Capstone Project - 3

Project Title – Coronavirus Tweet Sentiment Analysis

Team Members

Nivya T

Shaurabh Pandey

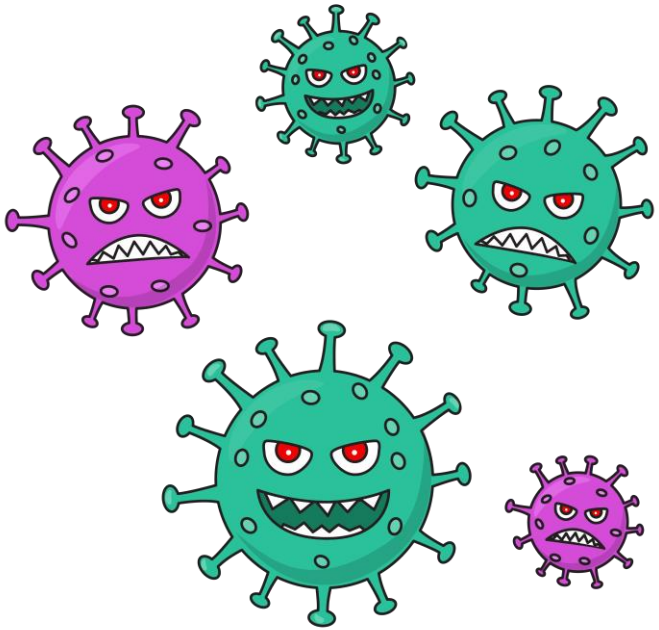
Manjusree K C

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.

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.

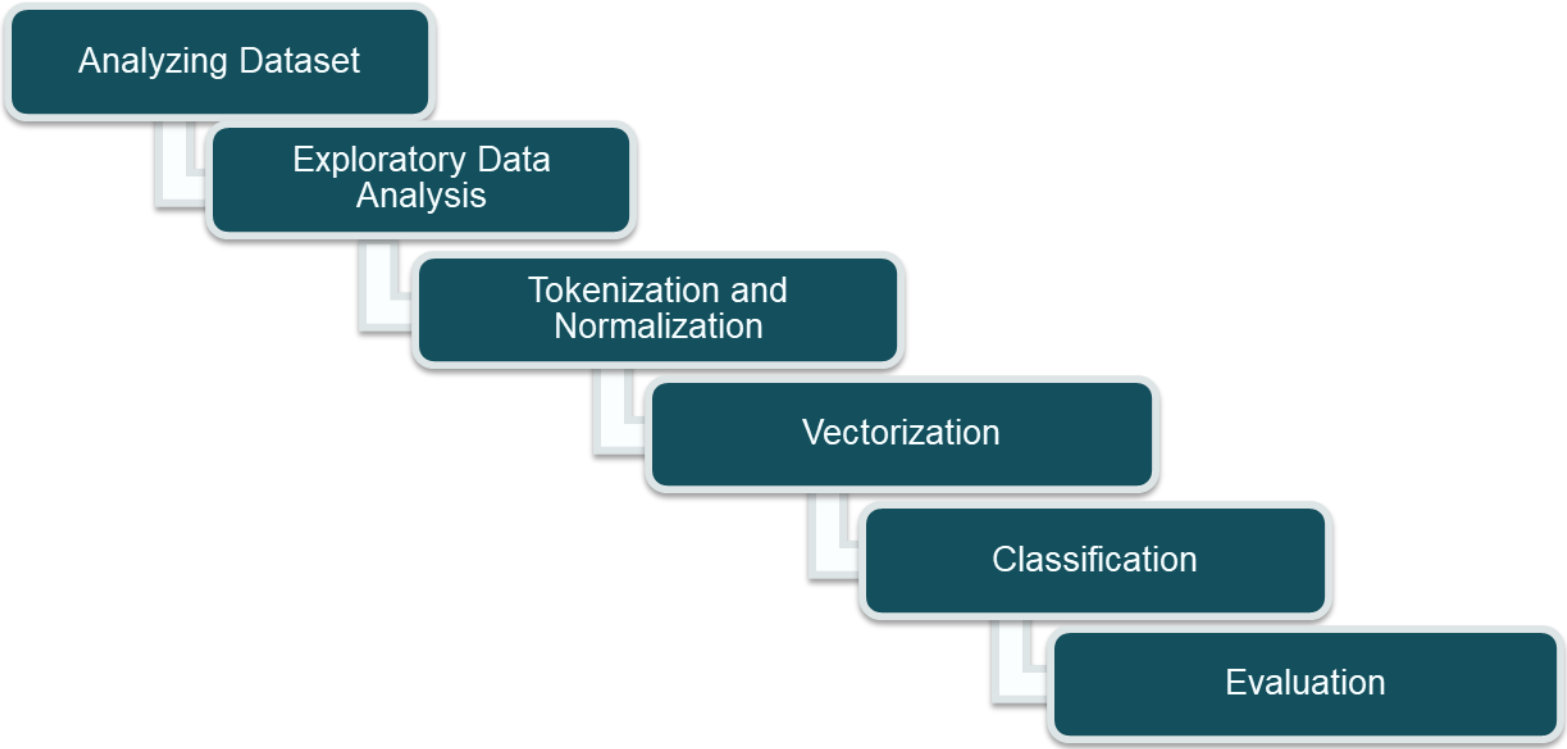
For this project we are provided with a coronavirus tweets csv file which contains more than 40000 tweets from people around the world on covid 19 and our aim is to analyze these tweets made on Covid-19 from around the world and predict the sentiment of each of the tweet by classifying them into three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

STEPS INVOLVED



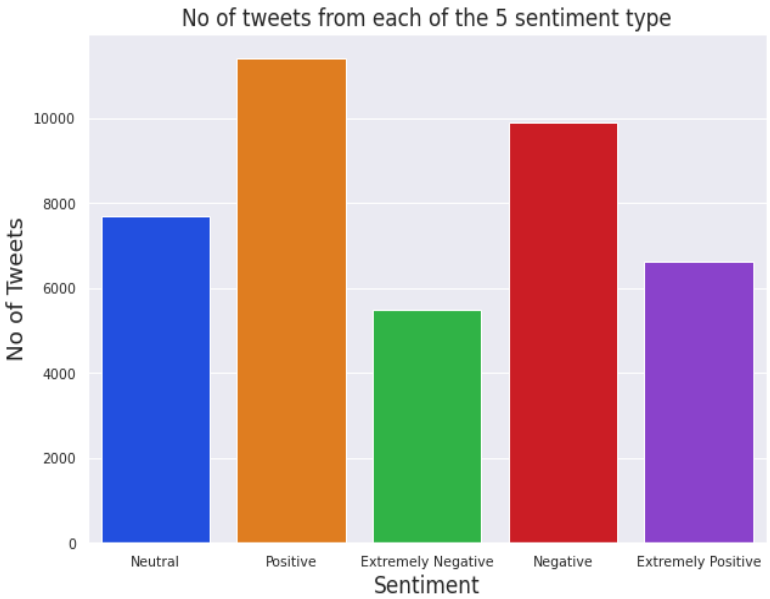
DEFINING THE VARIABLES

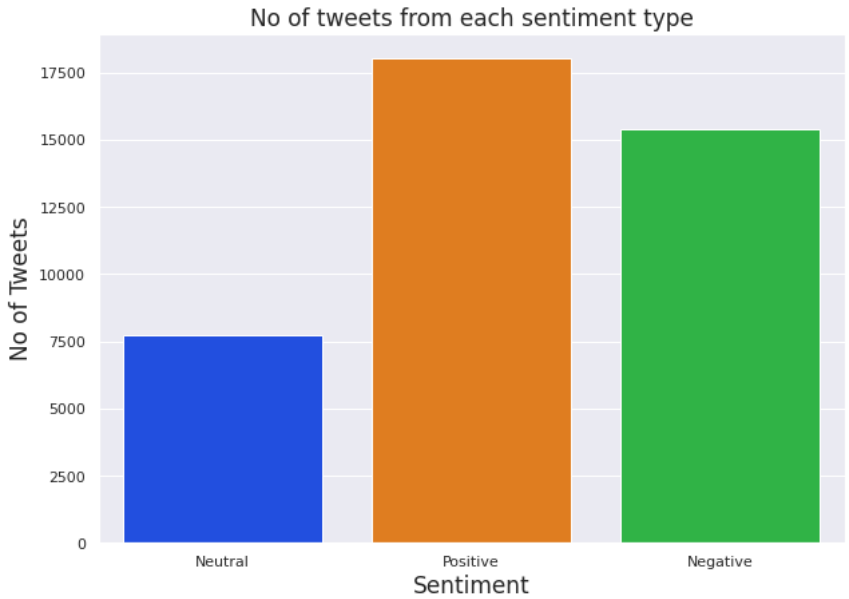
The shape of the dataset is (41157, 6).

- Username : The username of the person 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

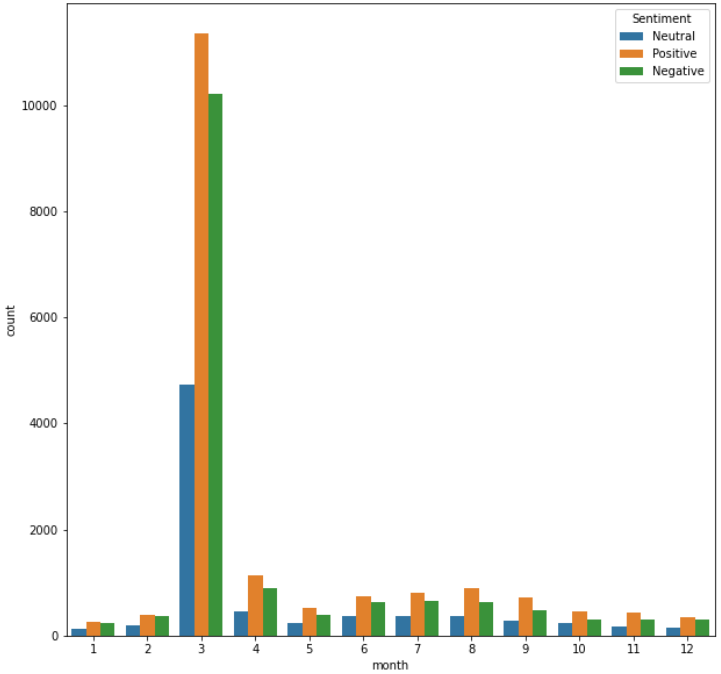
The target variable is **‘Sentiment’.**

TARGET VARIABLE STUDY

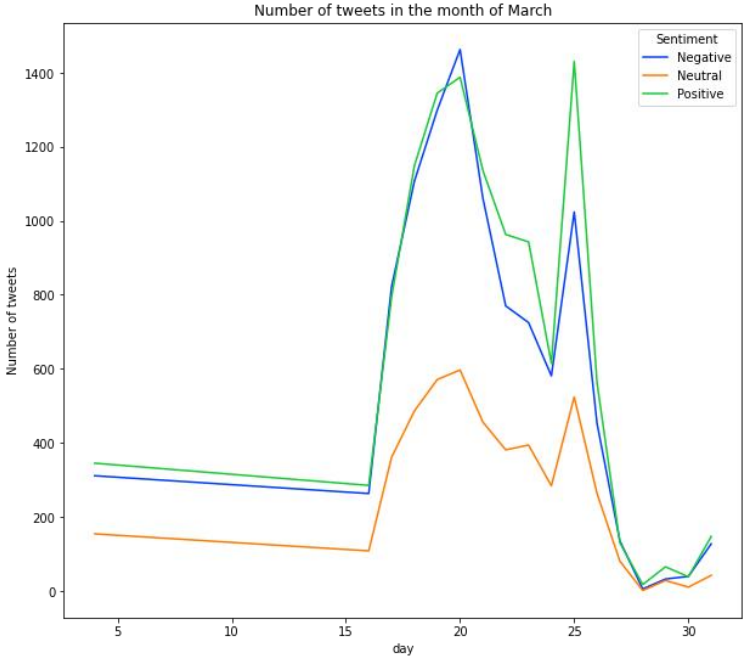




DATE ANALYSIS

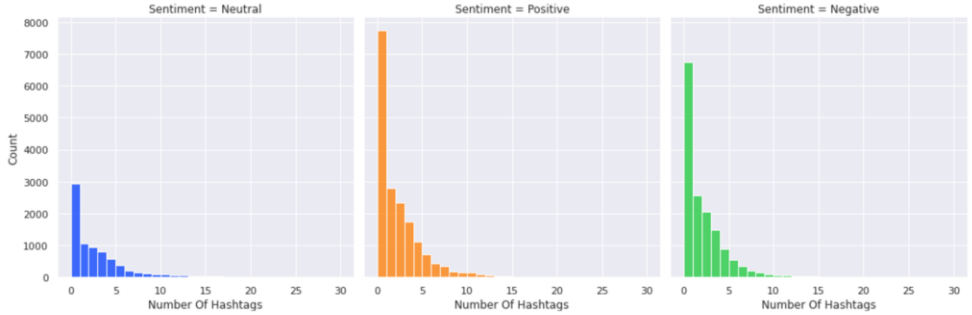


Majority of the tweets in our record are from the month of march whereas for all other months the number of tweets are more or less constant.

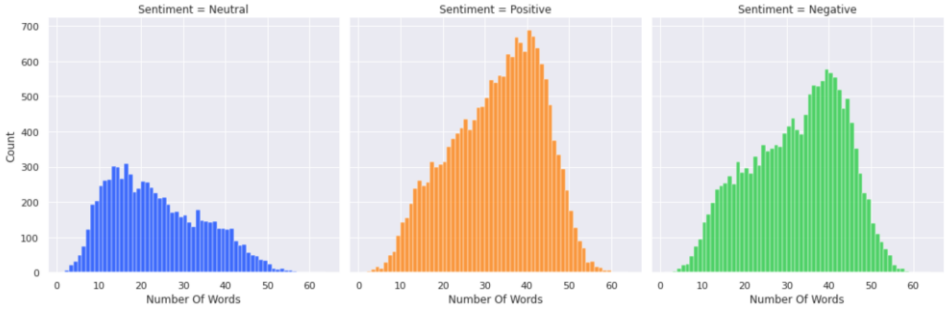


In the month of march from around the start of second week we can see a rise in number of both positive and negative tweets, and this continues till around 26th of march.

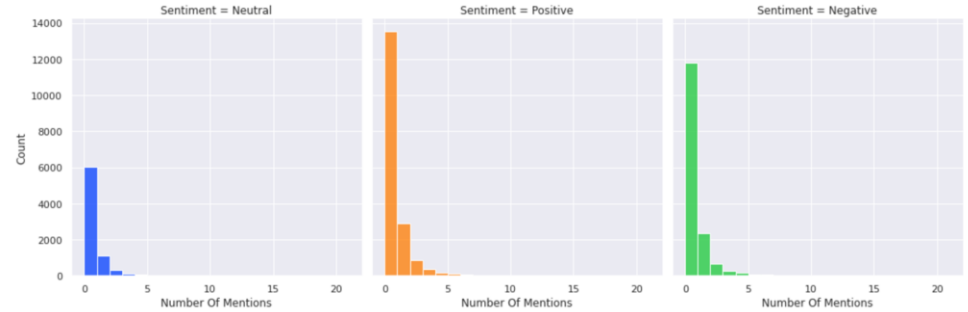
Tweets Analysis (Hashtags)



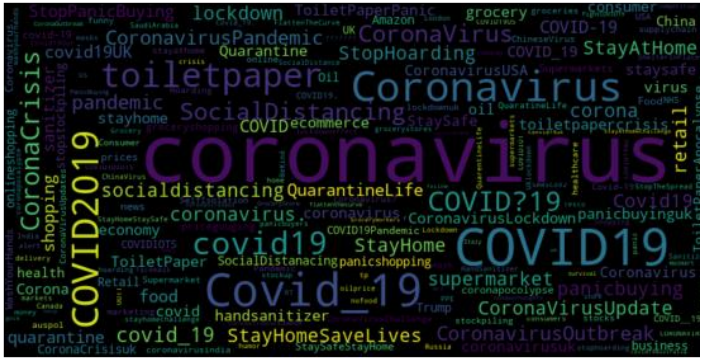
Tweets Analysis (Words)

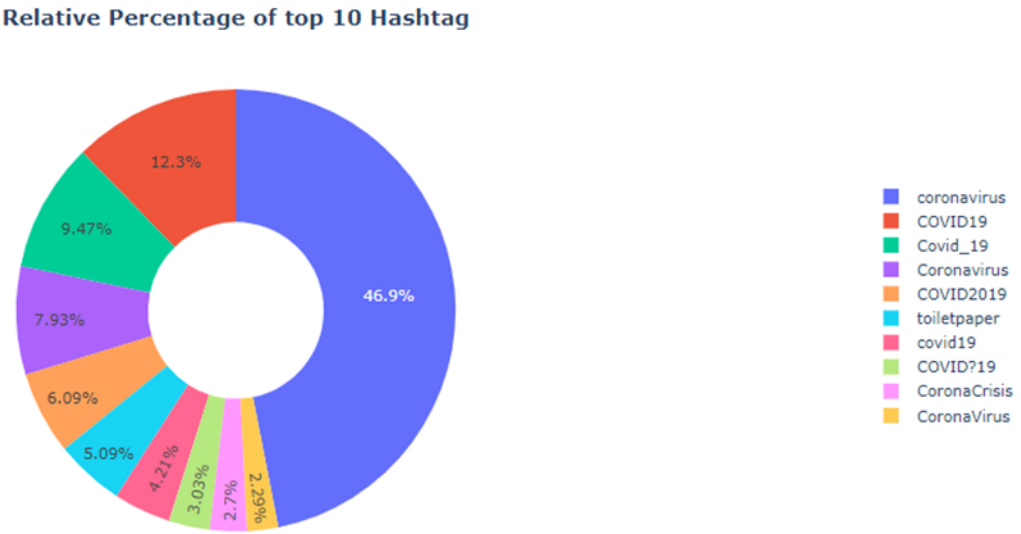


Tweets Analysis (Mentions)



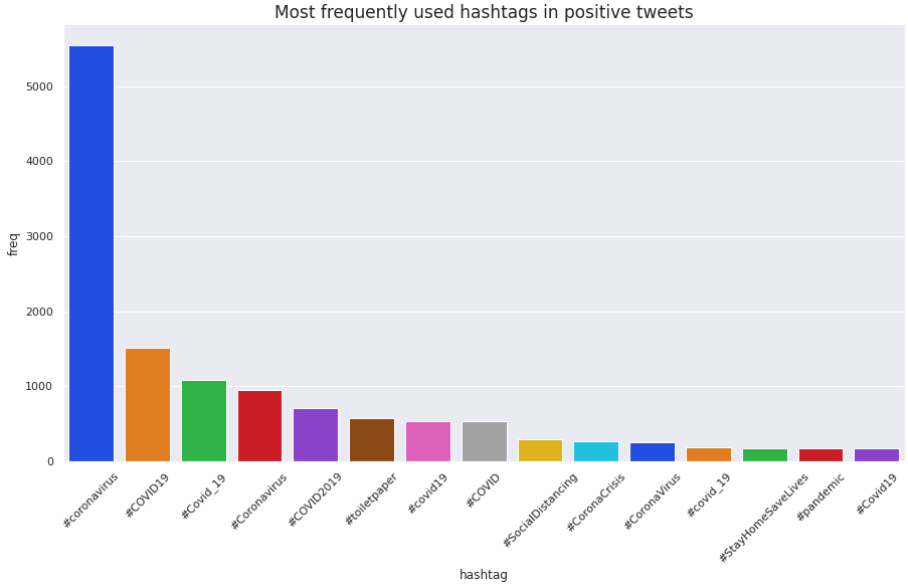
HASHTAG ANALYSIS

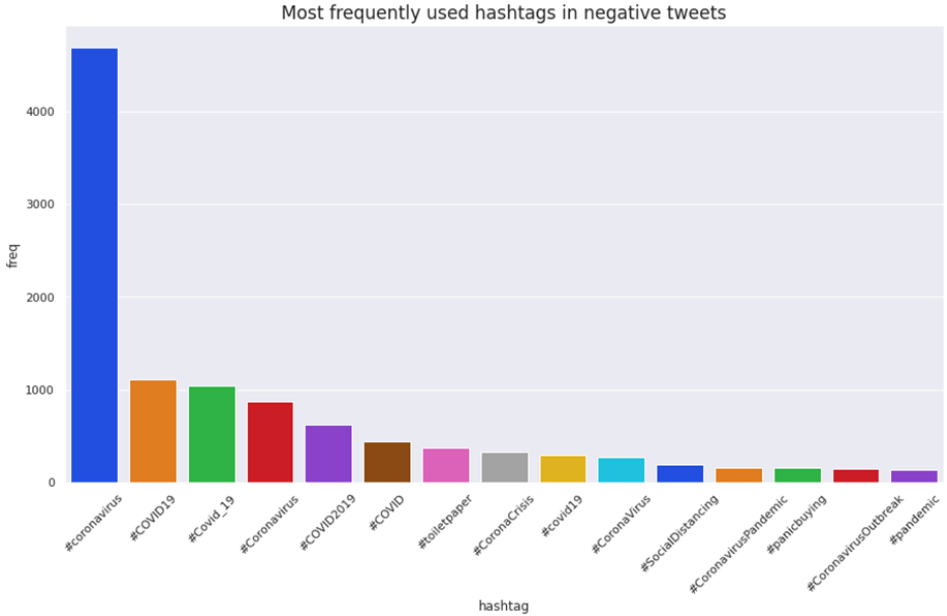




MOST FREQUENT HASHTAGS

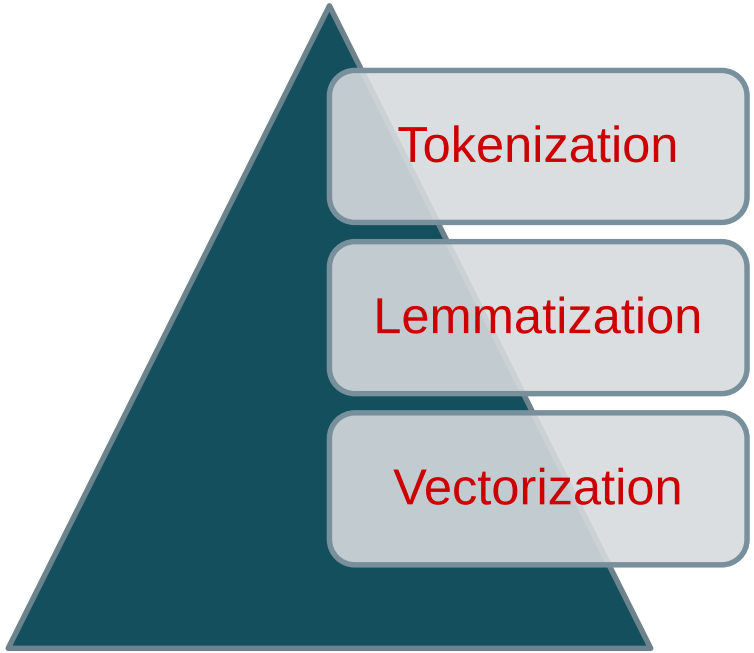
The most used hashtags are different versions of the name Corona virus itself.



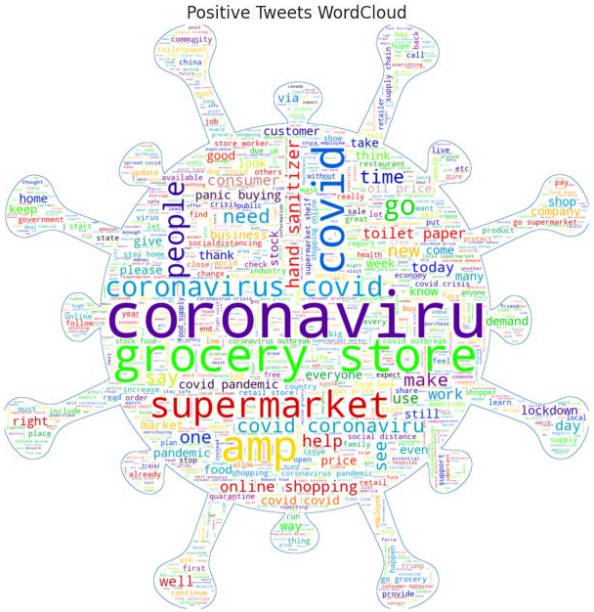


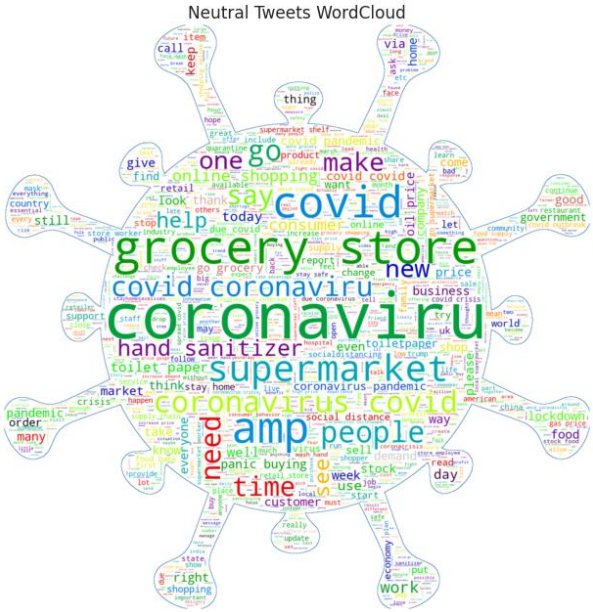
#StayHomesaveLives is much more used in tweets of positive sentiment than the negative sentiment tweets. On the other hand, hashtags like #panicbuying are more frequently used in negative sentiment tweets.

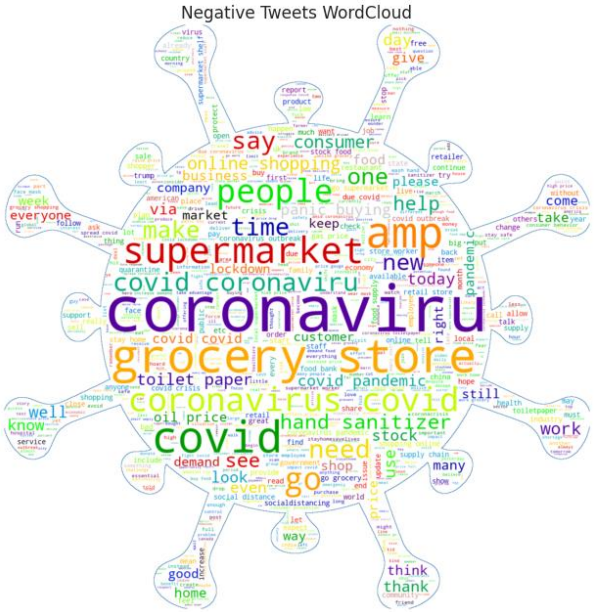
TEXT PREPROCESSING



TWEET WORDCLOUD

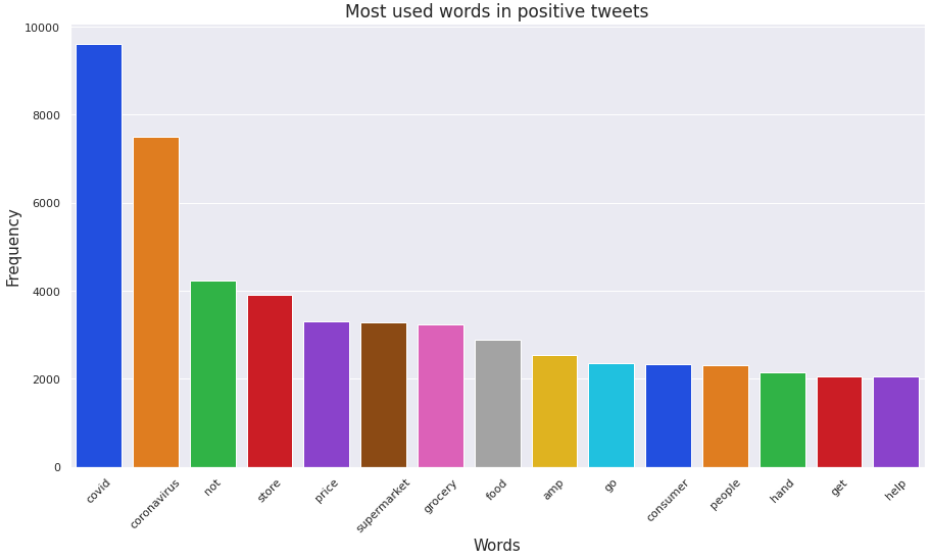




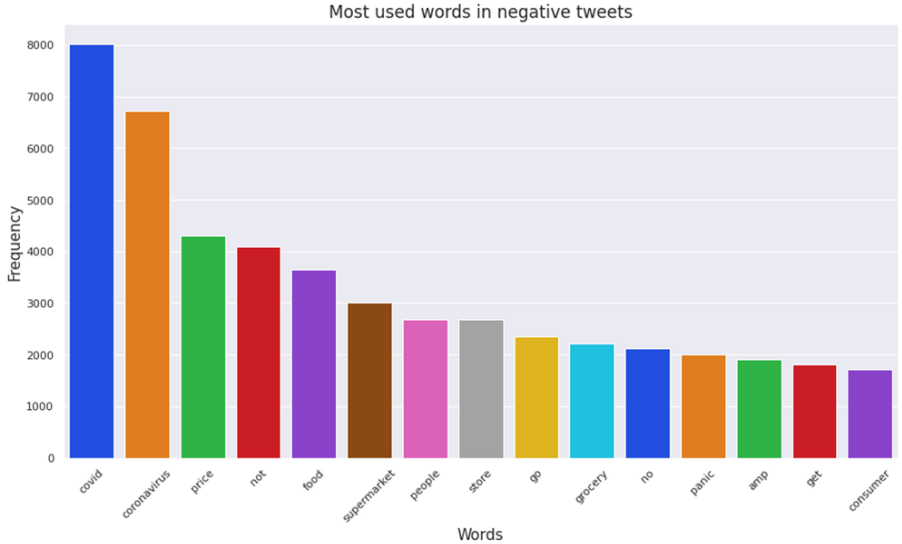


In all the tweets, irrespective of the sentiment the most frequently used words apart from the name of the disease are supermarket, grocery store, toilet paper, online shopping, food and price which signals how much of a significant cause of concern it was for the people even to get basic day to day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

MOST FREQUENT WORDS IN TWEETS



Positive tweets include words like store, grocery, supermarket, food, price etc.

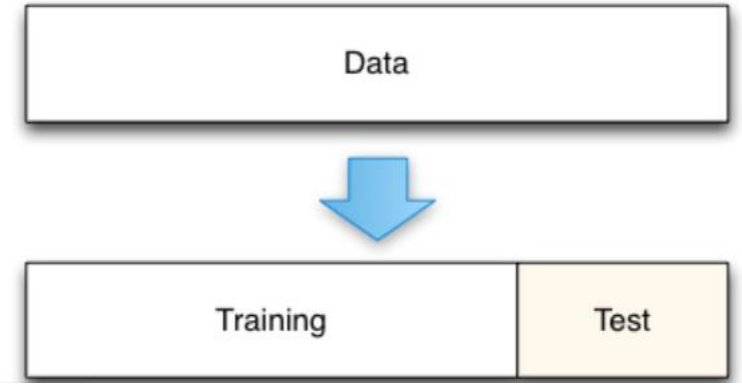


Words like food, groceries, supermarkets can also be found on the graph of negative tweets. Panic is a key word here, but it does not appear on the graph of positive tweets.

PREPARING DATASET FOR MODELLING

Task : Classification

Train Set:- (32925,5198)
Test Set:- (8232,5198)



APPLYING MODEL (BASELINE MODEL)

```
Training accuracy Score    :  0.7241609719058466
Test accuracy Score :  0.6722546161321672

 Naive Bayes-Classification Report for train dataset

              precision    recall  f1-score   support

          -1       0.76      0.77      0.77     12318
           0       0.83      0.27      0.41      6170
           1       0.69      0.88      0.77     14437

    accuracy                           0.72     32925
   macro avg       0.76      0.64      0.65     32925
weighted avg       0.74      0.72      0.70     32925

Naive Bayes-Classification Report for test dataset

              precision    recall  f1-score   support

          -1       0.70      0.72      0.71      3080
           0       0.71      0.19      0.30      1543
           1       0.65      0.84      0.73      3609

    accuracy                           0.67      8232
   macro avg       0.69      0.58      0.58      8232
weighted avg       0.68      0.67      0.64      8232
```

	scoring	Value
0	accuracy_score	0.672255
1	precision_score	0.686899
2	recall_score	0.582083
3	f1_score	0.579871

Our Baseline Model is Multinomial Naive Bayes which gives an accuracy score of 0.672, which is a low score. The recall value for neutral labels are pretty low for this model.

Model validation & Selection (Multinomial)

Training accuracy Score : 0.842672741078208
Test accuracy Score : 0.7978620019436345

SGD-Classification Report for train dataset

	precision	recall	f1-score	support
-1	0.90	0.81	0.85	13777
0	0.64	0.89	0.75	4415
1	0.88	0.86	0.87	14733
accuracy			0.84	32925
macro avg	0.81	0.85	0.82	32925
weighted avg	0.86	0.84	0.85	32925

SGD-Classification Report for test dataset

	precision	recall	f1-score	support
-1	0.88	0.77	0.82	3530
0	0.54	0.81	0.65	1017
1	0.84	0.82	0.83	3685
accuracy			0.80	8232
macro avg	0.75	0.80	0.77	8232
weighted avg	0.82	0.80	0.80	8232

Stochastic Gradient Descent

Training accuracy Score : 0.86958238420653
Test accuracy Score : 0.8118318756073858

LogisticRegression-Classification Report for train dataset

	precision	recall	f1-score	support
-1	0.88	0.87	0.88	12466
0	0.76	0.87	0.81	5420
1	0.91	0.87	0.89	15039
accuracy			0.87	32925
macro avg	0.85	0.87	0.86	32925
weighted avg	0.87	0.87	0.87	32925

LogisticRegression-Classification Report for test dataset

	precision	recall	f1-score	support
-1	0.83	0.83	0.83	3057
0	0.66	0.76	0.71	1352
1	0.86	0.82	0.84	3823
accuracy			0.81	8232
macro avg	0.78	0.80	0.79	8232
weighted avg	0.82	0.81	0.81	8232

Logistic Regression

Model validation & Selection (Multinomial)

Training accuracy Score : 0.8428549734244495
Test accuracy Score : 0.8126822157434402

CatBoost-Classification Report for train dataset

	precision	recall	f1-score	support
-1	0.84	0.86	0.85	11964
0	0.77	0.77	0.77	6160
1	0.88	0.85	0.86	14801
accuracy			0.84	32925
macro avg	0.83	0.83	0.83	32925
weighted avg	0.84	0.84	0.84	32925

CatBoost-Classification Report for test dataset

	precision	recall	f1-score	support
-1	0.81	0.83	0.82	2978
0	0.76	0.74	0.75	1583
1	0.84	0.83	0.84	3671
accuracy			0.81	8232
macro avg	0.80	0.80	0.80	8232
weighted avg	0.81	0.81	0.81	8232

CatBoost

Training accuracy Score : 0.8996203492786636
Test accuracy Score : 0.8182701652089407

LinearSVC-Classification Report for train dataset

	precision	recall	f1-score	support
-1	0.91	0.90	0.91	12521
0	0.82	0.89	0.85	5649
1	0.92	0.90	0.91	14755
accuracy			0.90	32925
macro avg	0.88	0.90	0.89	32925
weighted avg	0.90	0.90	0.90	32925

LinearSVC-Classification Report for test dataset

	precision	recall	f1-score	support
-1	0.83	0.84	0.83	3059
0	0.69	0.73	0.71	1443
1	0.86	0.84	0.85	3730
accuracy			0.82	8232
macro avg	0.79	0.80	0.80	8232
weighted avg	0.82	0.82	0.82	8232

Linear Support Vector Classifier

MODEL VALIDATION AND SELECTION (MULTINOMIAL)

	Model	Train accuracy	Test accuracy
0	Support Vector Machine	0.90	0.82
1	CatBoost	0.84	0.81
2	Logistic Regression	0.87	0.81
3	Stochastic Gradient Decent	0.84	0.80
4	Naive Bayes	0.72	0.67

	Model_Name	accuracy_score	precision_score	recall_score	f1_score
0	Support Vector Machine	0.82	0.80	0.79	0.80
1	CatBoost	0.81	0.80	0.80	0.80
2	Logistic Regression	0.81	0.80	0.78	0.79
3	Stochastic Gradient Descent	0.80	0.80	0.75	0.77
4	Naive Bayes	0.67	0.69	0.58	0.58

OBSERVATIONS

- The scores of Multinomial Naive Bayes are the lowest across all the metrics.
- Linear support vector machine has the highest test accuracy compared to other models.
- Recall score of CatBoost model is highest among all.
- Considering the overall performance and computation time we have chosen Linear support vector classifier as our model for multinomial classification.

LINEAR SVC AFTER CROSS VALIDATION AND HYPERPARAMETER TUNING

We have chosen Linear Support Vector classifier as the best model for our predictions

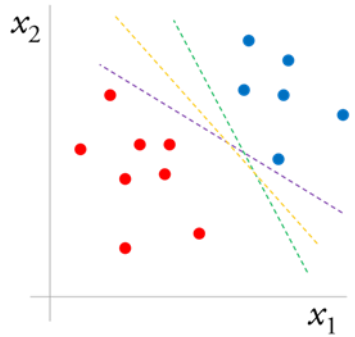
```
Fitting 5 folds for each of 4 candidates, totalling 20 fits
Best hyperparameters for the LinearSVC Grid Model are: {'C': 0.4, 'dual': False, 'penalty': 'l1'}
Training accuracy Score   :  0.8786636294608959
Validation accuracy Score :  0.8482750242954324
```

LinearSVC Grid Model-Classification Report for train dataset

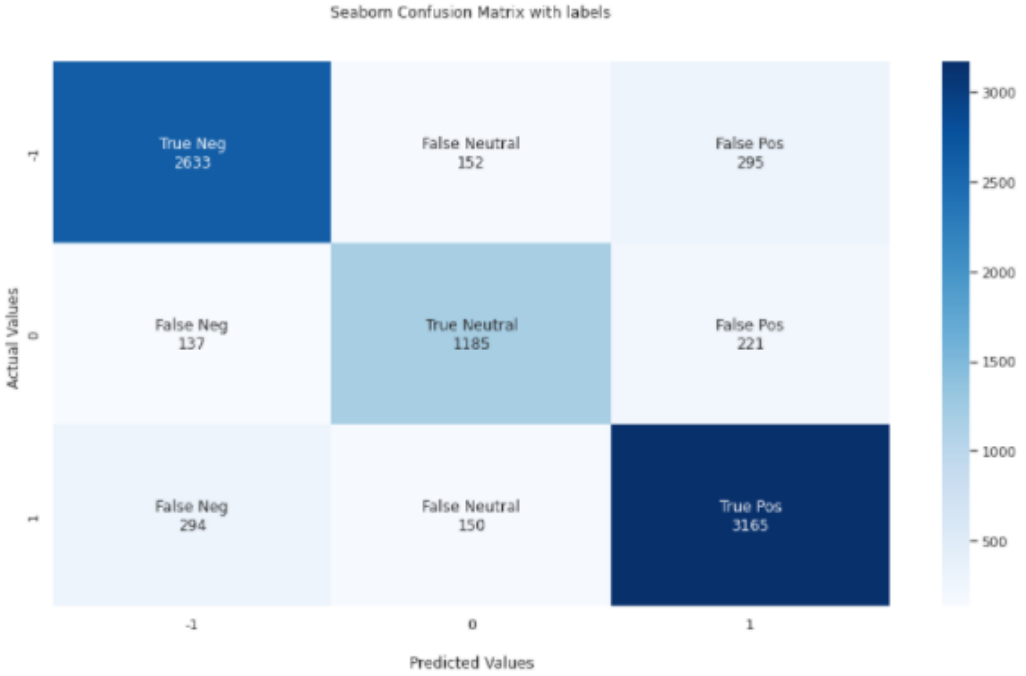
	precision	recall	f1-score	support
-1	0.89	0.88	0.88	12374
0	0.81	0.85	0.83	5821
1	0.90	0.89	0.89	14730
accuracy			0.88	32925
macro avg	0.87	0.87	0.87	32925
weighted avg	0.88	0.88	0.88	32925

LinearSVC Grid Model-Classification Report for test dataset

	precision	recall	f1-score	support
-1	0.85	0.86	0.86	3064
0	0.77	0.80	0.78	1487
1	0.88	0.86	0.87	3681
accuracy			0.85	8232
macro avg	0.83	0.84	0.84	8232
weighted avg	0.85	0.85	0.85	8232



LINEAR SVC AFTER CROSS VALIDATION AND HYPERPARAMETER TUNING



	scoring	Value
0	accuracy_score	0.848275
1	precision_score	0.838687
2	recall_score	0.833276
3	f1_score	0.835862

CONCLUSION

- Among the 5 classification algorithms we performed, Naive Bayes fails to be effective.
- The other 4 models SVM, SGD, Logistic Regression, and CatBoost perform well on the test data, SVM being the best model.
- Linear SVM with 85% accuracy after cross-validation and hyperparameter tuning performed well for multinomial classification.
- We hope this project will be helpful for the Government and NGOs to take adequate measures in policy making and rehabilitation respectively.
- Various profit organizations can make profit through the production and distribution of essential items during the pandemic by analyzing various sentiments.

THANK YOU