ABV INDIAN INSTITUTE OF INFORMATION TECHNOLOGY AND MANAGEMENT GWALIOR

A project report on

Sentiment Analysis of COVID-19 related tweets

Submitted in partial fulfillment of the requirements for the minor project of Big Data Analytics course

IPG-MTech (2017-2022)

Submitted to:

Dr. Anuradha Singh

Submitted By:

Pranay Kumar (2017IMT-028)
Danthala Deepak (2017IMT-029)
Karri Srinivas Rao (2017IMT-046)
Kavila Bhushanam (2017IMT-047)
Suggula Jagadeesh (2017IMT-082)
Dinesh Verra (2017IMT-089)
Venkat Sai Swarup (2017IMT-109)
Vishwajeet Kumar (2017IMT-111)

Chapter 1

Introduction:

The Corona Virus Disease (COVID-19), caused by a new coronavirus with higher reproductivity than SARS, first emerged in the People's Republic of China in December 2019. COVID-19 has created a large number of unknowns in the world. These unknown events occurred in small events of time that had led to an exponential rate of increment in cases, human-to-human transmission is also one of the reasons. It has brought serious challenges in the whole world. Soon momentary damage caused by COVID-19 spread to members of the medical staff organized in support of COVID-19 regulation and the public. Social media have been identified as a crucial medium for the public to collect knowledge and social learning to manage uncertainties and threats in the context of a public health crisis.

Origin of the proposal:

Since the pandemic began in early 2020, people took it to social media platforms like Twitter, Facebook, etc., to express their feelings about the situation. This project hopes to gain insights from the data collected by scraping these sites. Specifically, we try to explore the "Tweet data", analyze it, understand it, and finally train a model that can predict how the user must have felt while he was tweeting it.

The main goal of our project is to perform Sentiment Analysis on the "Tweet data" related to COVID-19. This helps us in monitoring the mental health of the people without having to directly meet them. We can then take the necessary measures to keep people happy and healthy.

Chapter 2

Review of the status of Research & Development in the subject

2.1 International Status:

COVID19 was a massive challenge for the whole world. During early 2020, insightful information on appropriate public health responses caused during the outbreak is used for public sentiment analysis.

One of the popular social media platforms of China called Sina Weibo, was posted with negative sentiments which were valuable for the analysis of public concerns. During the spring 2020, over a nine lakh randomly selected posts from Sina Weibo were analyzed[2].

The unsupervised model called BERT (Bidirectional Encoder Representations from Transformers) was adopted to organize sentiment categories (positive, neutral, and negative) and another model called TF-IDF (term frequency-inverse document frequency) was used to encapsulate the topics related to the posts. The two different analyses called Thematic analysis and Trend analysis were conducted to identify the characteristics of negative sentiment. So to summarize, the finetuned model BERT conducts sentiment analysis classification with extensive accuracy.

It was observed that the results obtained from Sina Weibo provided particular instructions on the health responses of each individual which might help in elevating public concerns through scientific guidance and information sharing.

2.2 National Status:

During the early 2020 there was a high increment in the data developed through online, and due to this many fields have attracted a large number of researchers. One of the fields was interested in the study of social media data. Later this year, sentiment analysis was continuing to grow at a rapid pace.

Using many machine learning algorithms, Jain and Dandannavar examined twitter sentiment analysis and listed a detailed approach.Mostly, the data is collected and then preprossed using the algorithms extracted from NLP. Sentiment analysis is done using decision tree models and multinomial naive bayes in the proposed method. It was observed that the best results were obtained with the use of a decision tree, with 100% accuracy rate.

Rajput et al. presented a statistical analysis of the twitter messages related to Coronavirus posted since January 2020. Two types of empirical studies were performed. The first one was on the word frequency and the second one was on sentiments of the individual tweet messages. Unigram, bigram, and trigram frequencies were modeled by a power-law distribution. Different metrics were used to validate the results, like the R2 score, the Sum of Square Error (SSE) and the Root Mean Square Error (RMSE). The results gave low SSE and RMSE along with a pretty good R2 score which proved that the model was good. According to the results, only 15% of the tweets were negative.

Barkur et al., analyzed sentiments of Indians which dealt with the lockdown announcements. The hashtags #IndiafightsCorona and #IndiaLockdown was extracted by the author from the Twitter tweet data during the latter month of march. Over a 24000 tweets were considered for the analysis using R statistical software. It was observed that we Indians took the battle against COVID19 positively and the majority of the people were in the favour of the Indian government regarding the decision of lockdown which was observed by flattening the curve.

2.3 Importance of the proposed project in the context of the current status:

In the current situation, where people around the world are battling for their physical health, it is important to know that the mental health of the people is important as well. People may fall into depression due to the current situation where they might be facing financial problems or the loss of a loved one. We need people to be mentally strong and help each other as much as possible to live through the pandemic. With that being said, it is impossible with our resources to keep track of the mental well-being of every person. Our project aims at drawing out the list of people who seem to be unhappy, by analyzing and predicting their sentiment from the tweets they have posted in the past few months. We can then take appropriate measures to help them and keep their mental health in check.

Although many projects and researchers are undergoing to find a vaccine for COVID-19, there's still time for the pandemic to be over. As many projects focus on the physical health of the people, the mental health of the people is being ignored. But, it must be noted that both physical and mental health is important for a person. So, we decided to focus on a project such that it helps in keeping the mental well-being of people in check indirectly.

Chapter 3

WORKPLAN

3.1 Methodology

3.1.1 Dataset Exploration:

The project uses the Twitter dataset[1] extracted from the IEEE dataport. From March 11, 2020 (pandemic outbreak), it was observed that there was a high rise in the contents related to the pandemic outbreak in the platforms like facebook, instagram, twitter and other social media. Twitter is observed to be indispensable in the extraction of situational awareness information relating to any crisis. The dataset contains over six lakh tweets that were posted by the people of India during the lockdown. The dataset contains five columns namely

1. Text ID - A unique identifier for each tweet.
2. Text - The tweet text corresponding to a particular tweet ID.
3. Date - The date on which the tweet was posted.
4. Location - The place where the tweet was posted.
5. Sentiments - The sentiment score for each tweet.

3.1.2 Data Cleaning:

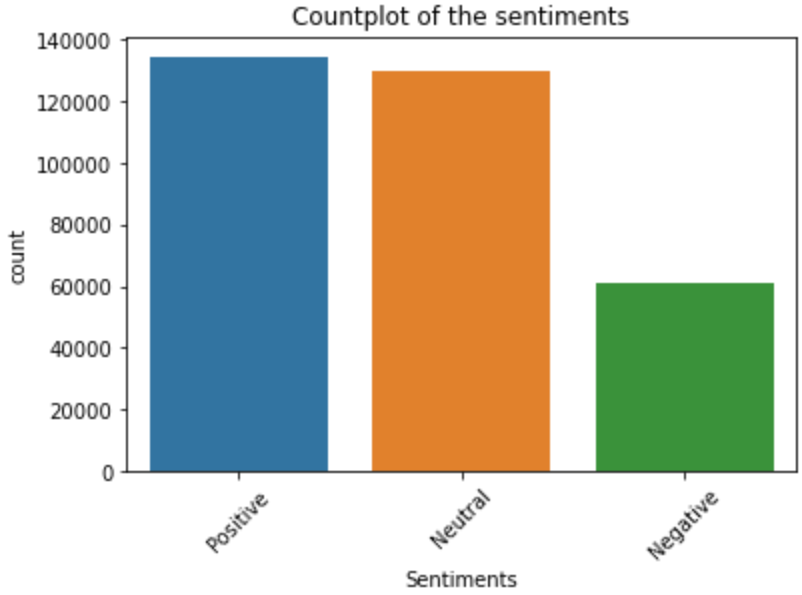
The dataset contains 648957 tweets posted by the people of India during the lockdown. It was observed that there are some duplicate tweets in the dataset. So to remove data inconsistency, we dropped all the duplicate tweets from the dataset and the total tweets were dropped down to 324960 posts. It was found that some tweets contain null values (missing values) in the dataset. So we replaced all the missing values with their mean values.

3.1.3 Exploratory Data Analysis:

To dig deeper and understand more about the dataset we performed initial investigations on the dataset to discover patterns, test hypotheses, and check assumptions using graphical representations. The different plots plotted using the dataset are

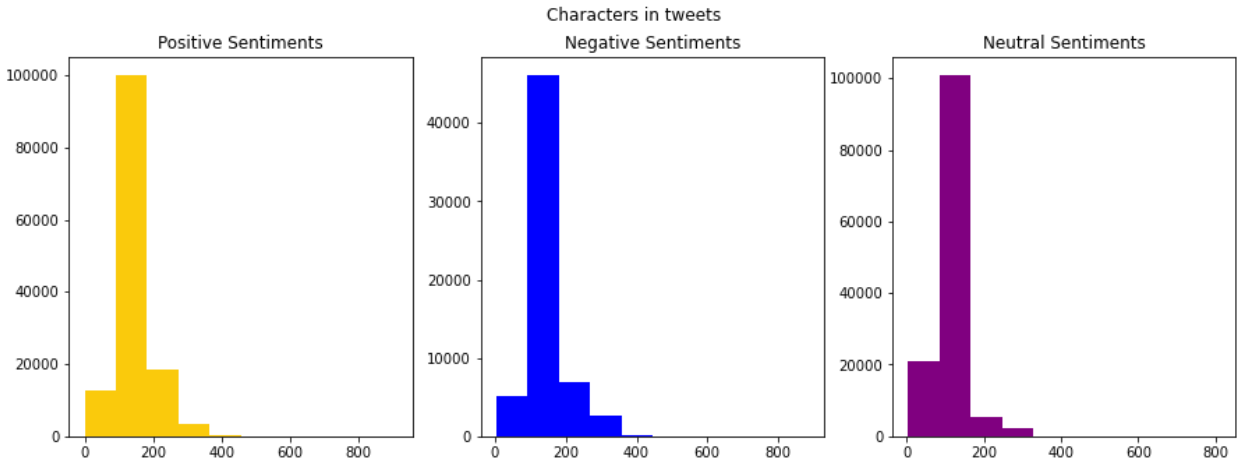
- Count Plot of the sentiment classes

  Here we plotted all the sentiment values (Positive [sentiment score >0], Negative [sentiment score <0] and Neutral [sentiment score = 0]) on the histograms categorised with the sentiments. The plot represents the count frequency of each sentiment. It was observed that there are more positive sentiments in the dataset followed by negative and then neutral.



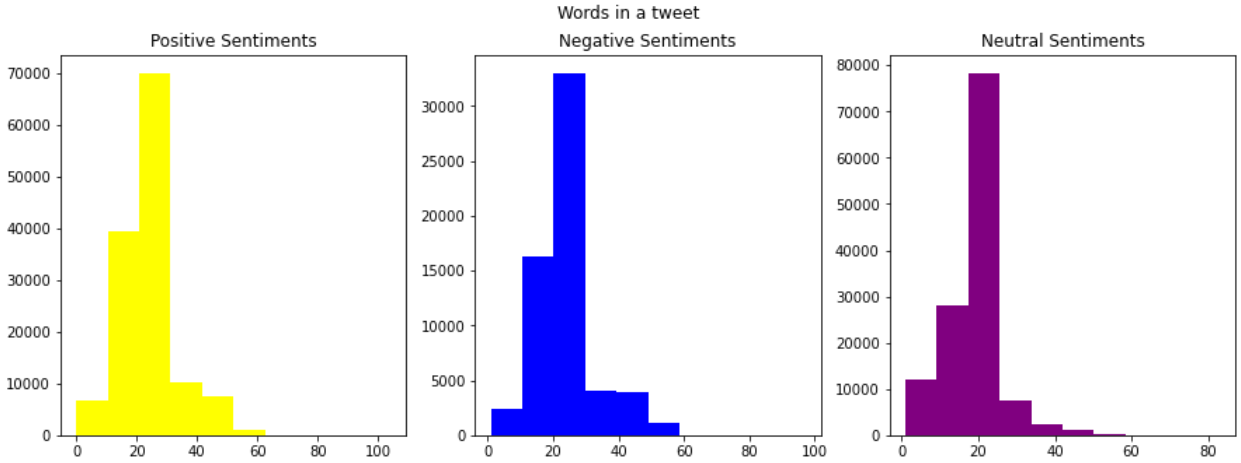
- Number of characters in a tweet per sentiment

  Here we plotted three graphs (Positive, Negative, and Neutral) of sentiments each referring to the number of characters in each tweet corresponding to its text ID in referral to the sentiment score. The plot represents the character frequency of each sentiment. It was observed that the character's distribution for positive and neutral sentiments are almost the same.



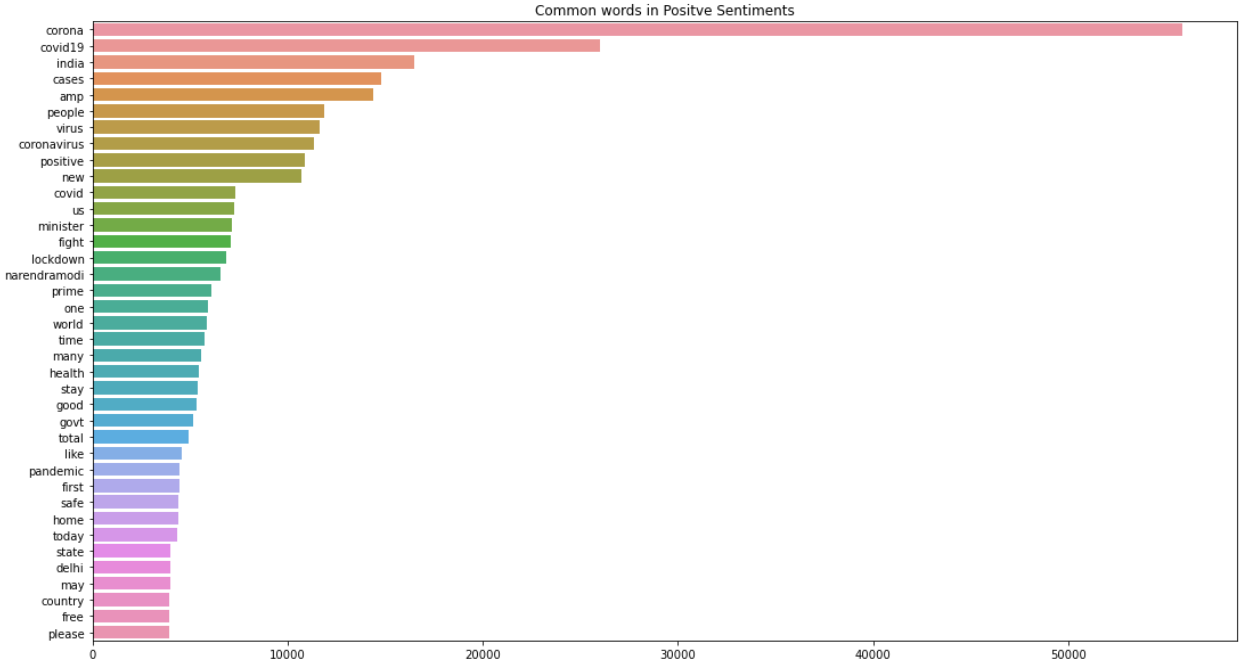
- Number of words per tweet

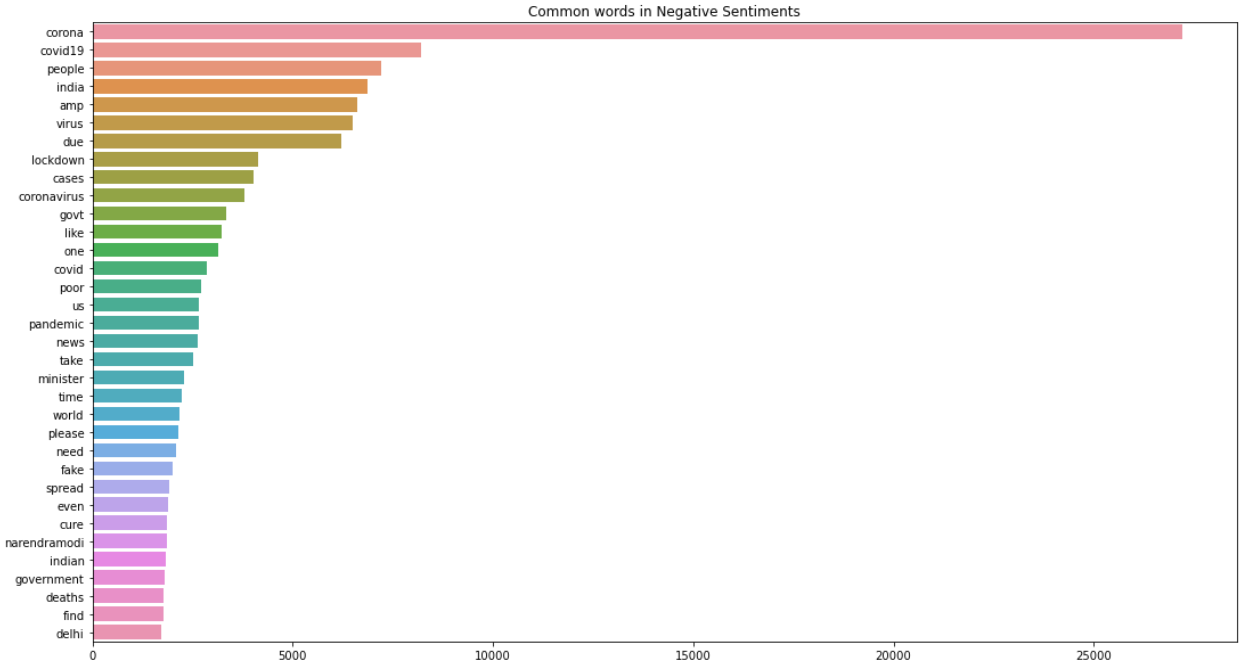
  Here we plotted three graphs (Positive, Negative, and Neutral) of sentiments each referring to the number of words in each tweet corresponding to its text ID in referral to the sentiment score. The plot represents the words frequency of each sentiment. It was observed that the words distribution for positive, negative and neutral sentiments are almost the same.

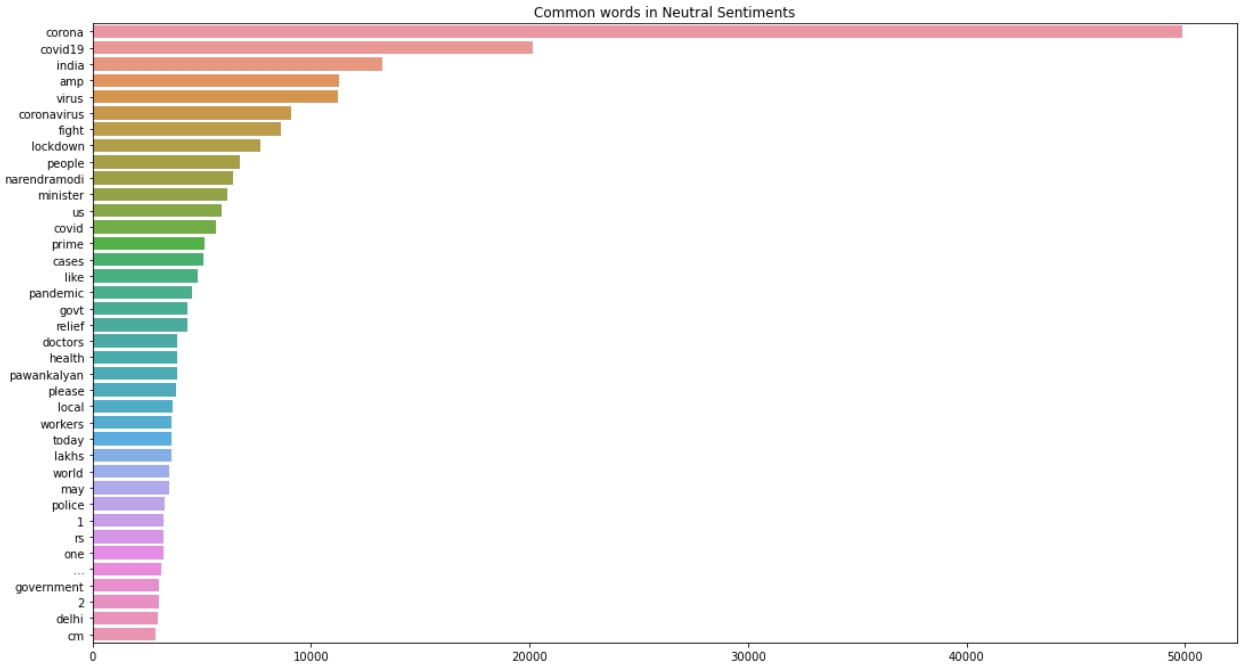


- Common words

  Here we plotted three graphs (Positive, Negative, and Neutral) of sentiments each referring to the common words used in each sentiment class. The plot represents all the common words used in each sentiment which are plotted against their frequencies. It was observed that the word “corona” was most used in all the sentiments.

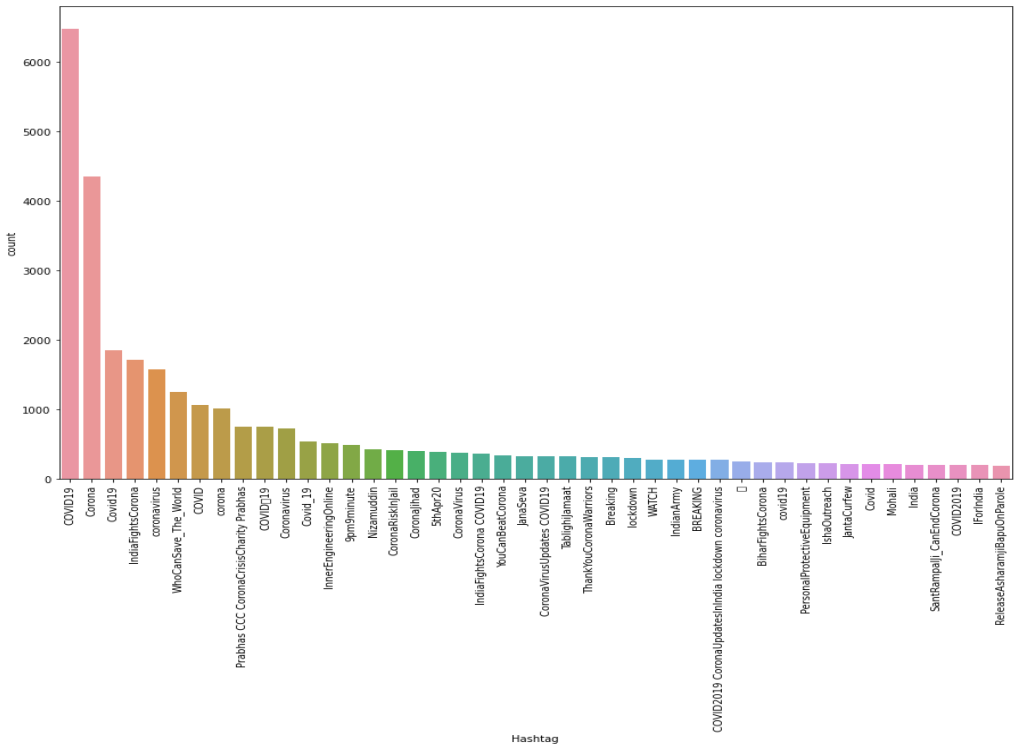

Common words in Positve Sentiments
corona
covid19
india
cases
amp
people
virus
coronavirus
positive
new
covid
us
minister
fight
lockdown
narendramodi
prime
one
world
time
many
health
stay
good
govt
total
like
pandemic
first
safe
home
today
state
delhi
may
country
free
please
0
10000
20000
30000
40000
50000


Common words in Negative Sentiments
corona
covid19
people
india
amp
virus
due
lockdown
cases
coronavirus
govt
like
one
covid
poor
us
pandemic
news
take
minister
time
world
please
need
fake
spread
even
cure
narendramodi
indian
government
deaths
find
delhi
0
5000
10000
15000
20000
25000



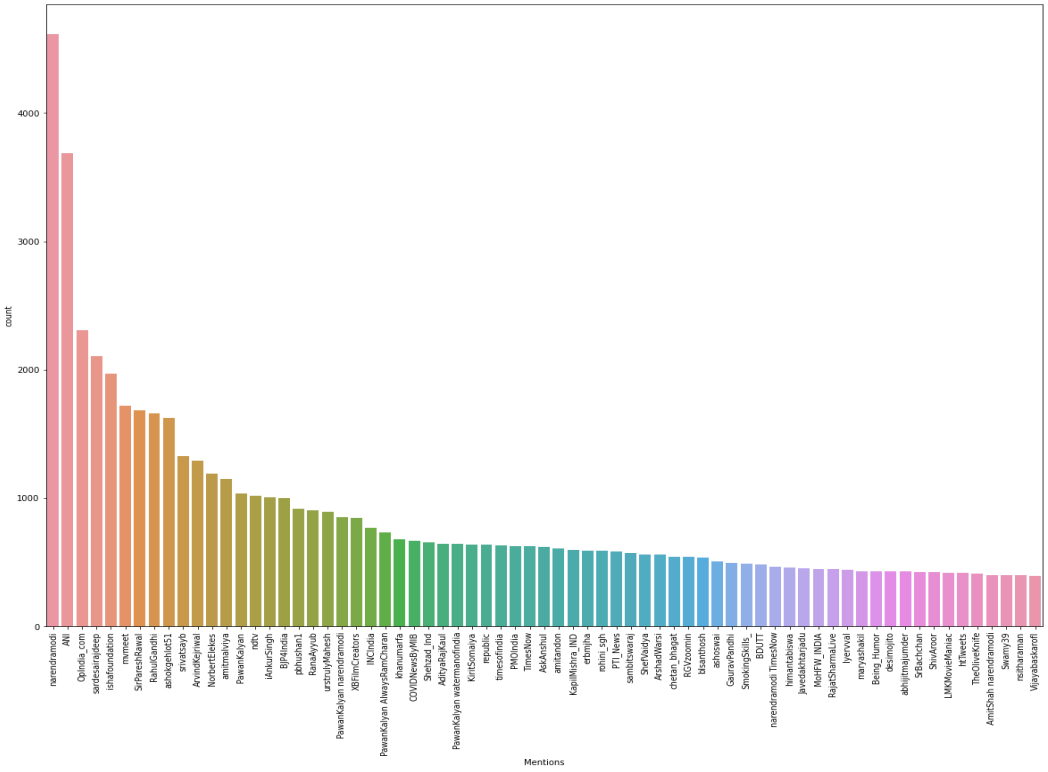
- Most popular hashtags

  Here we plotted a graph referring to the most popular hashtags used in the dataset for a specific period. The plot represents all the popular hashtags for all the sentiments which are plotted against their frequencies. It was observed that the hashtag “#COVID19” was most popular.



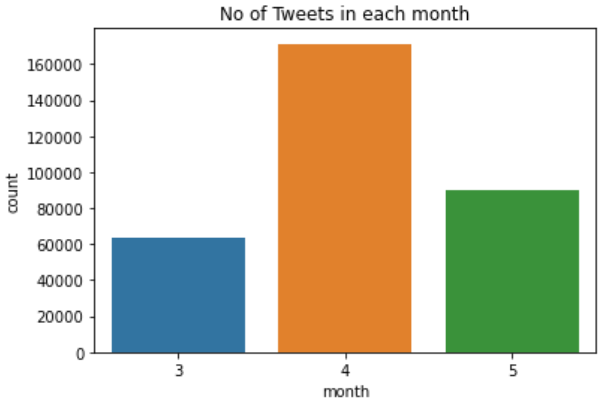
- Most popular mentions

  Here we plotted a graph referring to the most popular mentions used in the dataset for a specific period. The plot represents all the popular mentions for all the sentiments which are plotted against their frequencies. It was observed that “@narendramodi” was the most popular mention.



- No of tweets in each month

  Here we plotted a graph referring to the number of tweets posted in each month. The plot represents the month plotted against the number of tweets posted. It was observed that the highest tweets were recorded during the 4th month - April.



3.1.4 Data Preprocessing:

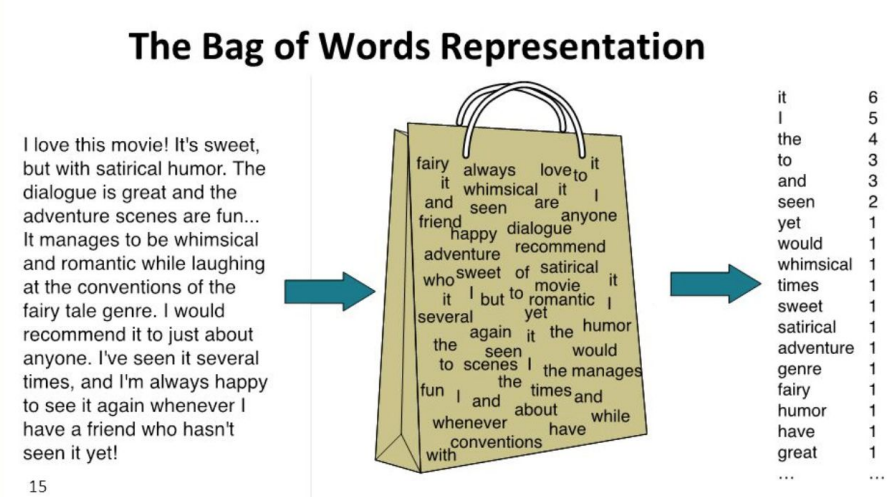
The data is totally in raw format. It contains URLs, HTML Tags, Hashtags, Punctuations, stopwords, etc, which are not useful while classifying the tweets. Rather they will increase the confusion for models. Even the machines can't understand text, It should be converted in a format that machines can understand.

Below are the steps used for data pre-processing in our project:

- Removing unnecessary text like URL's, hashtags, stopwords (like a, the, etc). Users also use shortcut words(like hru, gn, etc). These words have been replaced with their abbreviations.

- Lemmatization: Lemmatization is a process of capturing the various forms of a word together so that they'll be viewed as a single object. Lemmatization is similar to stemming, but it adds more factors to the words. It blends words of the same meaning to a single phrase.

- Bag-of-Words: Bag-of-words is a tool used to represent text in numbers. This model is used to pre-process the text by translating it into a word pack that maintains a count of the cumulative occurrences of the most commonly used words.



- TF-IDF vectorizer: TF-IDF stands for Term Frequency-Inverse Document Frequency. It is numerical statistics that are intended to reflect how important a word is to a document. It is also used for information retrieval.
  The idea behind it is that if a word appears several times in a text, its significance should be increased as it should be more relevant than other terms that appear less times (TF). At the same time, if a word appears several times in a text, but also along with many other documents, it could be because that word is just a frequent word; not that it was important or significant (IDF). Below is one example of a TFIDF vectorizer.

  Tweet 1: I am not feeling well.
  Tweet 2: I am good.
  Tweet 3: I am feeling back pain.

  Bag of Words Vector of tweet 1: [1,1,1,1,1,0,0,0]
  Bag of Words Vector of tweet 2: [1,1,0,0,0,1,0,0]
  Bag of Words Vector of tweet 3: [1,1,0,1,0,0,1,1]

  Term frequency is a measure of how frequently a term, “T” appears in the document, “D”.
  TF for the word “I” in tweet 2 = (no of times “I” appears in tweet 2)/(no of terms in tweet2).

  TF(“I”) = ⅓, TF(“am”) = ⅓, TF(“good”) = ⅓.

  Similarly, we can compute TF for other tweets.

  Inverse Document Frequency(IDF) is a measure of how important a term is. It is needed to calculate because computing TF alone is not sufficient to understand the importance of words.
  IDF for the word “I” in tweet 2 = log(no of documents/no of documents with term “I”).
  IDF(“I”) = 0.00, IDF(“am”) = 0.00, IDF(“good”) = 0.48.
  Similarly, we can compute IDF for other tweets.

  Then we will compute the TF-IDF score for each word in the vocabulary list. Words with higher scores are more important as compared to words with less score. Computing TF-IDF score for words in tweet 2:

  TF-IDF(“I”)= TF(“I”) * IDF(“I”)
  TF-IDF(“I”) = 0, TF-IDF(“am”) = 0, TF-IDF(“good”) = 0.16.

  Similarly, we can compute TF-IDF for other tweets.

3.1.5 Sentimental Analysis Modelling :

Multinomial Naive Bayes:

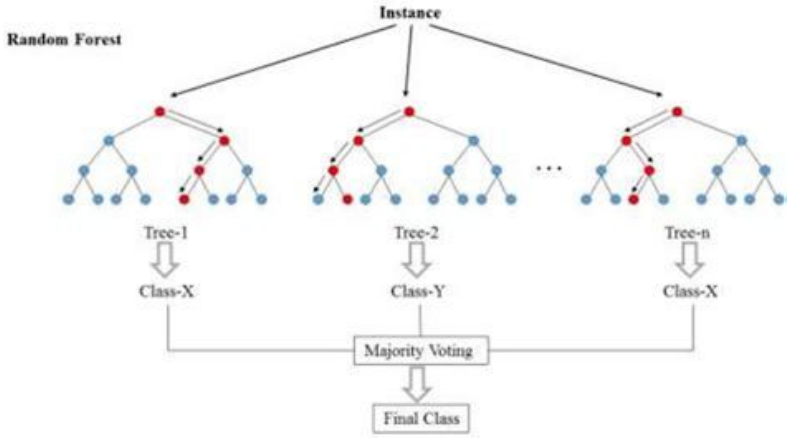
Naive Bayes works on the basis of Bayes' theorem. This algorithm treats that features are independent. The probability of occurrence of one feature doesn't get affected by the probability of occurrence of the other feature. It is useful for small datasets. Since it is a direct application of Bayes's theorem it is easy to implement.

The multivariate model of Naive Bayes is Multinomial Naive Bayes. It is mainly used for text documents.A simple naive Bayes theorem classifies a document as absence and presence of words, whereas the multinomial model computes the number of words and adjusts the computations accordingly.

Random Forest:

Random Forest is an ensemble algorithm which comes under the category of bagging. It uses a technique called bootstrapping, which trains multiple decision trees on various random subsets of features of the given dataset and takes the average of output of all trees to improve the predictive accuracy. More number of trees in the forest will help in achieving a good evaluation score and avoid the problem of overfitting.

Random forest also helps in identifying the best feature based on the relative importance. It is the most frequently used algorithm. It also gives competitive results when compared with neural networks. It is used for both regression and classification.

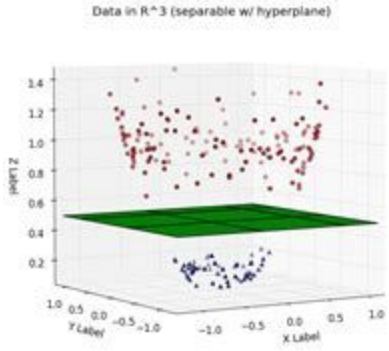


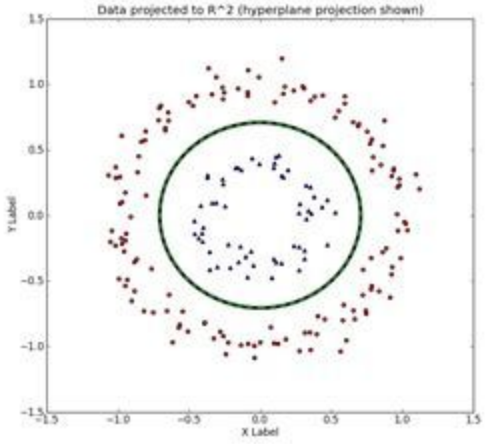
Support Vector Machine:

Support vector machines are algorithms that are supervised in nature. It creates the best boundary or plane that can categorize higher dimensional data so that we can classify a testing point accurately. This critical plane which is categorizing data is called a hyperplane. Support vectors are the critical points that lie very closely on either side of the hyperplane. Hyperplanes can be curvilinear also.

SVM are of two types. Linear SVM and Non - Linear SVM. Linear SVM means the data can be linearly separable. Nonlinear SVM is used for non-linear data that is not linearly separable. In nonlinear SVM we use kernel functions to convert them to linear separable. For example, polynomial, radial basis function (RBF), and sigmoid.

The RBF kernel is general -purpose kernel used to solve the problem of classifying datasets that cannot be separated linearly. Polynomial kernels are suitable for solving classification problems on normalized training datasets. Training an SVM with a Linear Kernel is simpler and faster than with any other Kernel.But it can be used if the dataset is linearly separable. In our project, we are using RBF , which is a general-purpose based kernel.

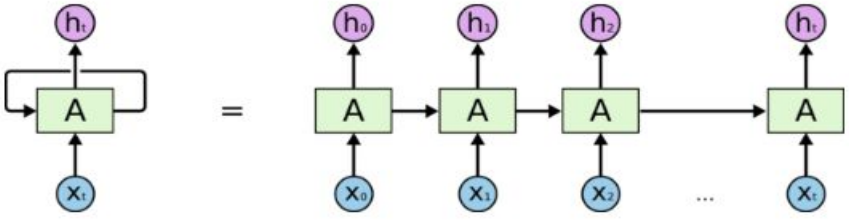




Recurrent Neural Networks (RNN):

Neural Networks that have internal memory and are a general form of a feedforward neural network are called Recurrent Neural Networks.RNN performs the same function for all inputs and nature of RNN is recursive. The output of an input which is being executed depends on the previous output computed. The output of every input is stored in its internal memory. Output of every input recurrently depends on the previous output.

The important state of RNN is the Hidden state It remembers a sequence for every input. A very long sequence cannot be processed. Like every neural network as we know the error sum is obtained i.e. target output - predicted output.The error is checked with the target error if it is greater than it is back propagated and weights are reseted accordingly and in a similar way RNN is also trained.An important observation is that inputs in a neural networks do not depend on each other.Whereas in RNN inputs are mutually correlated.



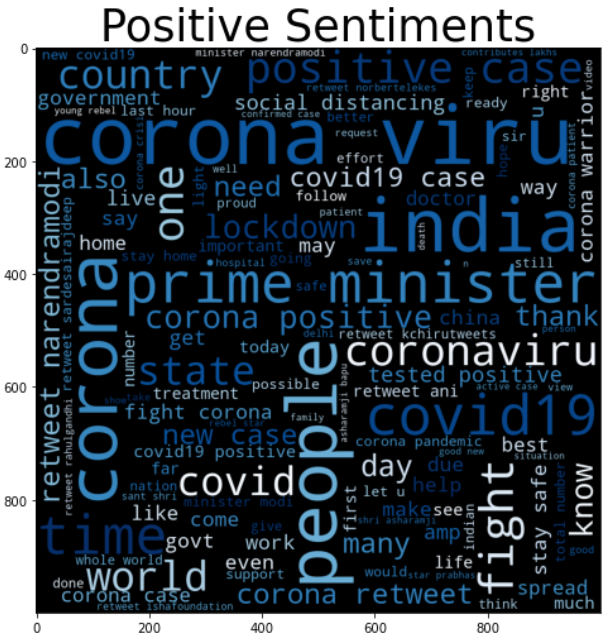
An unrolled recurrent neural network.

3.2 Results

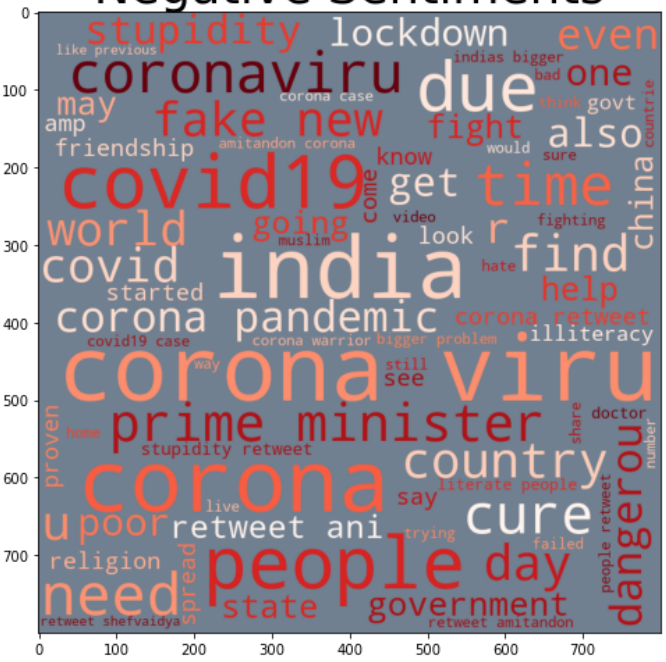
The data has been split into train and test in the ratio of 80% and 20%. We have trained different algorithms for classifying the sentiments on the training data. Accuracy has been used as the evaluation metric. Below are the accuracy scores obtained on the test data.

- Multinomial Naive Bayes: 78.83%
- Random Forest: 90.30%
- Support Vector Machine: 86.5%
- Recurrent Neural Networks: 93.2%

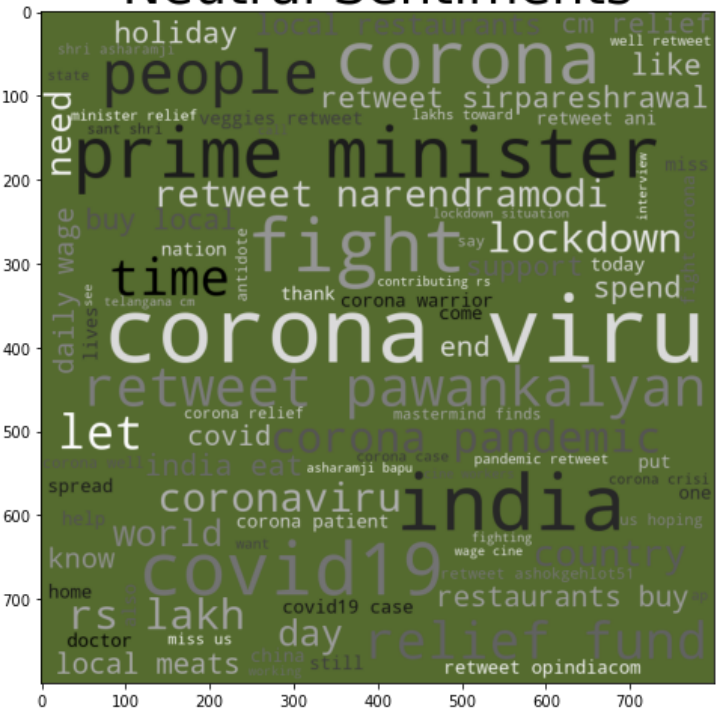
The RNN model has outperformed all other models. Below figures are the word clouds per sentiment based on our analysis of the data.



Negative Sentiments



Neutral Sentiments



3.3 Conclusion

During the course of the project, we have successfully explored the data, cleaned it, processed it, analyzed it, understood it, and finally trained a classifier to predict the sentiment based on the available features extracted from the dataset. We have used different approaches and algorithms to train the model and ultimately, chose the one with the best accuracy.

The project achieves its goal of classifying the tweets based on the sentiment of the user, as intended. This predicted data can be used in monitoring the mental health of the people and provide them with assistance if and when needed. Working on this project not only gave us an insight into the current situation caused due to the pandemic but also strengthened our Data analysis skills.

We can improve the performance of the existing models by using word embeddings and advanced models like Transformers and BERT. To train these advanced models, we require a huge amount of computational resources.