Capstone Project

Coronavirus Tweet Sentiment Analysis

Presented By:
Sandeep R

Introduction

The given challenge is to build a classification model to predict the sentiment of Covid-19 tweets. The tweets have been pulled from Twitter and manual tagging has been done. We are given information like Location, Tweet At, Original Tweet, and Sentiment.

One should know what is mean by 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.

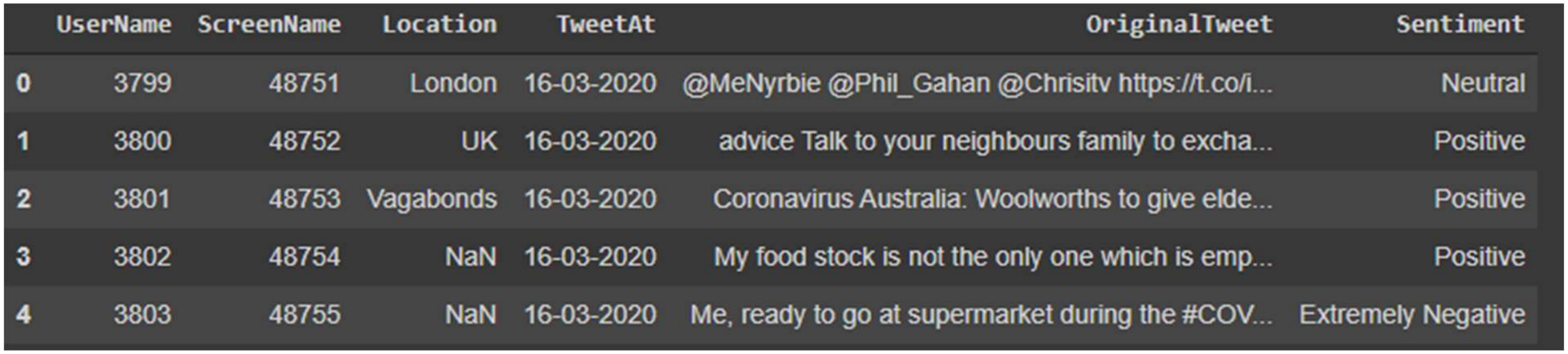
Standard Operating Procedure

Following is the Standard Operating Procedure to tackle the Sentiment Analysis kind of project. We will be going through this procedure to predict what we supposed to predict!

- **Exploratory Data Analysis.**
- **Data Preprocessing.**
- **Vectorization.**
- **Classification Models.**
- **Evaluation.**
- **Conclusion.**

Data Summary

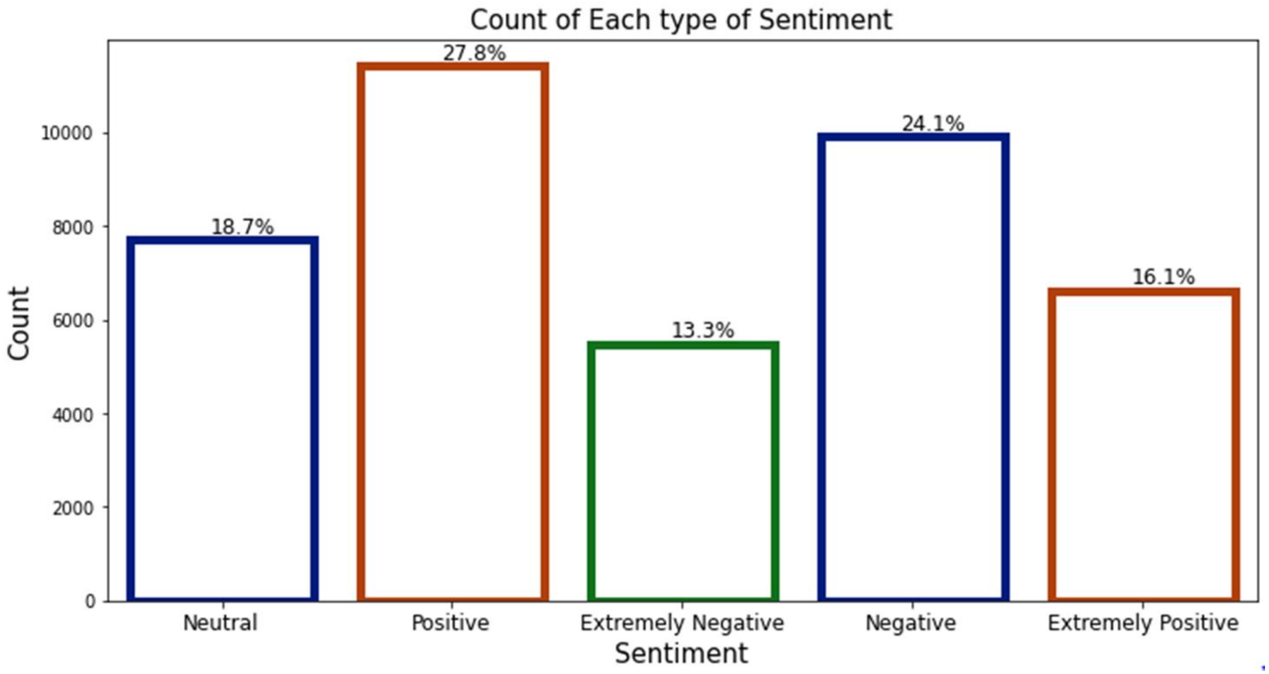
The original dataset has **6 columns and 41157 rows**. In order to analyze various sentiments, We require just two columns named Original Tweet and Sentiment. There are five types of sentiments- **Extremely Negative, Negative, Neutral, Positive, and Extremely Positive** as you can see in the following picture.



	UserName	ScreenName	Location	TweetAt	OriginalTweet	Sentiment
0	3799	48751	London	16-03-2020	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Neutral
1	3800	48752	UK	16-03-2020	advice Talk to your neighbours family to excha...	Positive
2	3801	48753	Vagabonds	16-03-2020	Coronavirus Australia: Woolworths to give elde...	Positive
3	3802	48754	NaN	16-03-2020	My food stock is not the only one which is emp...	Positive
4	3803	48755	NaN	16-03-2020	Me, ready to go at supermarket during the #COV...	Extremely Negative

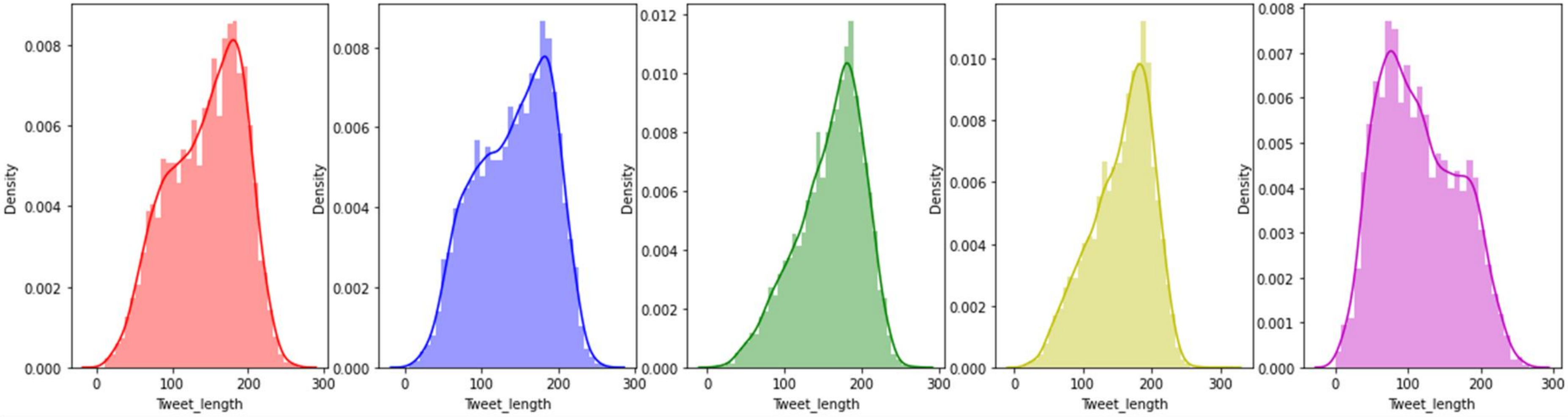
Basic Exploratory Data Analysis

When we try to explore the ‘Sentiment’ column, we came to know that most of the peoples are having positive sentiments about various issues shows us their optimism during pandemic times. Very few people are having extremely negatives thoughts about Covid-19.



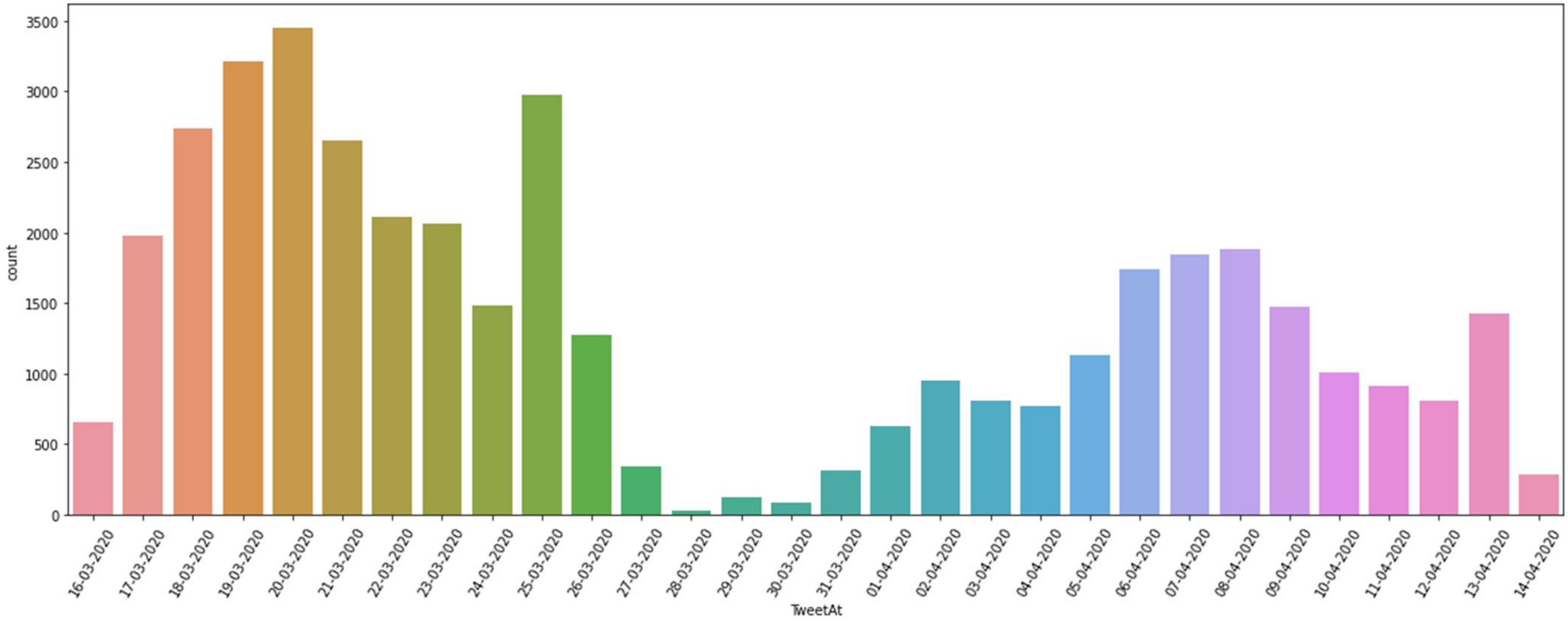
Distribution of message Length for each type of Sentiment

From the below Distribution plot it can be seen that all the distributions are equally Distributed. Length of the Message doesn't depend on the sentiment of the tweet.



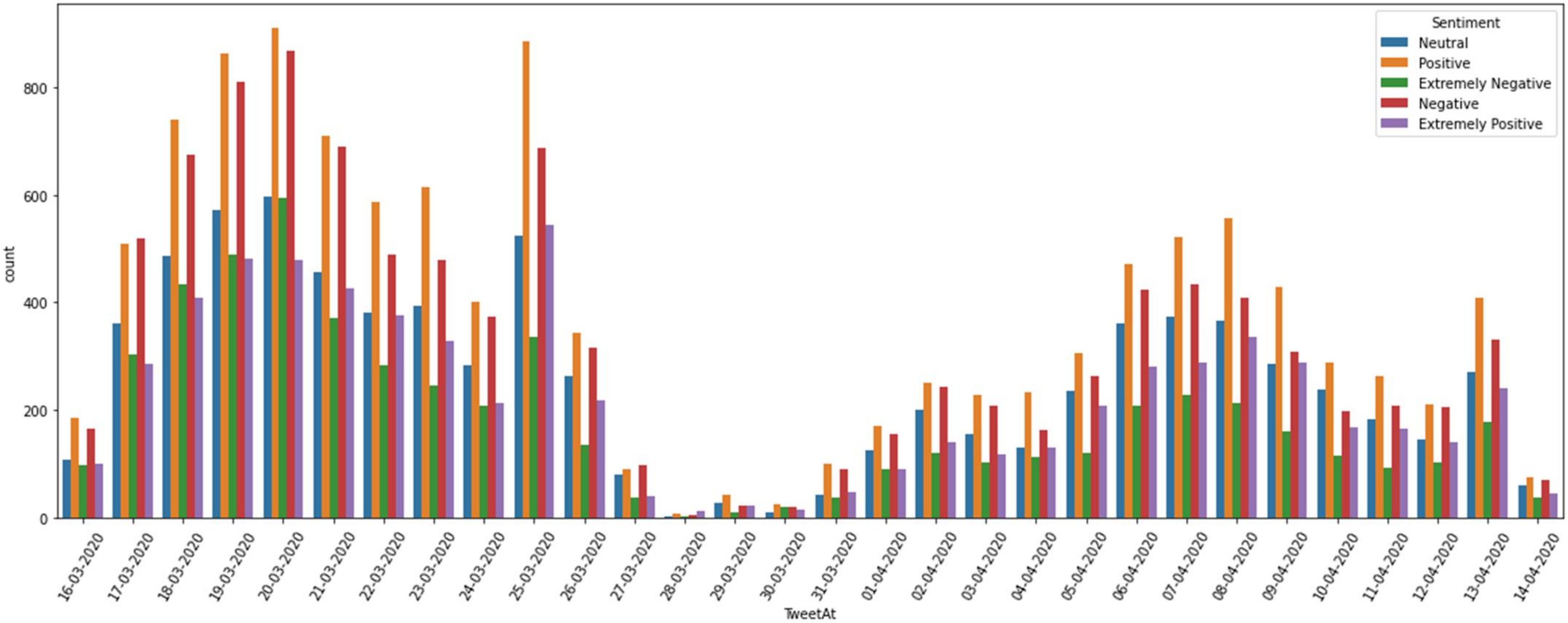
Count of Tweets For Each Day

Initially the Count of tweets were quite high but gradually the count of tweets related to corona started to decrease.

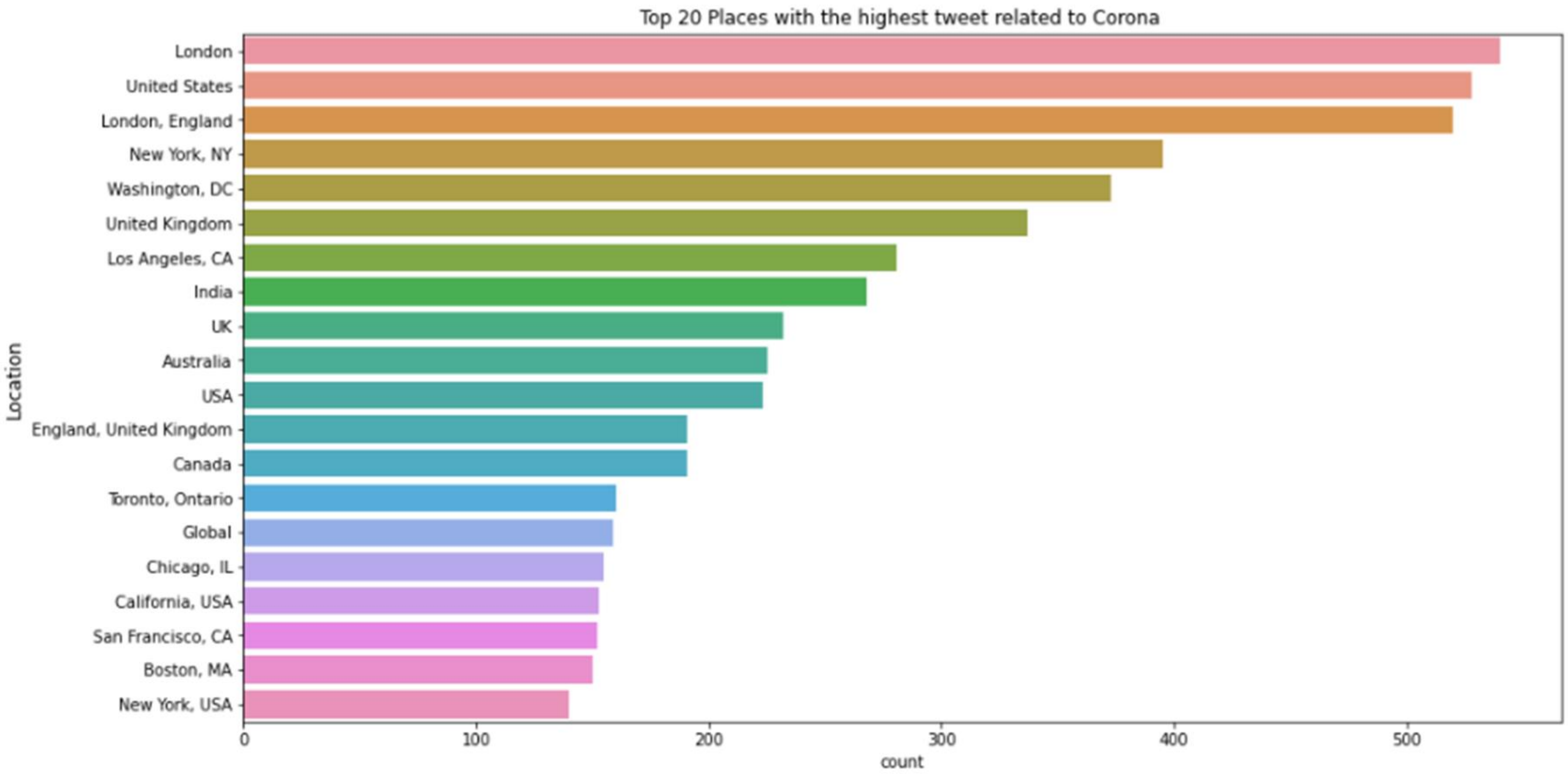


Count of Types Tweets For Each Day

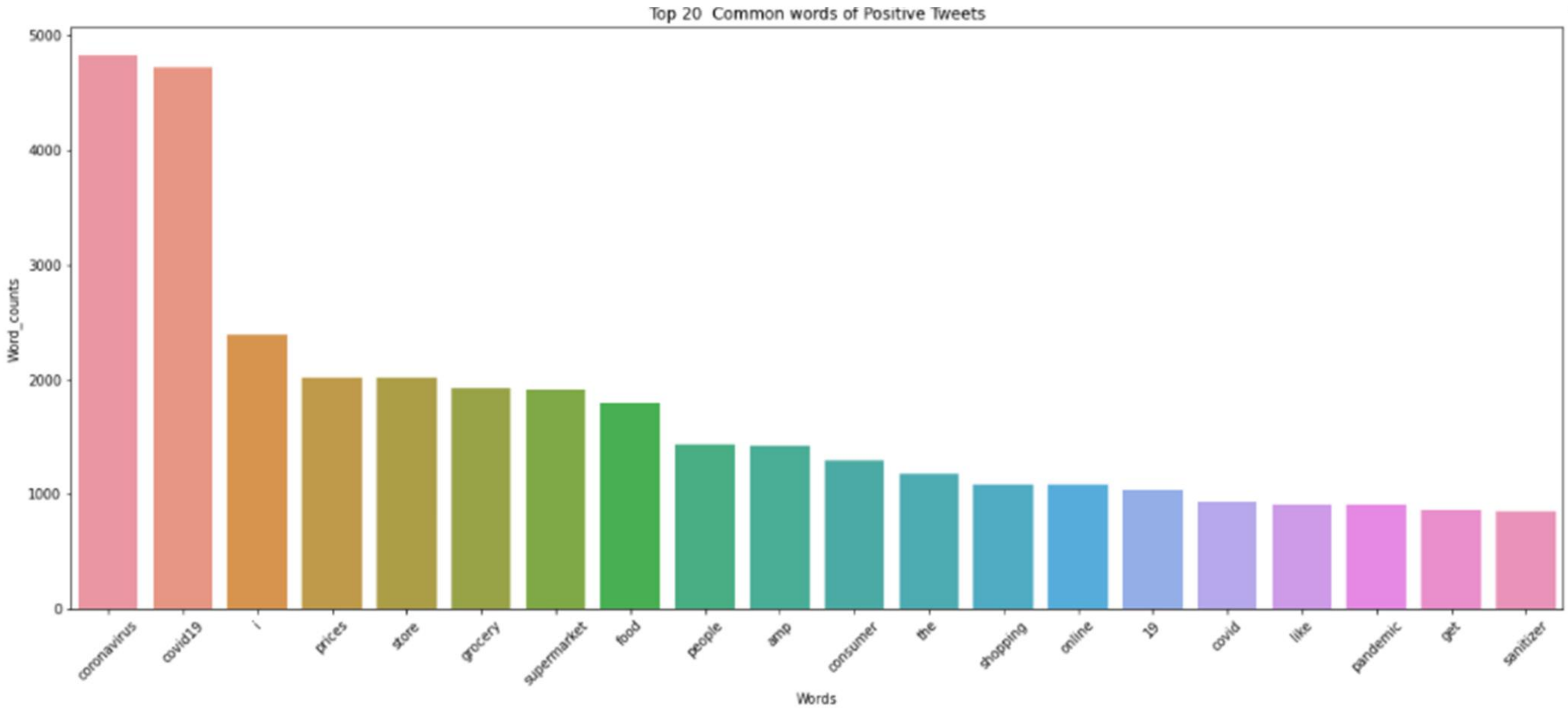
Almost all the days we can see that the Count of Positive tweet was high as compared with other kind of tweets.



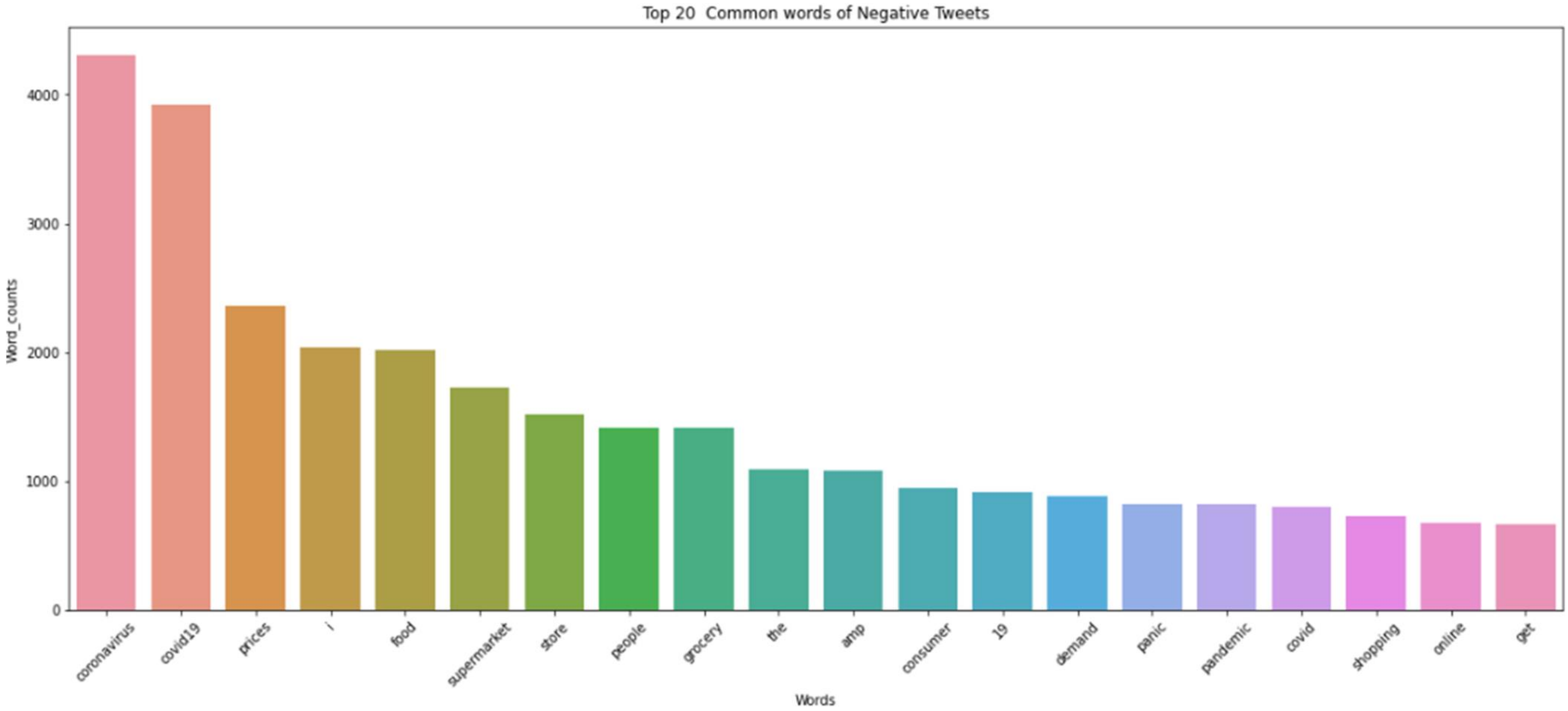
Geographical Location with Highest Tweets



Common Words in Positive Tweets



Common Words in NegativeTweets



Result of Training Different Models

Models	Test_score
SVC	0.603984
LogisticRegression	0.582483
RandomForestClassifier	0.528912
GradientBoostingClassifier	0.464893
DecisionTree	0.442906
XGboost	0.440233
KNN	0.214893

Conclusion

In this way, we can explore more from various textual data and tweets. Our models will try to predict the various sentiments correctly. I have used various models for training our dataset but some models show greater accuracy while some do not. **For multiclass classification, the best model for this dataset would be SVC.**

Thank You