# 1.Introduction:

The world’s struggle with the problems in health, social economy arises with the Covid-19 spreads. And it is very essential to make more accurate decision every moment during such condition like pandemic. Use of data from all the sources can be very key part in analysis and among them social networking sites can be very effective place of data to get information about public opinions, sentiments and any updates from authorities. Tweeter is one of the popular social networking sites where people and business share their opinions and updates. So, data from this can be useful for such analysis and hence in this project we are using twitter dataset for the sentiment analysis and to do some descriptive and diagnostic analysis.

Since our dataset is text so we are using here we are using natural language processing language (NLP) techniques for the text classification. The NLP is the branch of AI which enables computer to understand human language and which includes the text classification also where text is classified into predefined categories.

## 1.1 Accessing the data:

In this article we work with tweeter’s tweet dataset accessed from Kaggle.

For this project the source of the data is provided via https://www.kaggle.com/datasets/datatattle/covid-19-nlp-text-classification . Data is download on the Google Drive and work is performed through Google Colab.



Fig.1: Mounting google drive data to use in google colab

## 1.2 Data Dictionary

There are two dataset with the title of Corona\_NLP\_train and Corona\_NLP\_test in CSV format. Getting to know the variables:

Username: It is user code for individual tweeter user who is doing tweets.

ScreenName: It is also integer value which represents indexing for every individual tweet posted screen.

Location: It’s the location from where tweet was posted (Like London, Vagabonds etc.)

Tweet At: It’s the time at which user did tweets.

Original Tweet: It’s the text file or tweets that has been posted by user.

Sentiment: It’s the type of feelings that present in text file in the tweets.

## 1.3 Task and action

The main objectives or task of this project are as follows:

-To do descriptive analysis to know about tweets location and timing pattern as well as sentiment analysis quantitatively.

- To explain the differences in terms of sales of total quantity, products and dates with years.

- To create word cloud of tweets with positive, negative and neutral emotions.

- To get insights about top topics for the negative sentiments

- To make classifier model for sentiment analysis for the tweets.

# 2. Preparation

## 2.1 Packages

The Python 3 environment comes with many helpful libraries that are useful for analysis. So in this project we loaded library like Pandas library for data processing, Numpy library as np for linear algebra Matplot library and Seaborn library for data visualization tools. It is also required some of machine learning algorithm so library like Sk learn and also library for NLP has been imported as shown in figure below.

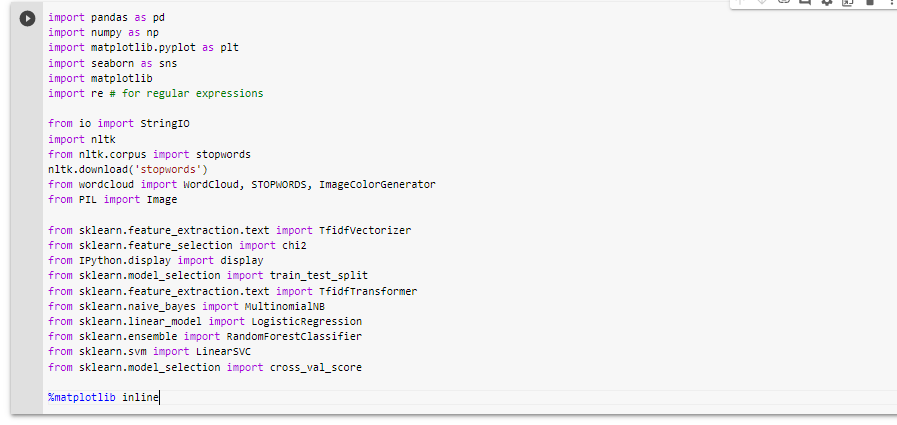
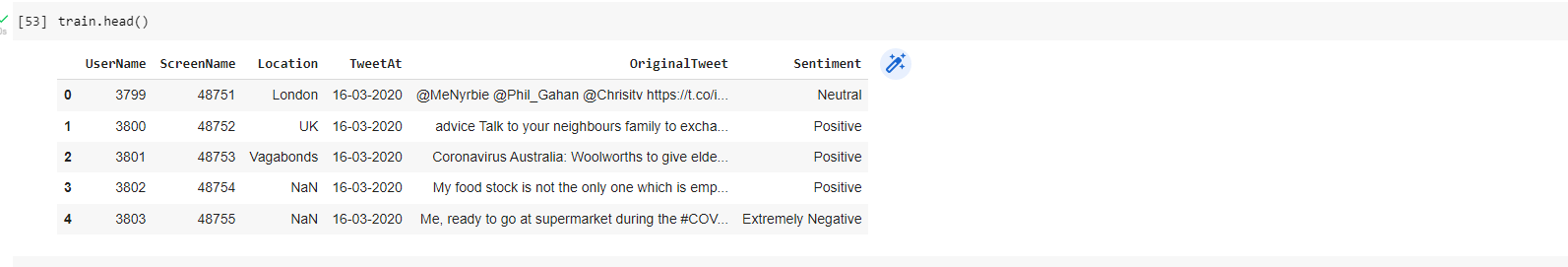
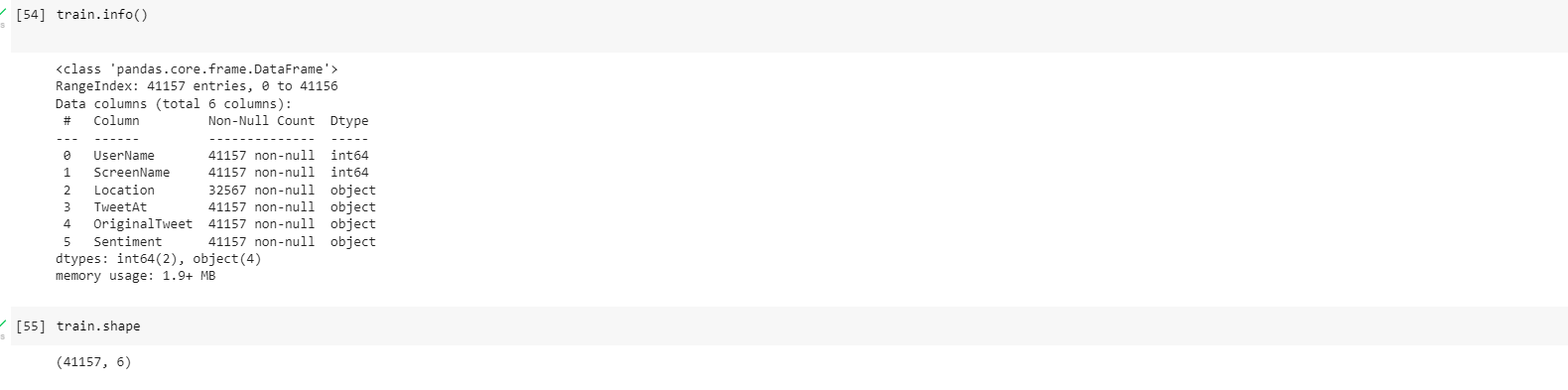


Fig 2. Importing libraries

## 2.2 Understanding Data

Both training and test dataset seems similar in columns number and data types. So after importing libraries now, we read data with code below and the we proceed for understanding characteristics of data like its shape, first 5 rows data, data types, summary statistics, unique values and missing values.







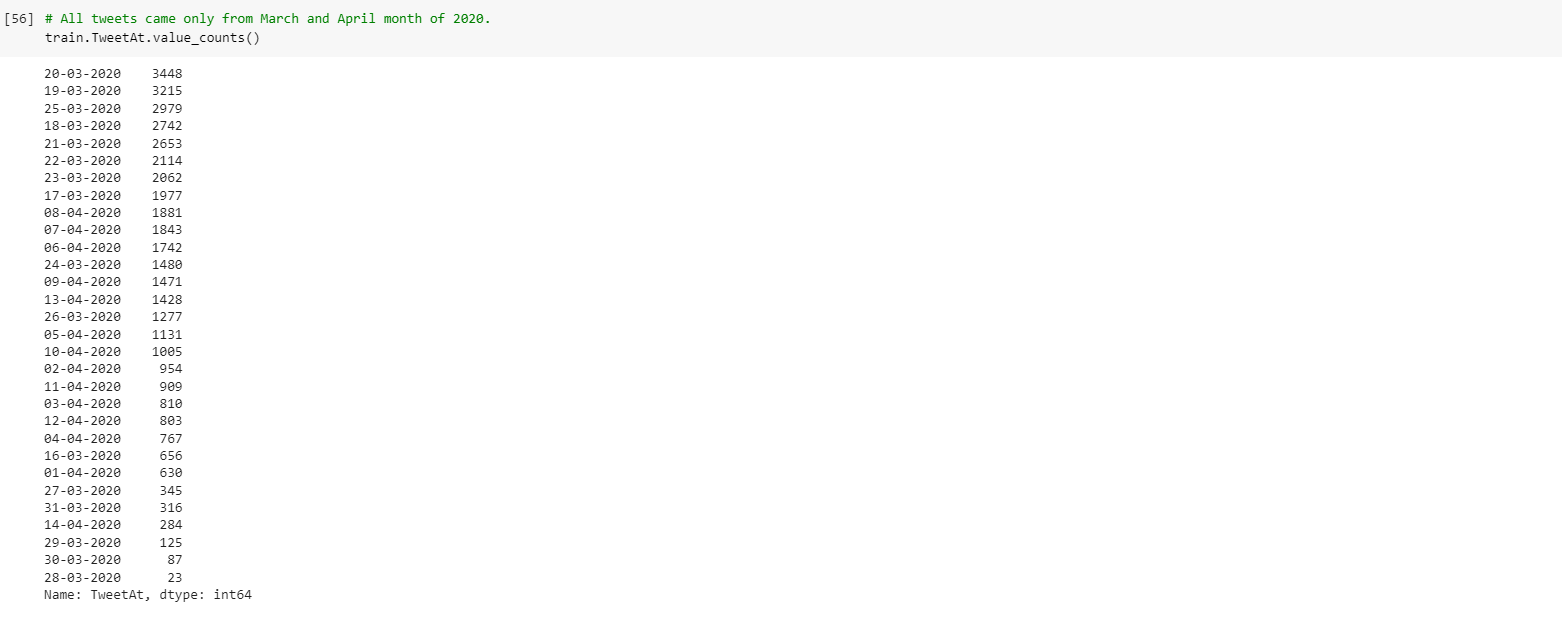


Fig 3. Understanding data

# 3. Exploratory Data Analysis

## 3.1 Data Wrangling

Especially training dataset might have duplicate value which can affect in accuracy of model build from it. So, we are using drop duplicate function of python to remove such duplicate values. And we check if there’s null or missing values with the use of isnull function as below but founds to be no missing value in required columns specially.





Fig 4. Data Cleaning process

Further on data processing since it is the text classification problem so we have to clean text file. As tweets may contains the hashtags, URLs , number, space, stop word, mentions etc. which has to clean so we have used the self-described functions to find those things in the data and remove with the way as per below.



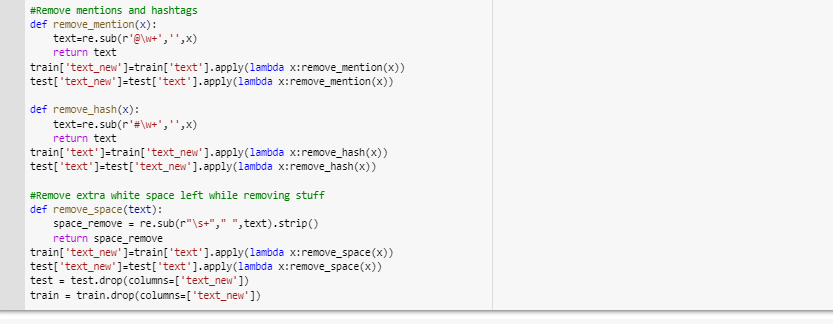


Fig 5: Data processing for tweets text data

## 3.2 Descriptive analytics

On the descriptive data analysis, we try to find valuable insightful information from the data. Here we tried to find out the number unique values within the features of our dataset and plotted it into bar graph as below.

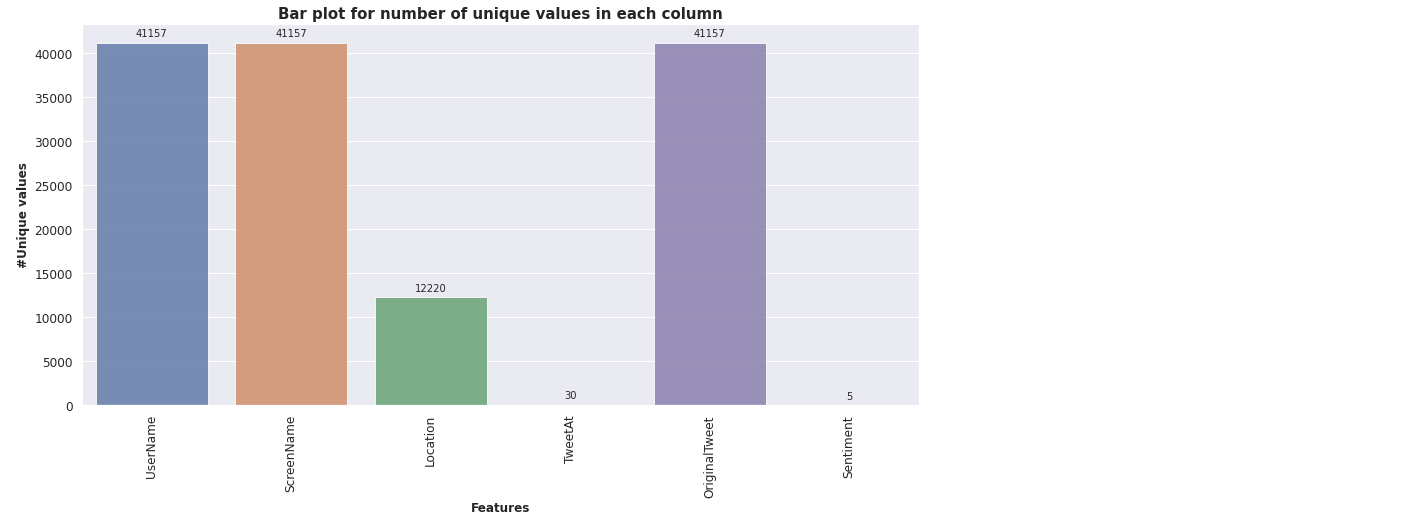


Fig. 6: Bar graph for unique values count in each features:

From above bar graph we found that there are total 12220 number of locations from where tweet was done so we try to find the top ten location among them with the use for value count and head function of pandas library.



Fig.7. Top 10 location of tweet

As we have total five distinct sentiment as per above bar- graph so to analyses the total count for those sentiments we generated further more bar graph for sentiments count.



Fig.8: Bar graph for Sentiment count

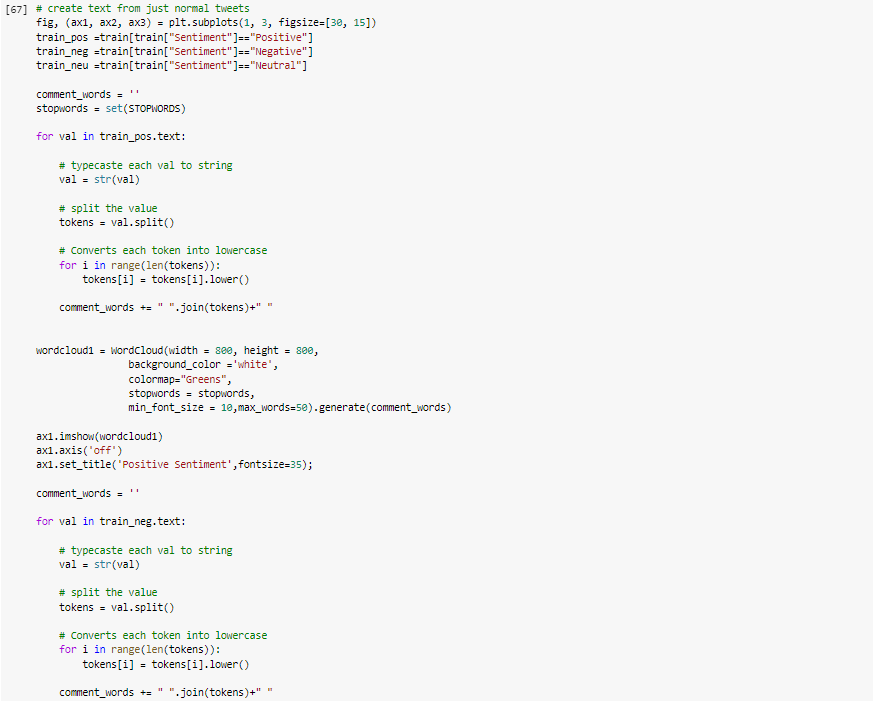
We found that almost 10,000 tweets from our dataset are with negative sentiment. So its easy that we can find top ten things that people are worried as per the tweeter data set and we have plotted count of those top words in the bar plot below.

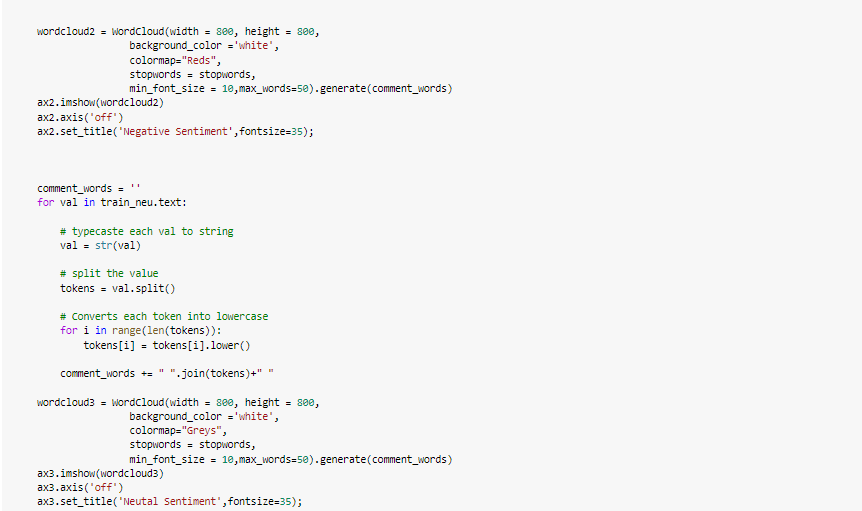
****

Fig.9: Bar plot of Tope 5 negative word

## 3.3 Diagnostic analytics:

Diagnostic analysis usually gives the very important information about why certain perticuler relationship occurred to support of decision of data. As ower project in about the text classifiication so draing the word cloud for our dataset can give very good support for diagnostic analysis. Since from word cloud we can see why certain sentiment on the tweets.

****

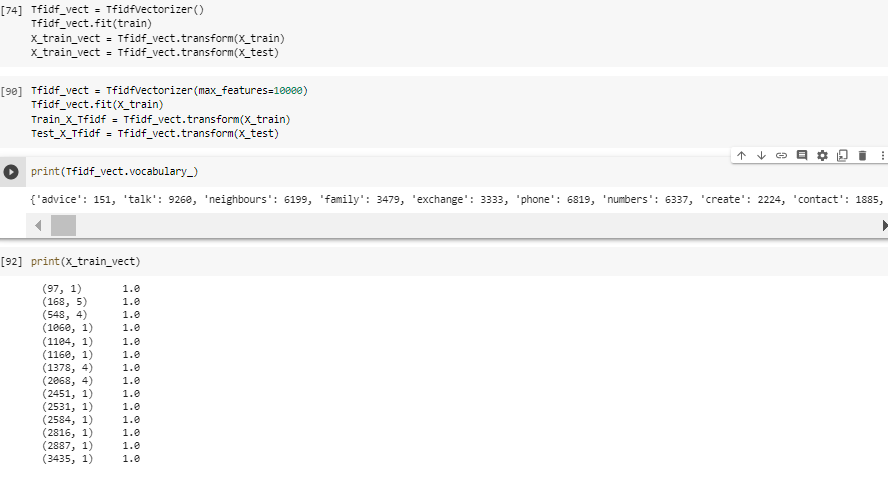


## 

Fig.10: Word cloud for sentiment

# 3.3 Predictive analytics:

Here for the predictive analysis of this project of tweets classfication we used TFIDF (Term frequency- inverse document frequency) with which we quntify or vetctorized text file.



We used three other populer model and cross validation is done and accuracy is calculate to find the best model by the way as per below. And we found linear support vector machine gives the best accuracy hence we use this classifier



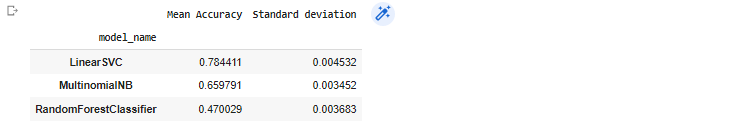


Fig.11: Accuracy of models comparision

Hence lets build the classifier model with linear support vector machine. And also out put of prediction can be seen in array as below for predicted sentiment from test dataset.

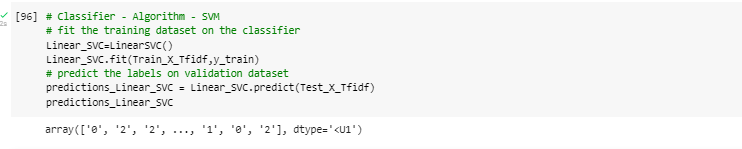


Fig.12: Linear SVC Classifer

# 4. Conclusion/ Recommendation:

From the descriptive analytics, diagnostic analytics and predictive analytics we point out the following conclusion:

1. The top location of tweets are London, New-York and Washingtons in our dataset. And number positive sentiment is higher than other sentiments.
2. The top five topics for the negative sentiments are coronavirus, COVID19, Covid\_19, Coronavirus and COVID2019. Here some words like covid 19 and covid 19 seems similar word so if we take distinct word than top five also includes the toilet paper, social distancing, pandemic. So it seems people are more worried about toilet paper. Social distancing, supermarket, food also.
3. Linear Support vector machine is giving best accuracy with accuracy rate of 78%.

From the dataset of covid tweets it is really hard to differentiate that whether particular word is positive and negative as so many words occurred in top numbers. Hence more advance method which can analyze as a whole sentence technique to get more accuracy in the model. So more than general machine learning algorithm it might more accurate with the use of neural network techniques for this predictive analysis in the future.

# References

* www.kaggle.com. (n.d.). *Coronavirus tweets NLP - Text Classification*. [online] Available at: https://www.kaggle.com/datasets/datatattle/covid-19-nlp-text-classification.

‌