique hashtag that were mostly found in positive sentiment tweets whereas for negative tweets it was #panicbuying and #lockdown
- For both multinomial classification and binary classification (just positive and negative) linear support vector classifier was our best performing model with an accuracy of 0.825 for multinomial classification and 0.88 for binary classification

Challenges

- As this was a text classification project , so our biggest challenge was to clean the tweets and remove all those words and character which do not add any value to our analysis and just induce noise into the algorithms
- Finding the right set of hyperparameters for both Vectorization estimator and ml model was tricky
- Computation time of some of the algorithms were high