Capstone Project-3

Coronavirus Tweet Sentiment Analysis

Lavanya M

CONTENTS

1. Problem statement
2. Introduction
3. About the dataset
4. EDA
5. Preprocessing the text
6. Model fitting
7. conclusion

Problem statement:

The challenge is to build a classification model to predict the sentiment of Corona tweets.The tweets have been pulled from Twitter and manual tagging has been done then.

The names and usernames have been given codes to avoid any privacy concerns.

The following information is given:

1. Location
2. Tweet At
3. Original Tweet
4. Sentiment

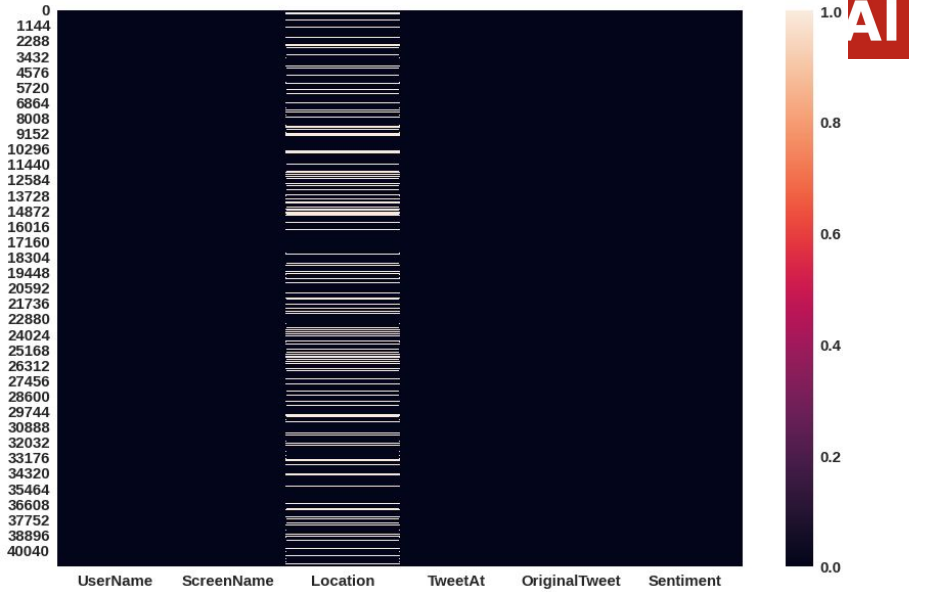
Sentiment analysis :

Sentiment analysis is used to determine whether a given text contains negative, positive, or neutral emotions. It's a form of text analytics that uses natural language processing (NLP) and machine learning. Sentiment analysis is also known as "opinion mining" or "emotion artificial intelligence".

As the Covid-19 outbroke rapidly all over the world day by day and also affecting the lives of millions, a number of countries declared complete lockdown to keep in check its intensity. During this lockdown period, social media platforms have played an important role in spreading information about this pandemic across the world, as people were expressing their feelings through the social networks.

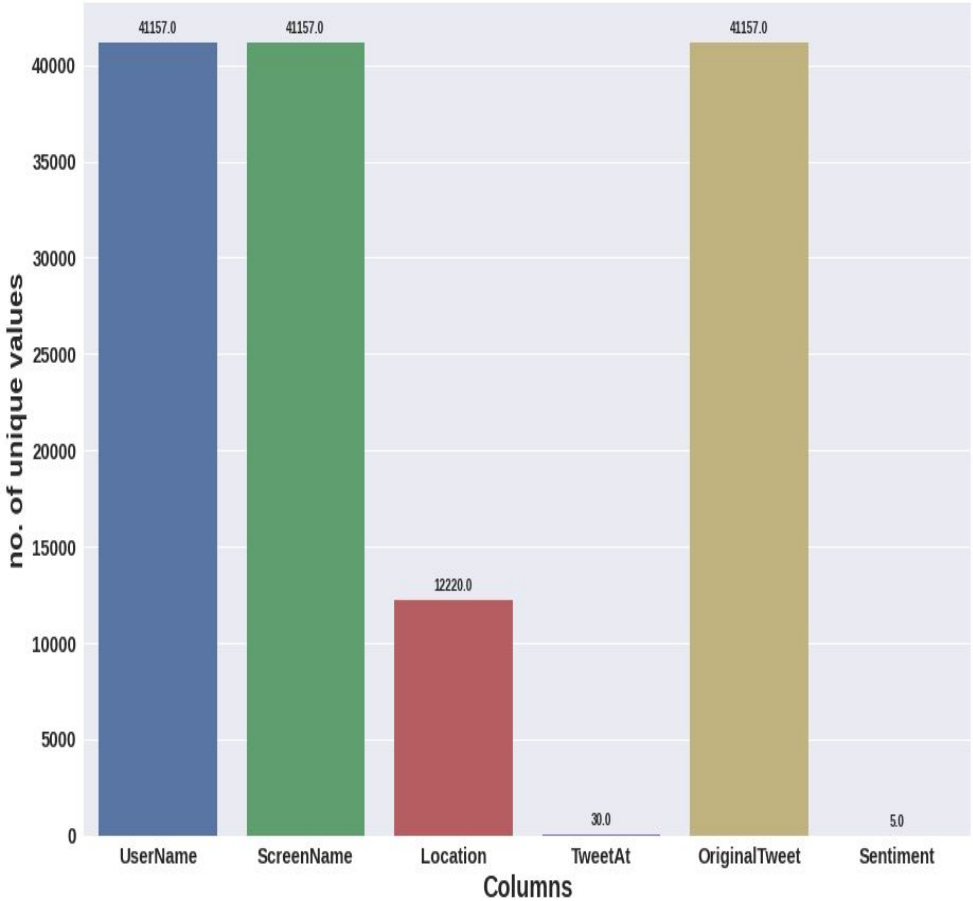
About the dataset:

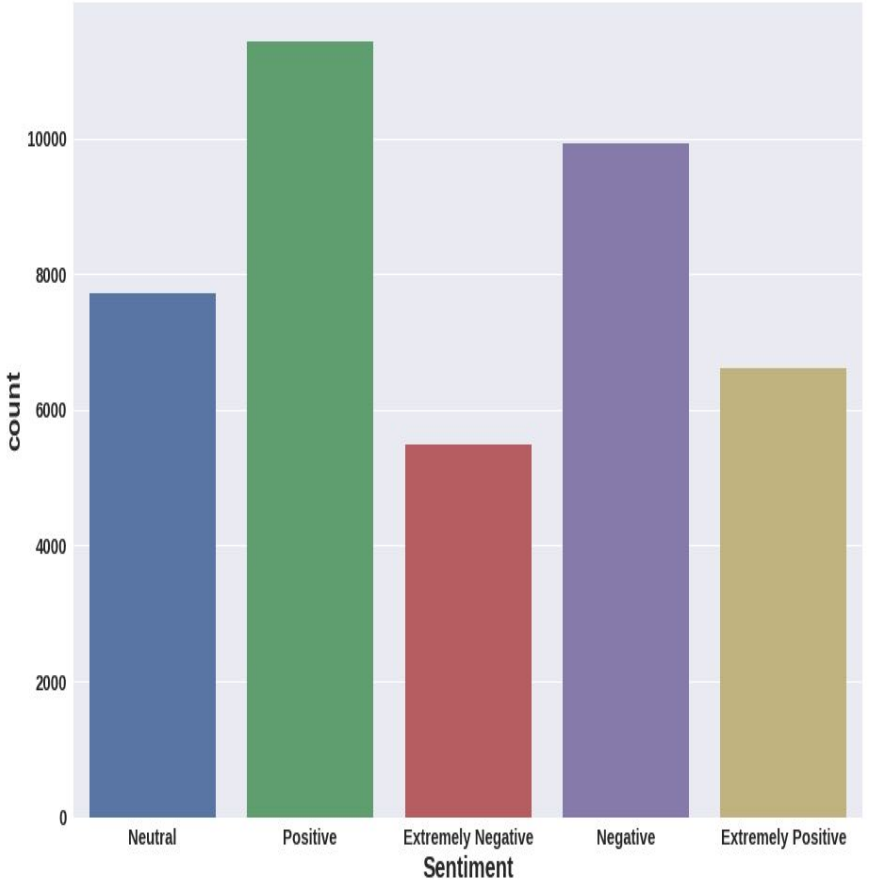
- Dataset has 6 columns and 41157rows.
- Size of dataset is 246942.
- There are no duplicate values in dataset .
- Column ‘Location has 8590 ie, 20.87% null values.
- Sample of the dataset is



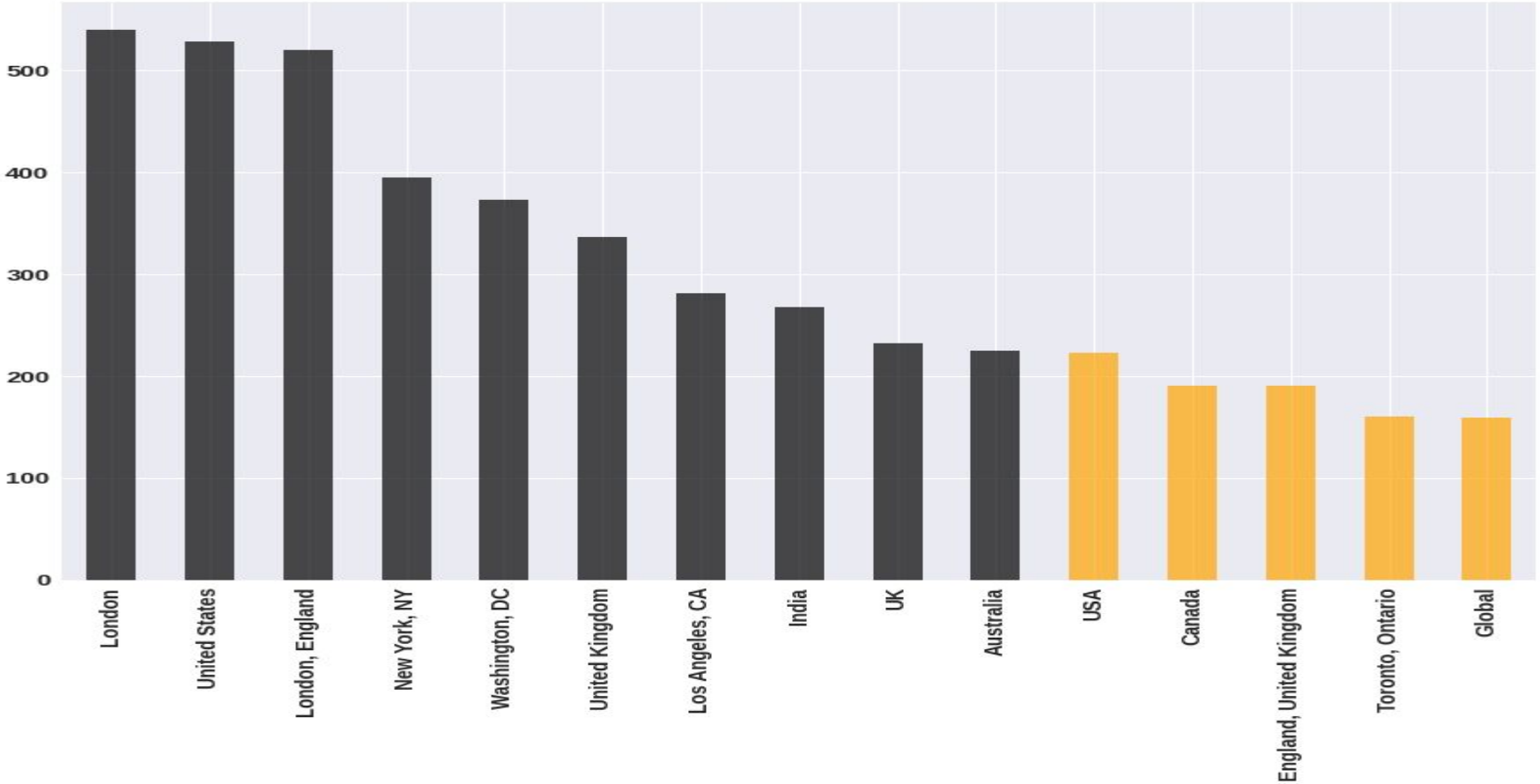
	UserName	ScreenName	Location	TweetAt	OriginalTweet	Sentiment
30969	34768	79720	United States	06-04-2020	@stacyherbert Have you noticed that being at home, not going to cinemas, theater or restaurants, just cooking your own food; not buying unnecessary stuff or clothing; we're spending by far much less money (saving) than before the covid 19 panic?	Negative
7416	11215	56167	Christchurch, New Zealand	19-03-2020	Ah, so that's why the supermarket was so busy when I walked down to do my normal shopping a couple of hours ago. I had not heard the "rumours." #coronavirus #notpanickingyet	Neutral
13687	17486	62438	??	21-03-2020	If you think you'll get #Coronavirus from your local takeaway just remember this: \r\r\n- It's no different to supermarket packaging\r\r\n- You can do contactless delivery\r\r\n- You can wash your hands\r\r\n- You've got more chance getting diarrhea and food poisoning than you do #COVID19	Negative

EDA

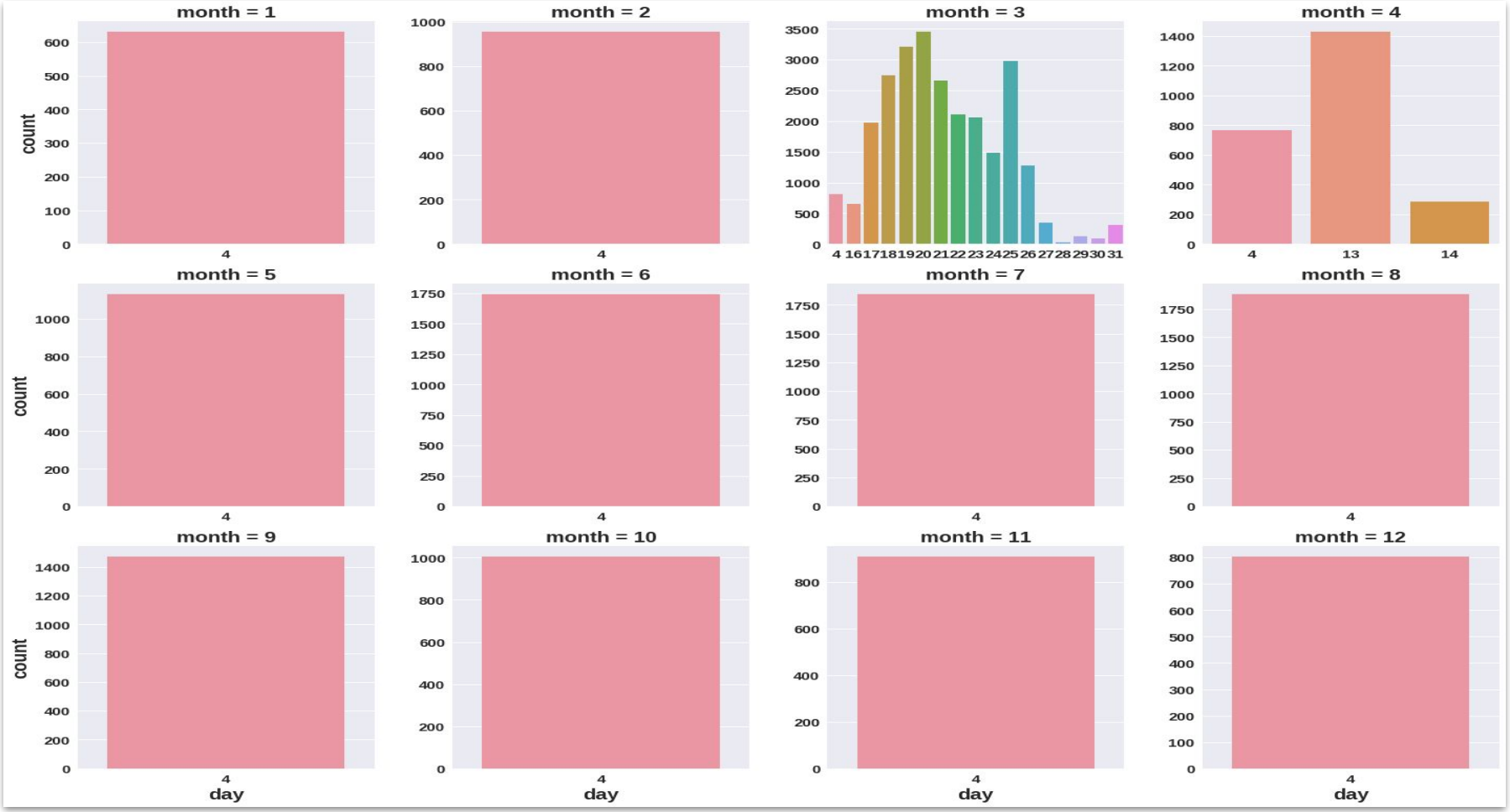

41157.0
41157.0
12220.0
30.0
41157.0
5.0
no. of unique values
0
5000
10000
15000
20000
25000
30000
35000
40000
UserName
ScreenName
Location
TweetAt
OriginalTweet
Sentiment
Columns


count
0
2000
4000
6000
8000
10000
Neutral
Positive
Extremely Negative
Negative
Extremely Positive
Sentiment

Most appeared Locations



Tweets in across months



Preprocessing data

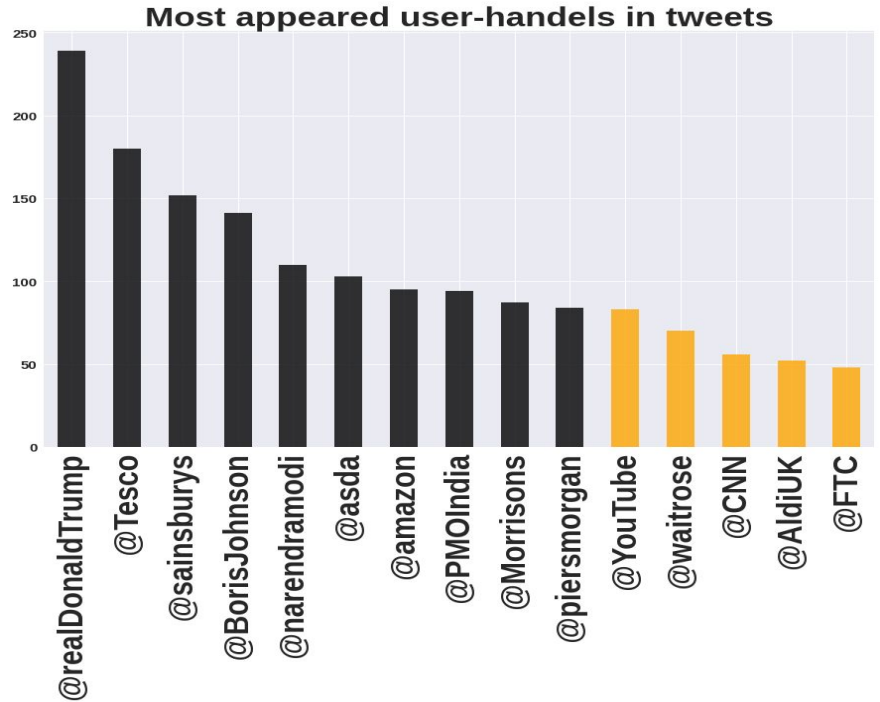
- Data preprocessing is the concept of changing the raw data into a clean data set. The dataset is preprocessed in order to check missing values, noisy data, and other inconsistencies before executing it to the algorithm.

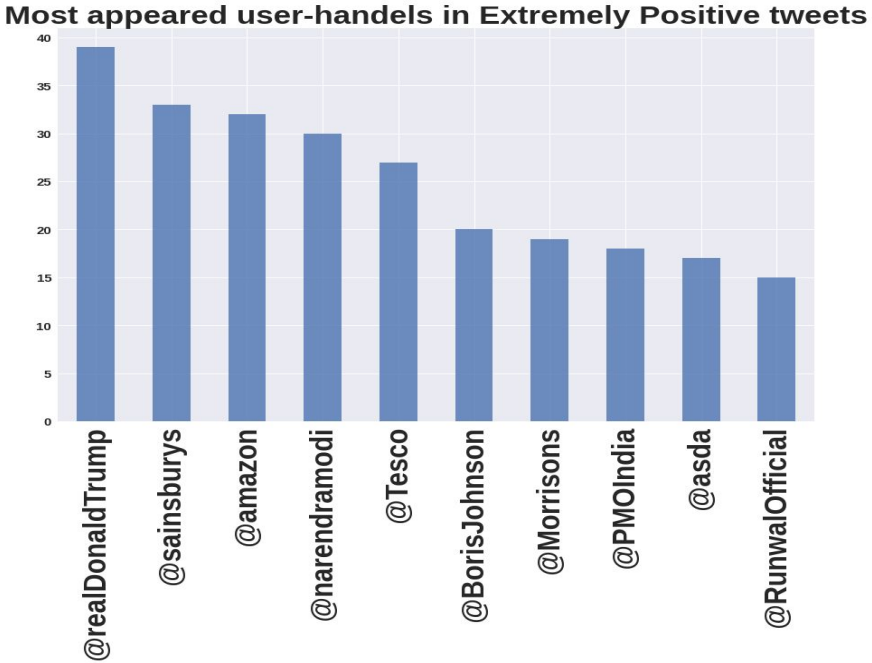
- For textual data there can be special characters ,emojis, urls, punctuation, numbers ,short-words, stop-words,multiple white spaces etc which can be considered as noise as they do not carry much weightage.

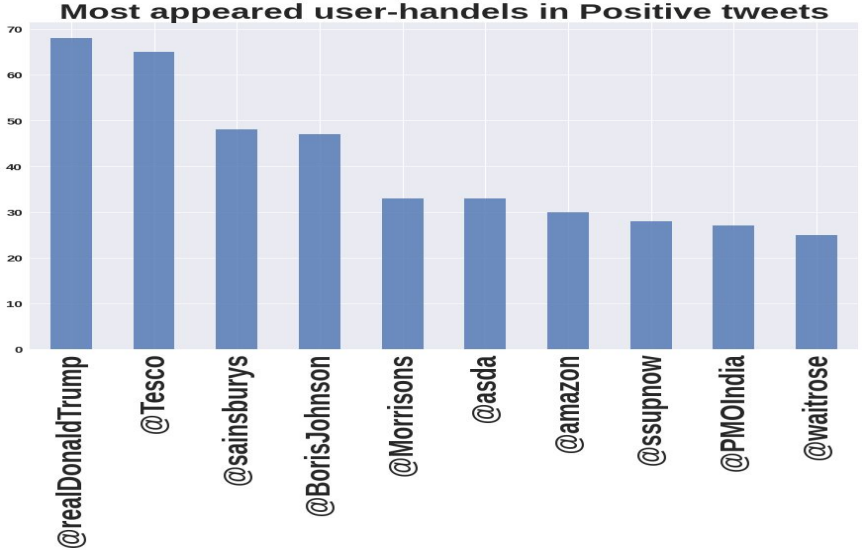
- Data preprocessing is done to remove mistakes, redundancies, missing values, and inconsistencies all that compromise the integrity of the data, we need to fix all those issues for a more accurate outcome before fitting in machine learning algorithm.

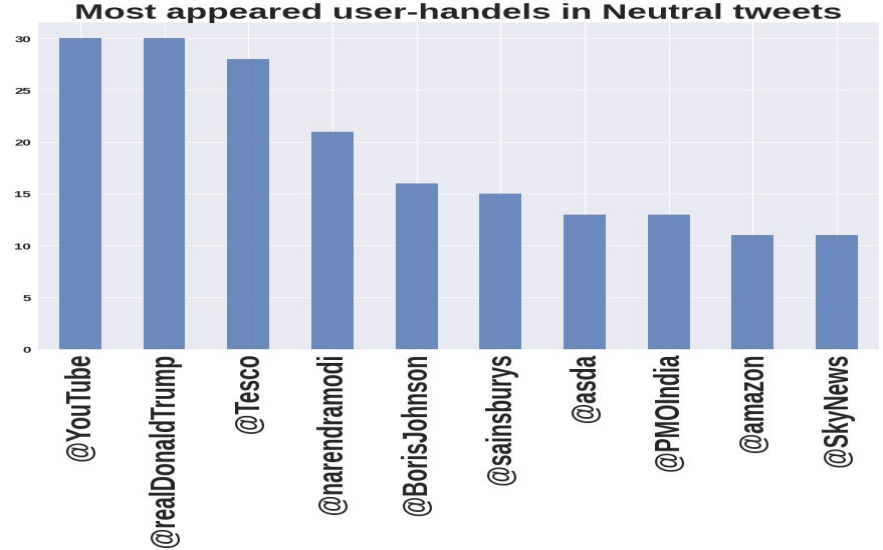
Steps in preprocessing:

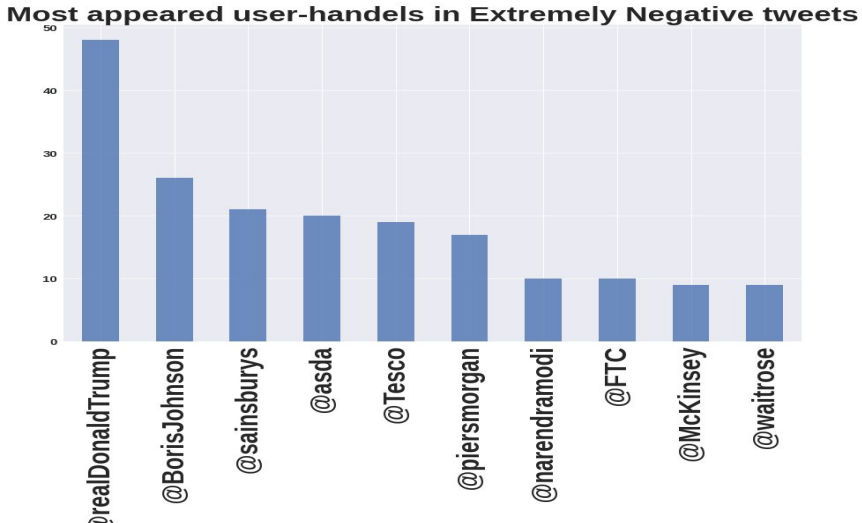
1. Extraction and removal of user-handels form tweet:

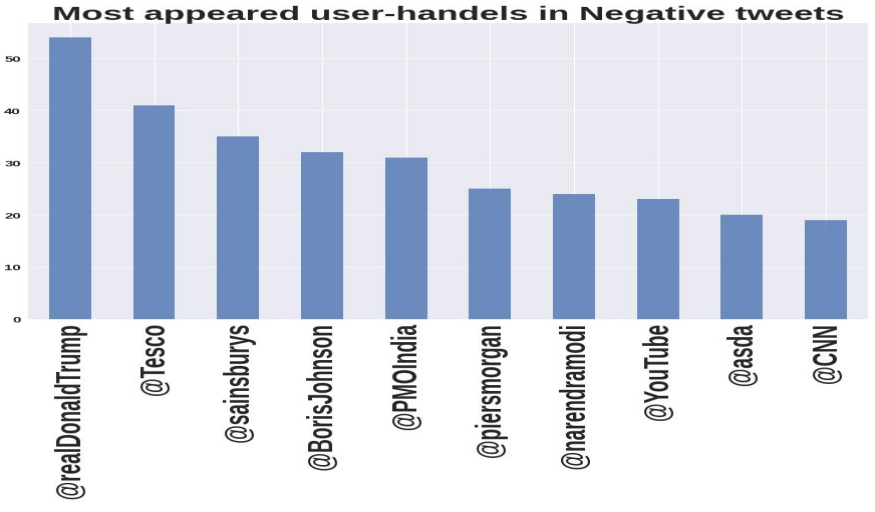





Most appeared user-handels in Positive tweets
70
60
50
40
30
20
10
0
@realDonaldTrump
@Tesco
@sainsburys
@BorisJohnson
@Morrisons
@asda
@amazon
@ssupnow
@PMOIndia
@waitrose


Most appeared user-handels in Neutral tweets
30
25
20
15
10
5
0
@YouTube
@realDonaldTrump
@Tesco
@narendramodi
@BorisJohnson
@sainsburys
@asda
@PMOIndia
@amazon
@SkyNews

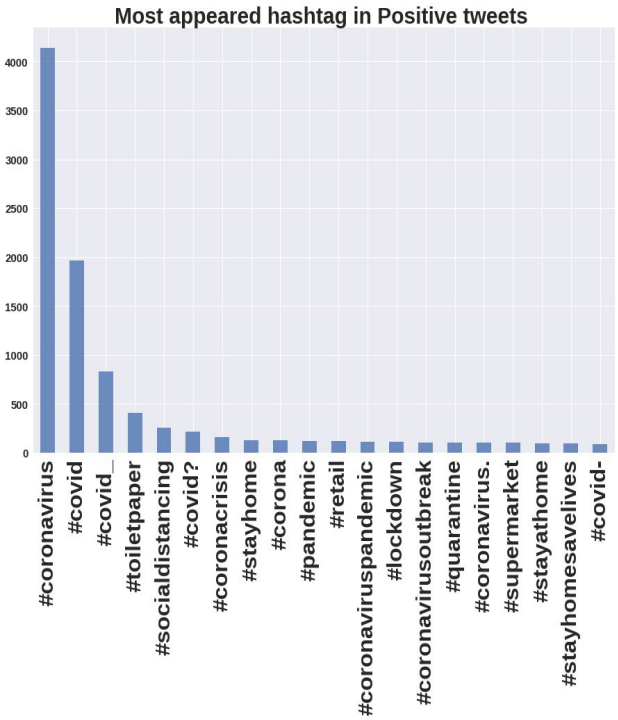

Most appeared user-handels in Extremely Negative tweets
50
40
30
20
10
0
@realDonaldTrump
@BorisJohnson
@sainsburys
@asda
@Tesco
@piersmorgan
@narendramodi
@FTC
@McKinsey
@waitrose

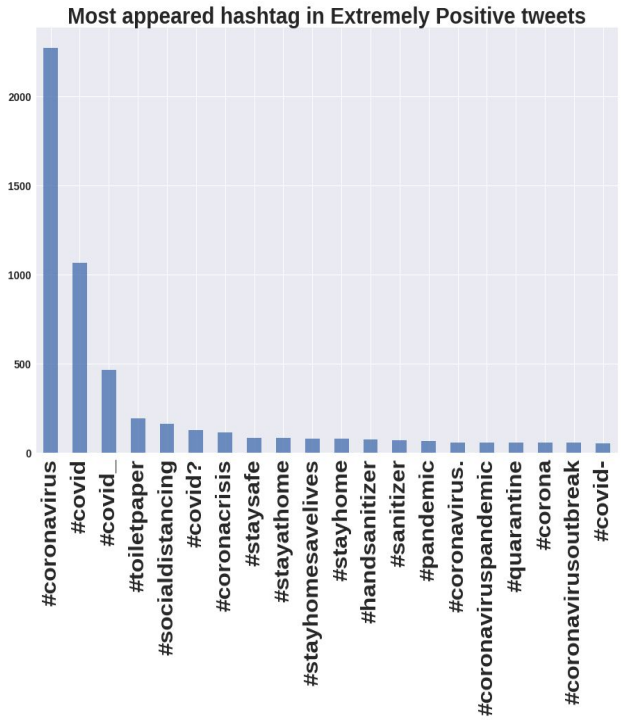

Most appeared user-handels in Negative tweets
50
40
30
20
10
0
@realDonaldTrump
@Tesco
@sainsburys
@BorisJohnson
@PMOIndia
@piersmorgan
@narendramodi
@YouTube
@asda
@CNN

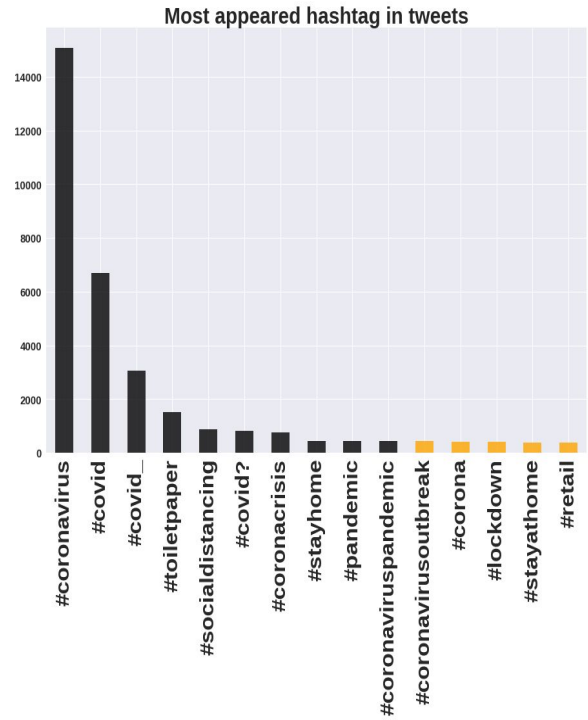
2. Removing urls

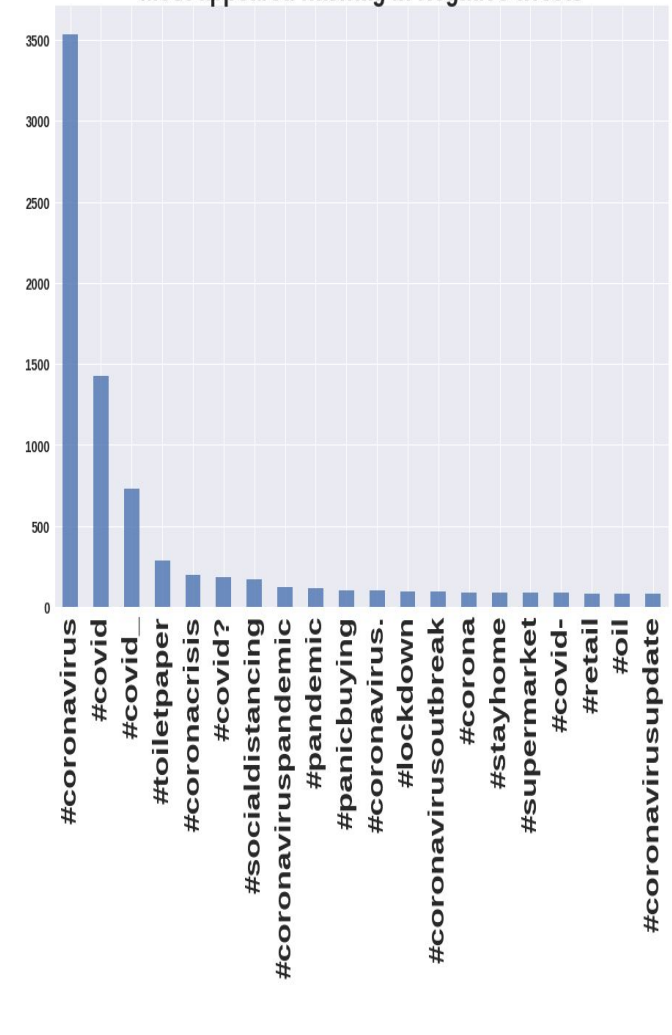
3. Removing numbers

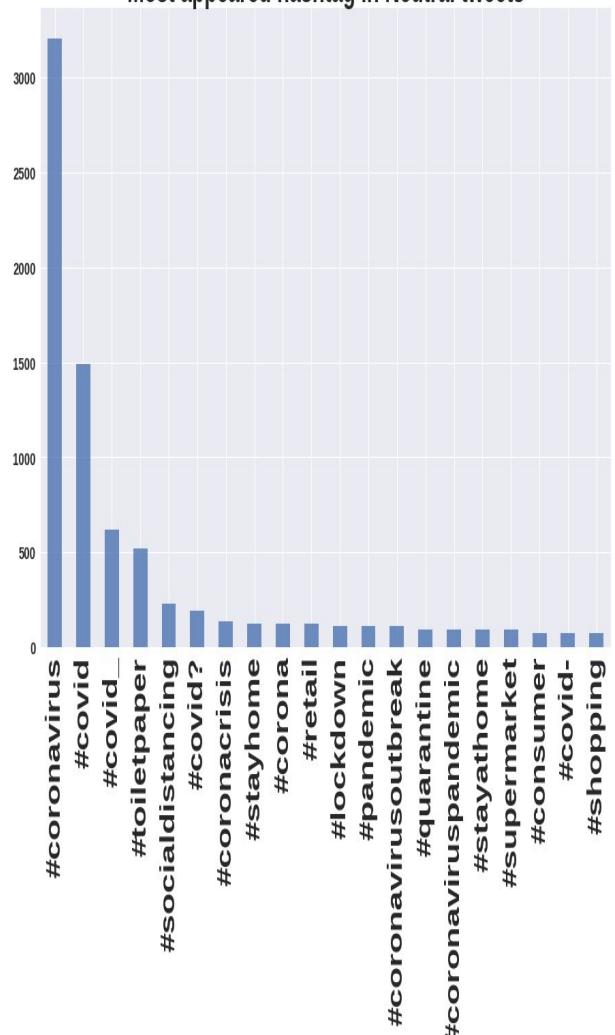
4. Removing special characters (#, /r, /n, -, ?, !, $, Â□, (,), /, * etc and extracting hashtags:

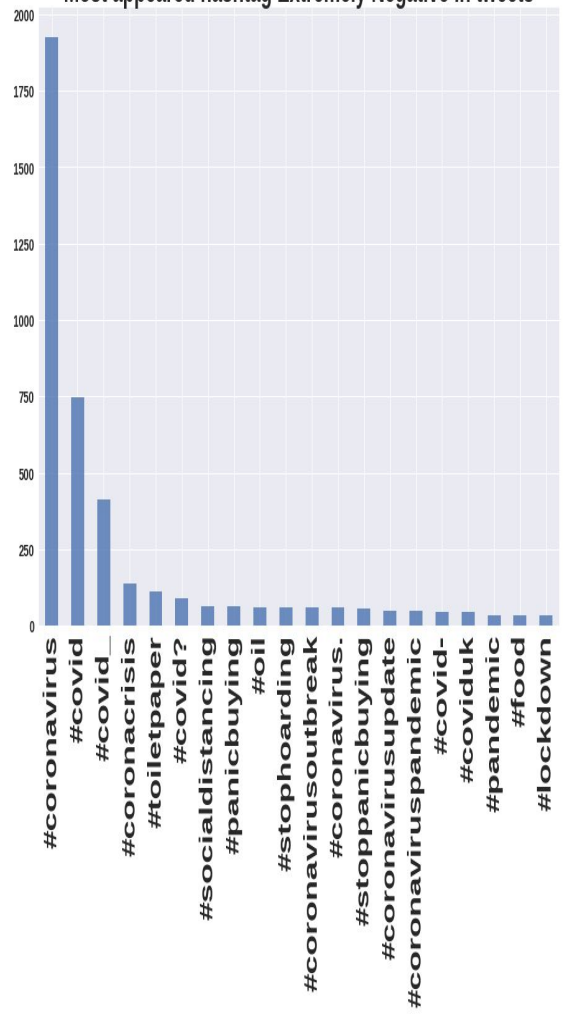







Most appeared hashtag in Negative tweets
3500
3000
2500
2000
1500
1000
500
0
#coronavirus
#covid
#covid_
#toiletpaper
#coronacrisis
#covid?
#socialdistancing
#coronaviruspandemic
#pandemic
#panicbuying
#coronavirus.
#lockdown
#coronavirusoutbreak
#corona
#stayhome
#supermarket
#covid-
#retail
#oil
#coronavirusupdate


Most appeared hashtag in Neutral tweets
3000
2500
2000
1500
1000
500
0
#coronavirus
#covid
#covid_
#toiletpaper
#socialdistancing
#covid?
#coronacrisis
#stayhome
#corona
#retail
#lockdown
#pandemic
#coronavirusoutbreak
#quarantine
#coronaviruspandemic
#stayathome
#supermarket
#consumer
#covid-
#shopping


Most appeared hashtag Extremely Negative in tweets
2000
1750
1500
1250
1000
750
500
250
0
#coronavirus
#covid
#covid_
#coronacrisis
#toiletpaper
#covid?
#socialdistancing
#panicbuying
#oil
#stophoarding
#coronavirusoutbreak
#coronavirus.
#stoppanicbuying
#coronavirusupdate
#coronaviruspandemic
#covid-
#coviduk
#pandemic
#food
#lockdown

5. Removing multiple whitespaces

6. Removing short words

7. Removing stop words

8.Tokenization: In tokenization, we convert a group of sentences into tokens. It is also called text segmentation or lexical analysis. It is basically splitting data into a small chunk of words.

9.Stemming: “Stemming” is a rule-based process of stripping the suffixes (“ing”, “ly”, “es”, “ed”, “s” etc) from a word. For example — “play”, “player”, “played”, “plays” and “playing” are the different variations of the word — “play”.

10.Vectorization: Count Vectorizer will create a sparse matrix of all words and the number of times they are present in a document.

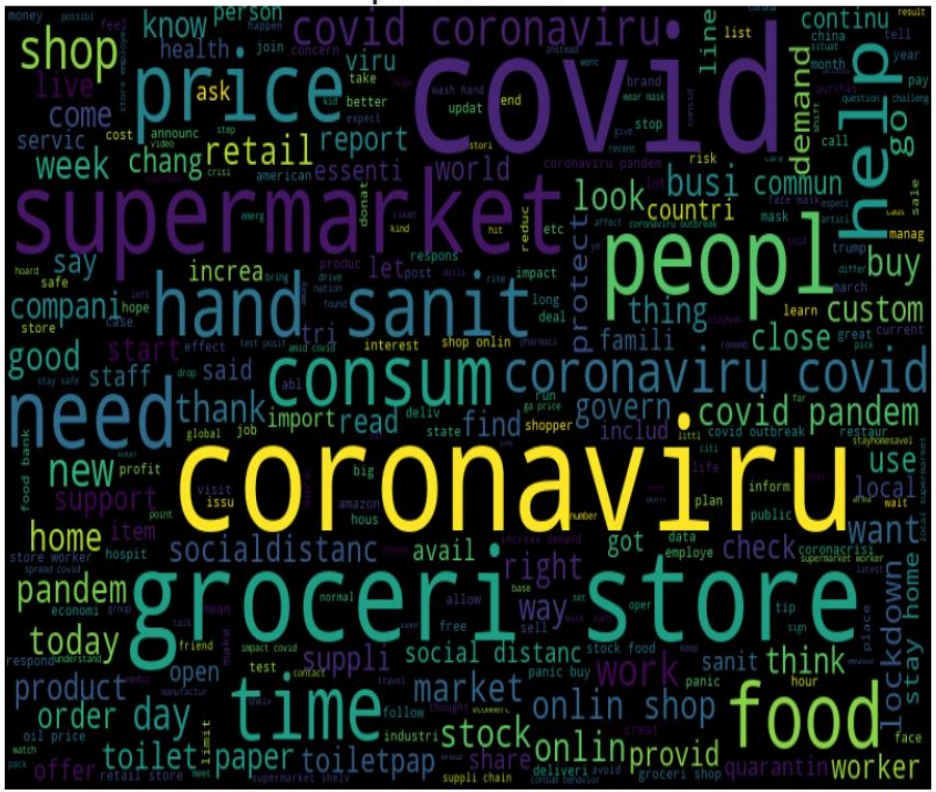
California real estate agents, back into the field with masks, gloves and plenty of sanitizer #coronavirus @mercnews \r\r\nhttps://t.co/0lySOdepuN	california real estate agents back into the field with masks gloves and plenty of sanitizer coronavirus
Older (high-risk for coronavirus) volunteers, increased demand for food and decreasing grocery store donations....The Tri-Cities Food Bank is dealing with multiple consequences of the #coronavirus outbreak, and they need your help! \r\r\n\r\r\nFULL STORY >>> https://t.co/7rUGjdV79e https://t.co/yyCUpZaj4u	older high risk for coronavirus volunteers increased demand for food and decreasing grocery store donations the tri cities food bank is dealing with multiple consequences of the coronavirus outbreak and they need your help full story gt gt gt

Story Generation:

All words wordcloud!

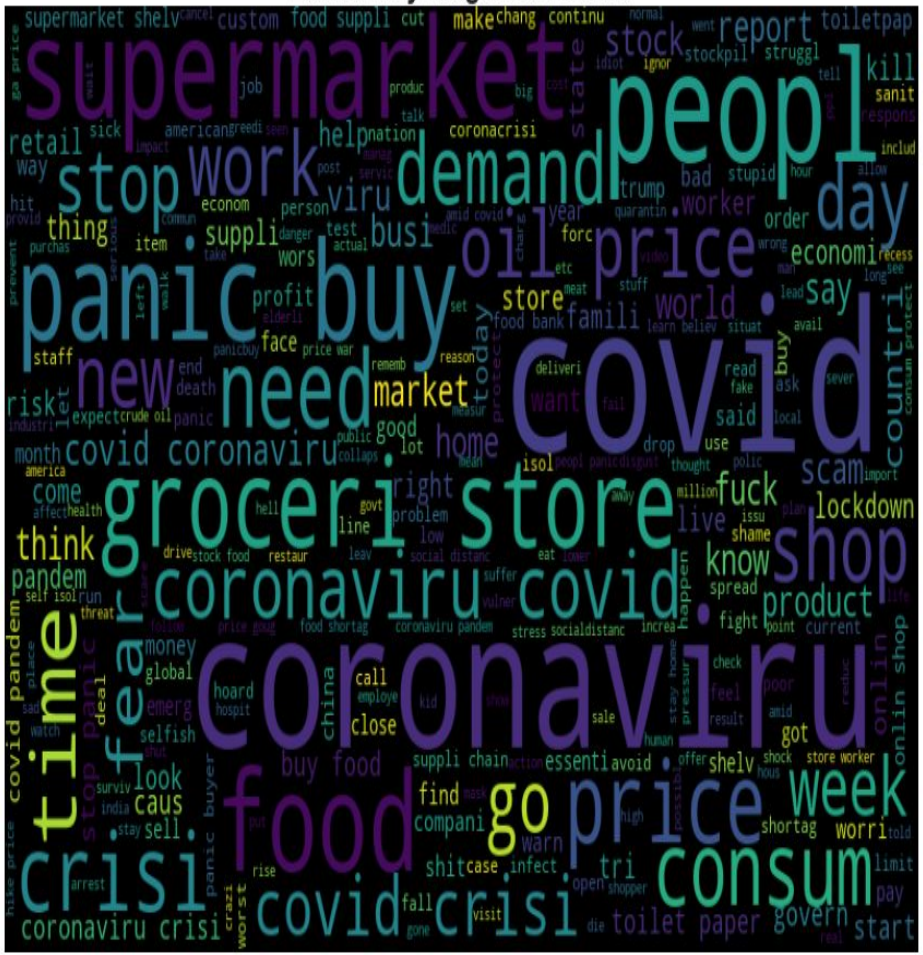
Extremely positive words

positive words

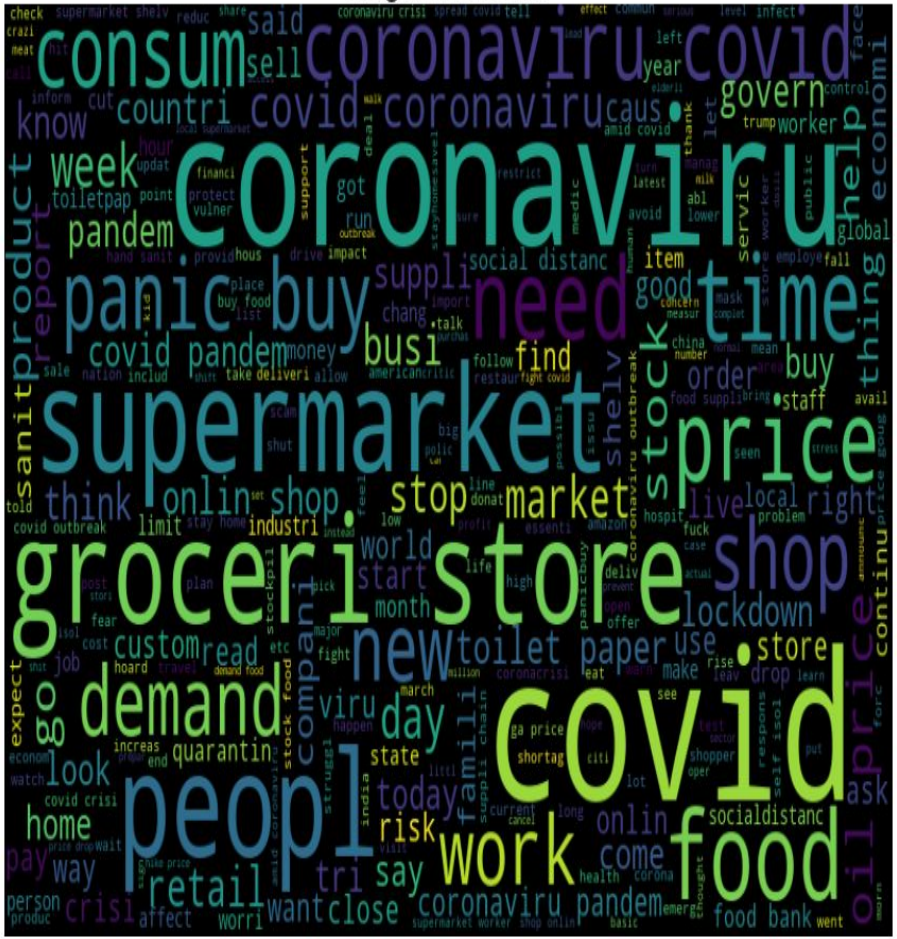


Neutral words

Extremely Negative words



Negative words



Multi-class and binary classification:

- Multiclass- classification is done with 6 classes , ie: Extremely Positive, Positive,Neutral,Extremely Negative and Negative.
  Following algorithms are used in Multiclass classification and binary classification:
  1. Support Vector Machines
  2. Logistic Regression
  3. Naive Bayes
  4. Stochastic Gradient Decent
  5 . XGBoost
  6. CatBoost
  7.Random forest

Evaluation:

Binary Classification

Model	Accuracy
Stochastic Gradient Descent	**0.858844**
Logistic Regression	0.856414
CatBoost	0.844145
Support Vector Machines	0.836978
Random Forest	0.825923
Naive bayes	0.787658
XGboost	0.740889

Muticlass- Classification

Model	Accuracy
CatBoost	**0.609208**
Logistic Regression	0.602648
Support Vector Machines	0.597425
Stochastic Gradient Descent	0.565598
Naive Bayes	0.46694
XGBoost	0.465500

Evaluation contd.

Multiclass Classification catboost

```
Training accuracy Score    :  0.6559453302961276
Validation accuracy Score :  0.6092079689018465
                  precision    recall  f1-score   support

Extremely Negative       0.55      0.73      0.63       835
Extremely Positive       0.55      0.78      0.64       932
          Negative       0.51      0.56      0.54      1809
           Neutral       0.81      0.59      0.68      2124
          Positive       0.62      0.56      0.59      2532

          accuracy                           0.61      8232
         macro avg       0.61      0.64      0.62      8232
      weighted avg       0.63      0.61      0.61      8232
```

Binary Classification Stochastic Gradient Descent

```
Training accuracy Score    :  0.9364312832194381
Validation accuracy Score :  0.858843537414966
              precision    recall  f1-score   support

           0       0.77      0.84      0.80      2811
           1       0.91      0.87      0.89      5421

    accuracy                           0.86      8232
   macro avg       0.84      0.85      0.85      8232
weighted avg       0.86      0.86      0.86      8232
```

Conclusion

- A social media sentiment analysis tells us how people feel about certain brand, issue and person online.
- For multiclass classification, the best model for this dataset would be CatBoost for our dataset
- For binary classification, the best model for this dataset would be Stochastic Gradient Descent for our dataset.

Thank you