BIG DATA PROJECT REPORT

TWITTER SENTIMENT ANALYSIS FOR

COVID-19

A Team Project By:

Varsha Raghavendra

Chhavi Kumar

Shrey Kshatriya

Table of Contents:

Introduction

The novel coronavirus that first emerged in Wuhan, China back in November has now been declared a pandemic. Also known as the COVID-19 virus, it has brought business and daily life to a standstill. It has been over 2 months and people are still into lockdown, economies are failing, and deaths tolls are not looking to stop.

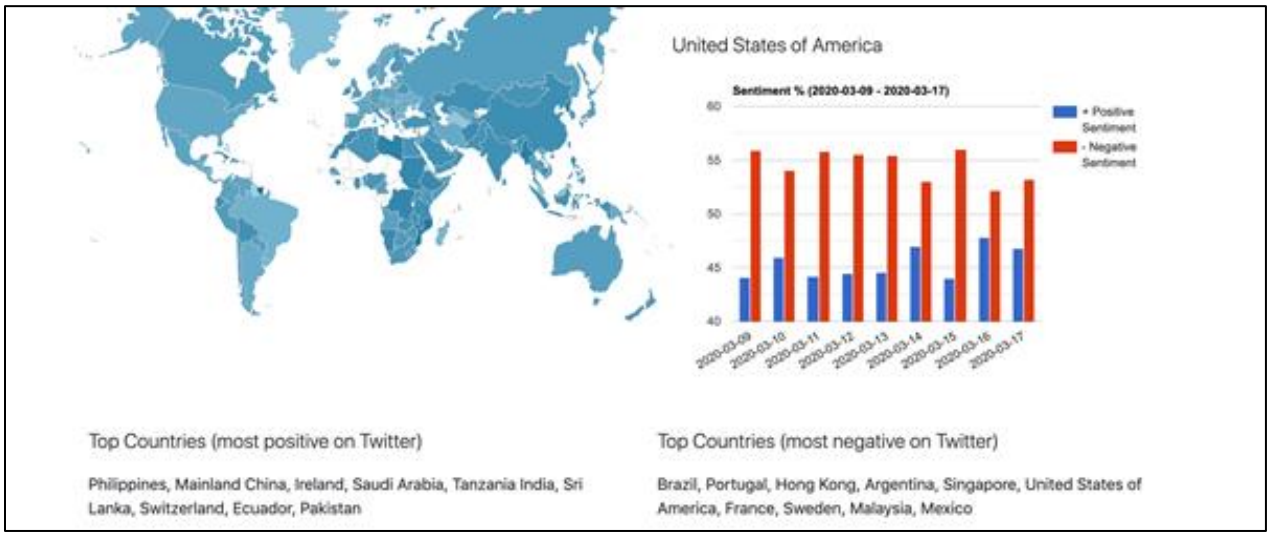
Social media plays a pivotal role in communicating information during global crisis events. We are interested in improving detection and analysis of misinformation and exposing as much false, misleading and clickbait content as possible, so that people can guard themselves against it. Moreover, in terms of research, we want to further research understanding about information spread during pandemics, and whether social media can provide insights into public perception about policy measures in real time.

To understand what sentiment analysis is, we first need to understand what a sentiment is. A sentiment is an emotion or an opinion towards something. It can be a topic or something in general. This can be either positive, negative, or neutral. This is known as sentiment. Now, when we understand the sentiment of masses, with a "score" attached to it, then it is called sentiment analysis.

However, why is it so important right now? And why use twitter? As we know the current situation where everyone is locked inside their home, the only way to know what they feel is through social media. Performing sentiment analysis will help the government and the higher organizations to carry out social media monitoring. It will also help them to get a wider view of the public opinion.

So, if the response is positive, they can continue to do what they are doing, or if the response is negative, they can carry out countermeasures.

For example: Looking at the beginning of the outbreak through the first day that the first person outside of China was diagnosed with the virus (January 1-13) there were 48.7K Tweets mentioning the coronavirus with 27% of those tweets expressing fear. There were 46.7 million more tweets mentioning the coronavirus between January 14-February 28, and during that period the level of fear in Tweets talking about the coronavirus has remained remarkably static, between 15%-16%.



But how do we perform sentiment analysis? We cannot just look at some few 100 tweets and say 40 were positive, 40 were negative and the rest were neutral. We need to scrape real-time tweets from twitter to keep the sentiments up to date. We extracted tweets from March, April, and early week of May with the keywords, “#corona”, “#coronavirus”, and “#COVID-19”. We also visualized tweets to show the trends in the sentiments. After that we calculated the tweet sentiment and measured the performance of the algorithms against time, scale, and accuracy.

Earlier we aimed to prove that the number of negative tweets will decrease, and the number of positive tweets will increase. However, with our analysis of the tweets on this subject we detected that negative sentiment of the tweets indeed decreased, but in turn the number of positive tweets also decreased.

Data Collection

To perform sentiment analysis on COVID-19 tweets, the first step was to gather the data. We utilized Tweepy, an open source library which enabled us to gather Tweets automatically through the Twitter API across the months of March, April, and May.

This is the data that we used for:

1. Training the machine learning model, and
2. Analyzing the trend of opinion of masses on the pandemic.

The data set consisted of 12344 rows and 6 columns which referred to:

1. Tweet ID
2. Tweet content
3. Date of creation
4. Month of creation
5. Sentiment: Score given to each tweet using TextBlob Library
6. Classes
    a. The classes variable was the target variable for the machine learning model.
    b. It is a categorical variable that classified tweets into 0, 1, 2 categories which corresponded to Negative, Neutral and Positive sentiment, respectively.

Data Understanding

The next step was exploring the data and analyzing it for any patterns. As can be seen in the Fig 1 below, the number of tweets with neutral sentiment outnumbered the ones with the negative and positive sentiment.

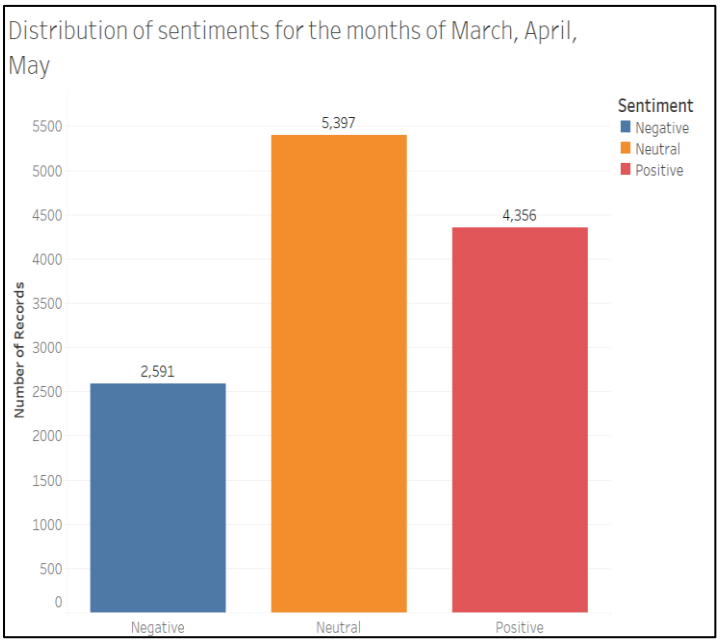


Fig 1

We also observed that neutral sentiment decreases in April and begins to rise as of the first week of May whereas negative sentiment decreases drastically in the first week of May. The positive sentiment occupies about 30-33% of the total tweets across all three months. (Fig 2)

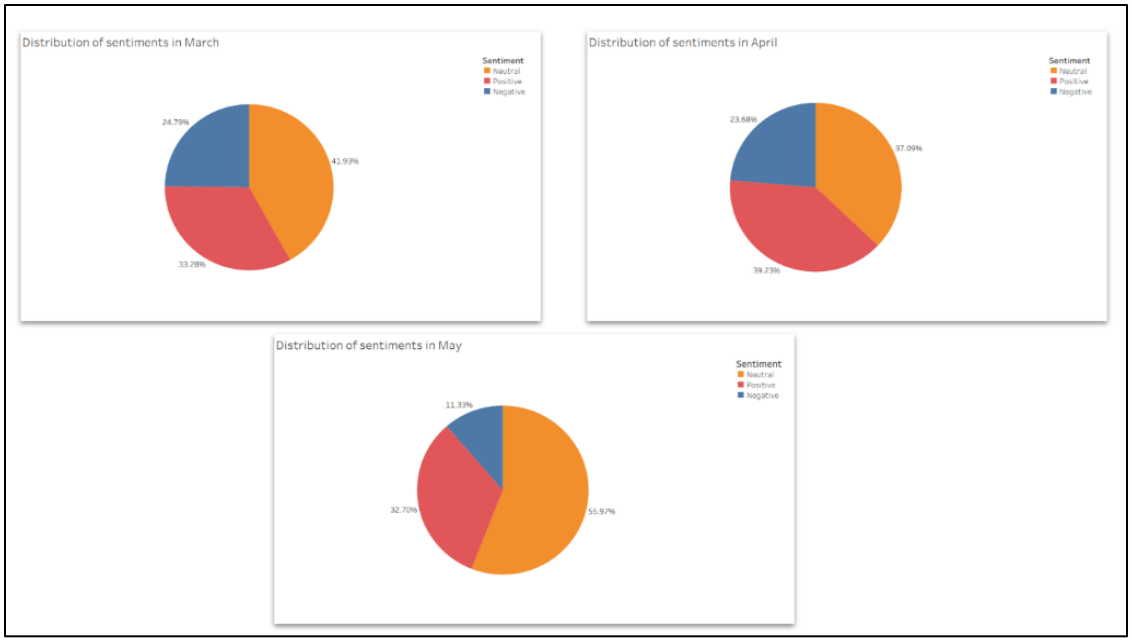


Fig 2

Data Preparation and Cleaning

Once we gathered the tweets, we needed for sentiment analysis, we moved ahead with preparing the data. Social media data are not structured. In other words, it's raw, noisy and needs to be cleaned before we can begin to use it for modelling. This is a significant step because the quality of the data will yield in more reliable results.

Preprocessing of the dataset involved a series of tasks such as removing all types of irrelevant information like emojis, special characters, and extra blank spaces. We also made format improvements, deleted duplicate tweets, or tweets that were shorter than three characters.

Fig 3 shows the raw form of the tweet text and the cleaned tweet.

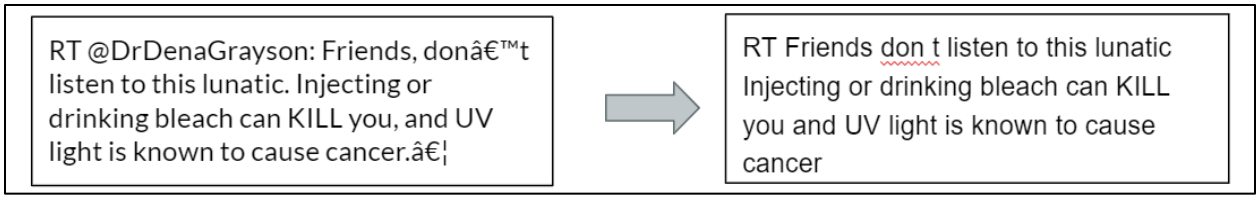
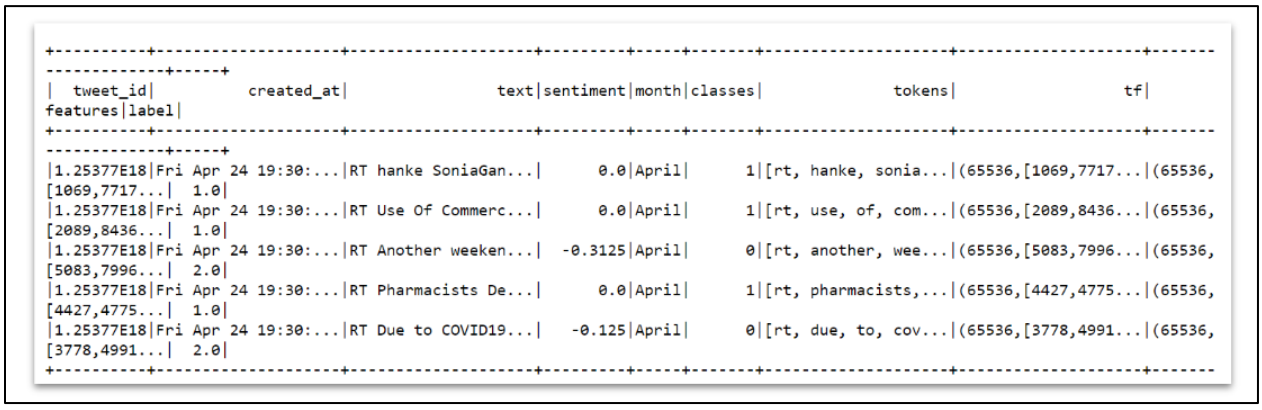


Fig 3

We utilized Tfidf Vectorizer to convert textual data of the tweet to numeric form, so that it can be further used in the modelling process**.** Fig 4 shows the head of the data after it was processed through TFIDF Vectorizer.



```
+----------+--------------------+--------------------+---------+-----+-------+--------------------+--------------------+-------
------------+-----+
|  tweet_id|          created_at|                text|sentiment|month|classes|              tokens|                  tf|
features|label|
+----------+--------------------+--------------------+---------+-----+-------+--------------------+--------------------+-------
------------+-----+
|1.25377E18|Fri Apr 24 19:30:...|RT hanke SoniaGan...|      0.0|April|      1|[rt, hanke, sonia...|(65536,[1069,7717...|(65536,
[1069,7717...|  1.0|
|1.25377E18|Fri Apr 24 19:30:...|RT Use Of Commerc...|      0.0|April|      1|[rt, use, of, com...|(65536,[2089,8436...|(65536,
[2089,8436...|  1.0|
|1.25377E18|Fri Apr 24 19:30:...|RT Another weeken...|  -0.3125|April|      0|[rt, another, wee...|(65536,[5083,7996...|(65536,
[5083,7996...|  2.0|
|1.25377E18|Fri Apr 24 19:30:...|RT Pharmacists De...|      0.0|April|      1|[rt, pharmacists,...|(65536,[4427,4775...|(65536,
[4427,4775...|  1.0|
|1.25377E18|Fri Apr 24 19:30:...|RT Due to COVID19...|   -0.125|April|      0|[rt, due, to, cov...|(65536,[3778,4991...|(65536,
[3778,4991...|  2.0|
+----------+--------------------+--------------------+---------+-----+-------+--------------------+--------------------+-------
```

Fig 4

Data Analysis

In data analysis, we get into the modeling process and the process of building a classifier to predict whether the tweet is negative, neutral, or positive. We decided to run our tweets through three popular algorithms used for classification: Logistic Regression, Decision Trees and Random Forests.

We first started our implementation by creating a cluster on AWS EMR with 3 instances. This is to assess the performance of our models using parallel processing as compared to using a single processor. We then fetched data from an Amazon S3 bucket. We had stored March, April and May

tweets on it and uploaded these datasets onto a notebook instance running on our cluster. We followed the preprocessing steps explained in the previous sections and after we have our tf-idf vectors, we are ready to run this through our models.

One important performance measure is to see the time vs scale and accuracy vs scale comparisons. To do this, we split our data into training and test data four times each time with a different train size. We split the training data into 30%, 50%, 70% and 90% of the data and the rest was considered test data. We imported the pyspark modules of our three algorithms and ran our training datasets with its features (tf-idf vectors) with the labels (0, 1, 2 classes) on these models. We noted down the accuracies and time taken to run these steps. We terminated the cluster and repeated the same steps over by creating a new cluster with just one instance and hence no parallel processing.

Results

In our results, we finally arrive at certain conclusions from what we observed through the entire process of data mining, data preparation, data analysis and model implementation.

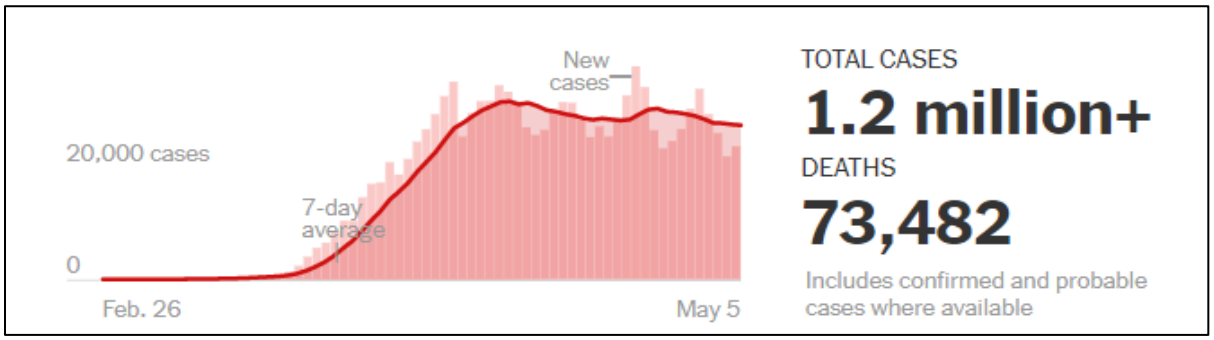


Fig 5

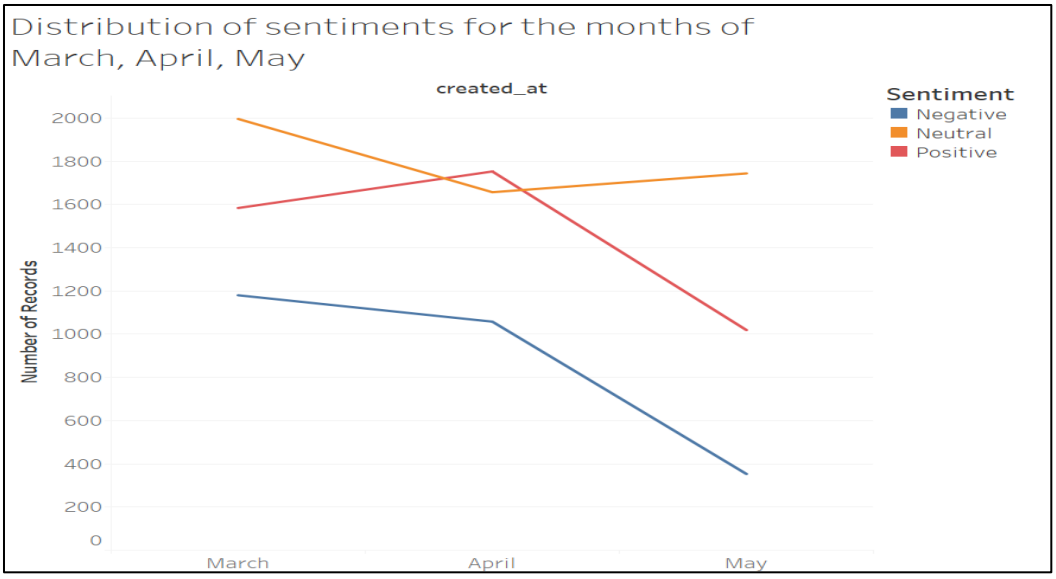
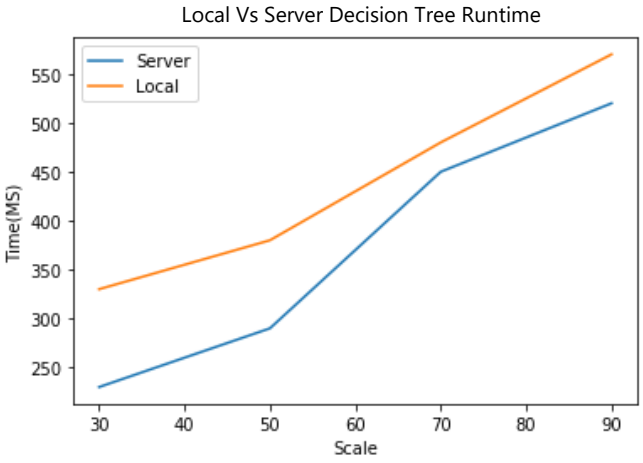


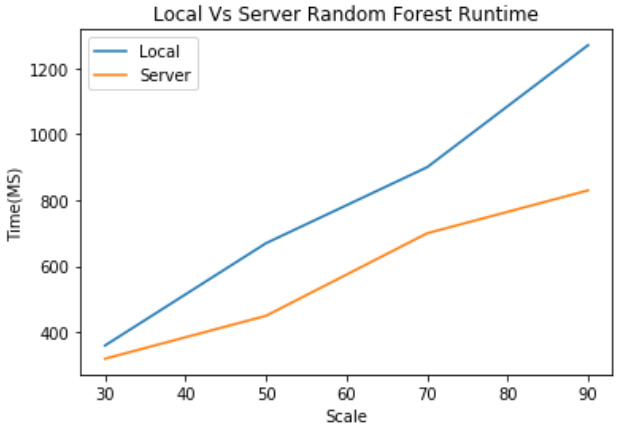
Fig 6

In the Fig 5 above, we can see the graph of the number of cases per day in the whole United States. We see a peak reaching in mid-April and a decrease in the number of cases as we approach May. Just like we explained in our introduction, our hypothesis stated that as we see the situation getting better and the number of cases decrease, we would see a decrease in negative sentiments. According to the results published, it is evident that this indeed is the case. The Fig 6 shows how number of negative tweets has decreased, the number of positive tweets has also decreased, and we see an increase in neutral sentiments among the masses. This means that as the pandemic hit its peak and is now on the verge of decline, people are tweeting more about news, facts, government guidelines and advisories rather than resorting to fear mongering or spread of false hope and false news. This proves our theory of people starting to understand the pandemic, the guidelines, and the virus itself better with time.

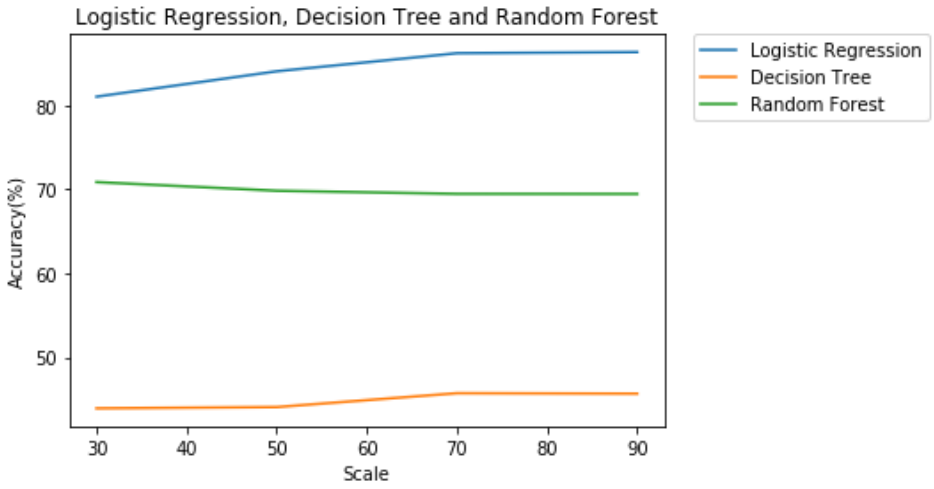
Performance Evaluation

To evaluate the performance of our algorithms against time and scale when we use both 3 instances as well as one instance, we generated graphs. We can see in the graphs below that using parallel processing or a server is much faster compared to local computation using only one instance.





The graph below represents the accuracy vs scale for all three of our algorithms. We can see that Logistic > Random Forest > Decision Trees. This is the case because the performance of Random Forest degrades when the number of features is less. Here the text of the tweet is our only feature! Logistic regression on the other hand is known to perform better with lower dimensional data. Also, random forests are an ensemble of weak decision trees. Hence together they perform better than just a single decision tree.



Next we also evaluated the time vs scale graph for each of our algorithms to see which performs the fastest and the slowest and as expected saw that decision trees are much faster than random forests and logistic regression as the model expects a traversal through one tree to get the output

whereas a random forest is an ensemble of trees considering the majority of the output from all the trees.

Conclusion

In conclusion, we learned the following:

1. Sentiment Analysis of COVID-19 related tweets gave us an insight into the minds of people as we deal with this pandemic worsening, reaching the peak and going back into normal.
2. Logistic Regression gave us the highest accuracy of 86.18%
3. A comparison of local vs server performances was made and observed accurately that running parallel is faster
4. In conclusion, as this situation is getting better, people are starting to trust facts rather than negatively being impacted or wrongly believing in an “all is well” agenda. We see this clearly in the sentiments hidden in the tweets.

References

1. How to collect tweets from the Twitter Streaming API using Python
2. How to Extract Tweets from Twitter in Python
3. Sentiment Analysis with PySpark
4. Machine Learning with PySpark and MLlib — Solving a Binary Classification Problem
5. USC Researchers Analyze Coronavirus Misinformation on Twitter
6. Sentiment Analysis: Definition, Uses, Examples + Pros /Cons
7. Corona Virus (COVID-19) Tweets Dataset