**“Mental Health Of India During COVID**”

Covid-19 is an infectious disease caused by newly discovered coronavirus which has brought the whole world to halt. At this time, there are no specific vaccines or treatment for COVID-19. In such a situation, the best way to prevent and slow down transmission is to be well informed about the COVID-19.

**Problem Description:-**

This Covid-19 pandemic has severely affected countries around the world. The intensity of the pandemic is increasing very fast in India. The number of new cases is increasing every day, every week. In a span of six months, the total number of cases crossed 50 lakh and total number of deaths is almost I lakh. It has been observed that the sudden outbreaks of such pandemics affect public mental states and emotions. This pandemic also results in either constructive or destructive behavioural changes among people. Anger, Sadness, fear are the most common emotions witnessed among the people during several pandemics. Social media platforms like Twitter and others have rich sources of information from people.

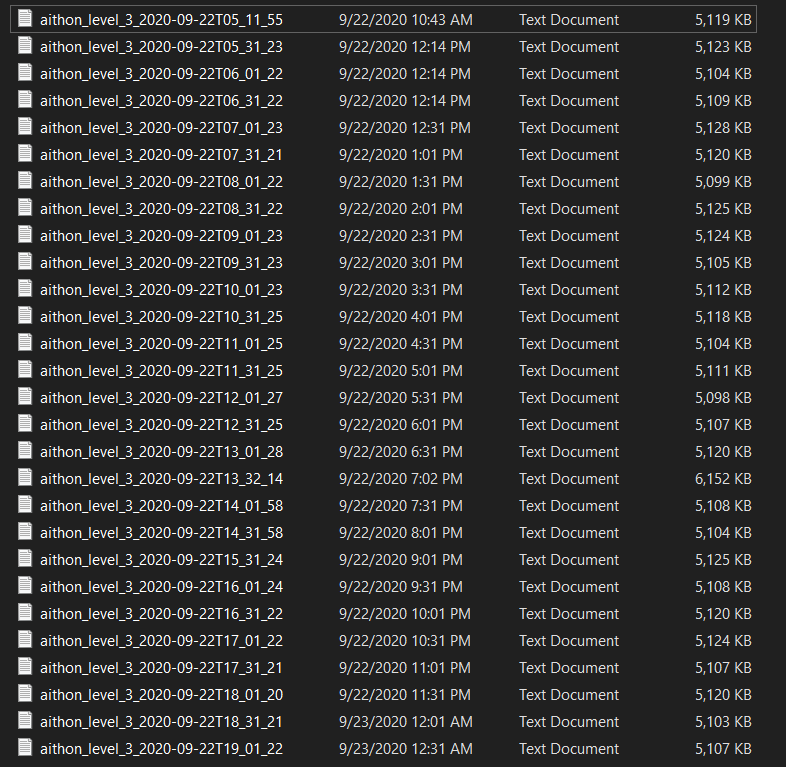
Here, we are to study given twitter data to understand the emotions of people against Covid-19.

Finally, a detailed exploratory analysis on emotions analysis based on a given dataset.

**Introduction**

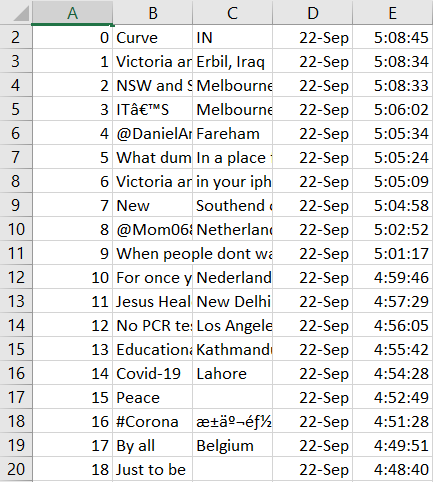
**Dataset Provided →**

We were given a dataset in the form of 28 text documents. Each of them had twitter data in Json format. There were a total of 4,96,448 tweets which were collected from a 10 day period from 13th to 22nd september. The json data included the tweet text, location, date and time.



**Data Pre-Processing →**

The first task we had was to convert our data into a format we can understand. So we used a python library called pandas and converted our json file into a csv file.

In general a tweet needs to be pre-processed, because it consists of @mentions, #hashtags, URLs, emojis, lemmatized words, etc. Our preprocessing steps included:-

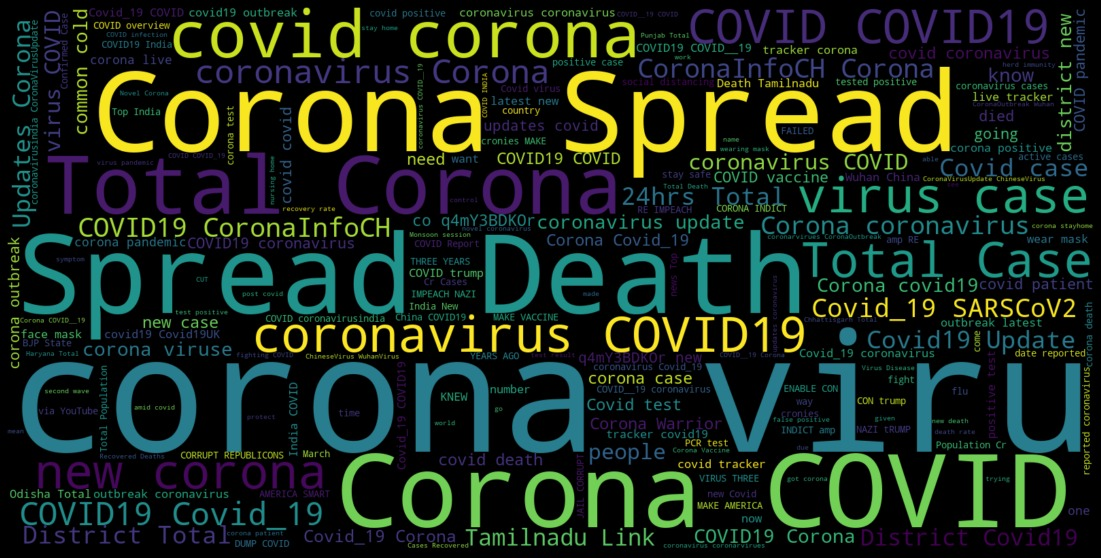
1. Removing mentions, hashtags, URLs
2. Converting emojis to words
3. Reverting lemmatization
4. Reduced lengthening

**Data Filtering →**

A tweet acquired after data processing still has a portion of raw information in it which we may or may not find useful for our application. Thus, these tweets are further filtered by removing stop words, numbers and punctuations.

Stop words: For example, tweets contain stop words which are extremely common words like “is”, “am”, “are” and hold no additional information. These words serve no purpose and this feature is implemented using a list stored in stopfile.dat. We then compare each word in a tweet with this list and delete the words matching the stop list.

**Feature Extraction →**

**Fig. 1** *WordCloud generated from the given dataset*

TF-IDF is a feature vectorization method used in text mining to find the importance of a term to a document in the corpus. Feature extraction involves the “mlib” library of Apache Spark. The recommended API is the Data Frame based API. This feature is useful for a case where we need to find trending topics or to create word clouds. However, this project is more focused towards finding sentiment in twitter streams so TF-IDF is not implemented.

**Sentiment Analysis →**

Sentiment analysis, also known as opinion mining, is a sub machine learning task where we want to determine what is the general sentiment of a given text. Using machine learning techniques and natural language processing we can extract the subjective information of given text and try to classify it under the labels on which we trained our model.

**Model Description →**

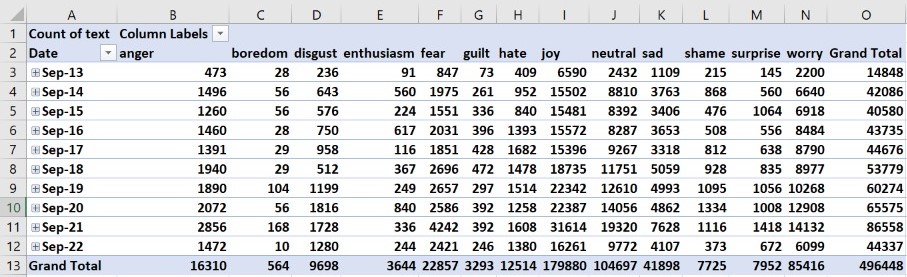
We used a machine learning model for classifying the given data into 13 emotions. These emotions are joy, neutral, worry, sad, fear, anger, hate, disgust, surprise, shame, enthusiasm, guilt and boredom.

First we used count vectorisation to convert our tweet text data into arrays of 1s and 0s. This makes our data easy to process.

For a model we need training data, so we downloaded some data with labeled emotions and used it to train our model. To get a good result, we tried a number of machine learning algorithms and chose the one with the best accuracy on our training set. We also reserved some part of our downloaded data for cross-validation.

After cross validation, we applied the best model to the given dataset and classified them accordingly. Finally we added the emotion labels to our csv file and then started exploratory analysis. (The code for our model will be attached with our submission.)

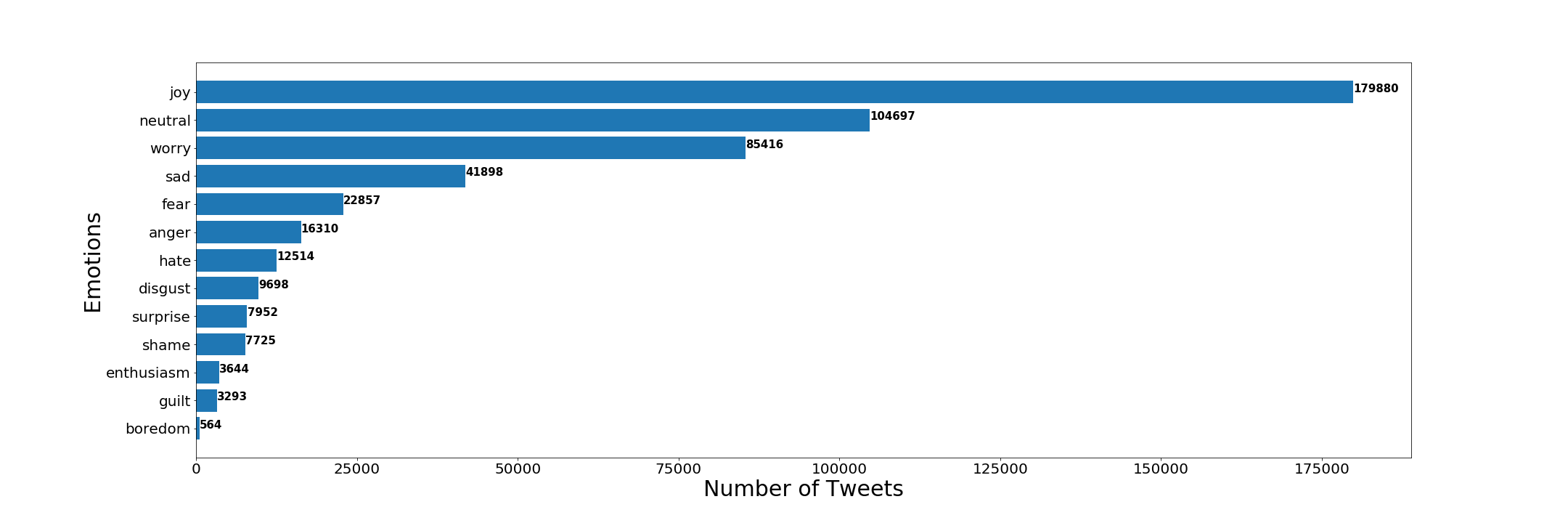
Given below is the table which depicts our result after the classification of the dataset.



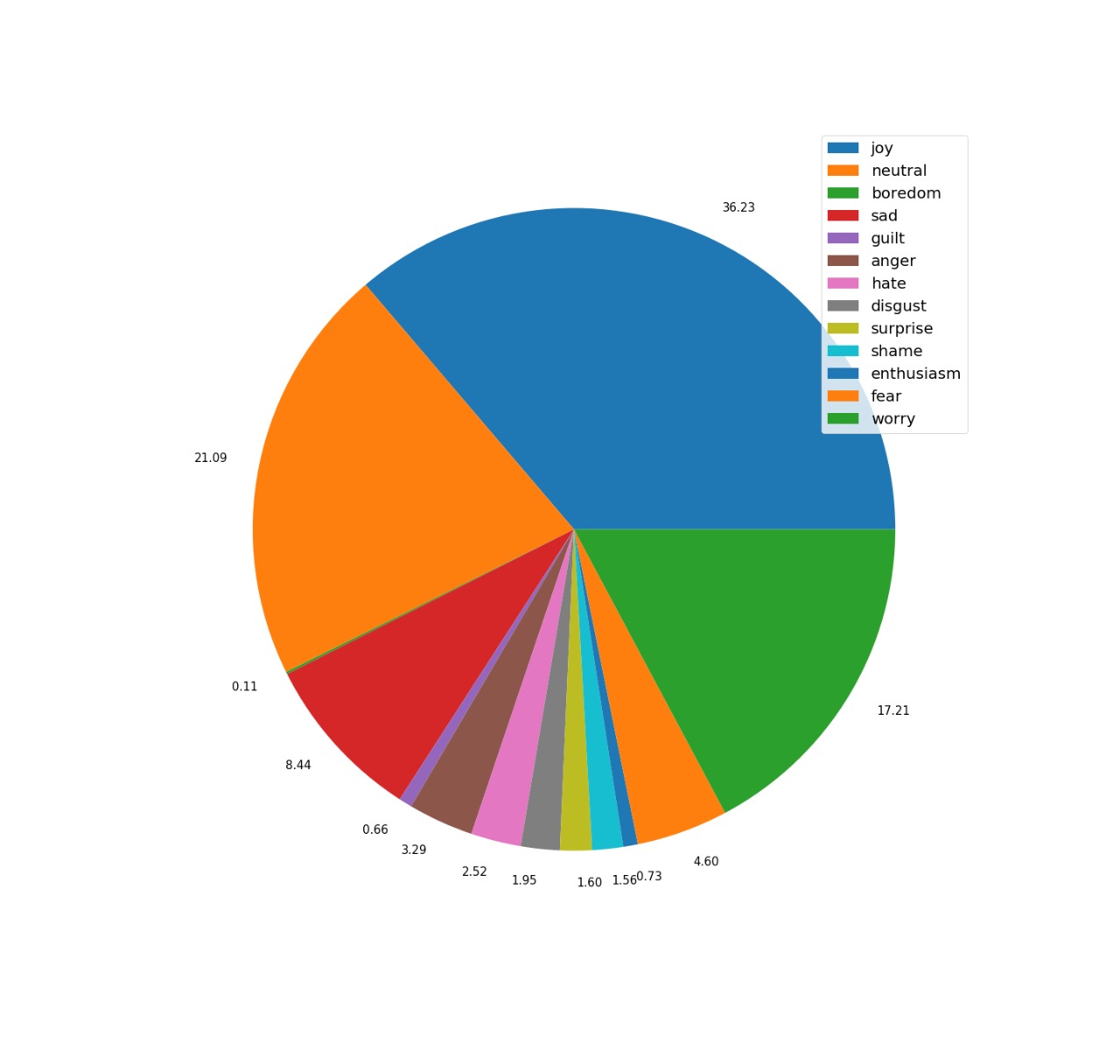
**Exploratory Data Analysis**

**1.1** To show the result of our classification and to show how many tweets belonged to a single emotion, we created a column and pie chart. These charts also depict which emotion turns out to be the most prevalent during the pandemic.

The rows of the column chart show different emotions, which have the values in form of the count of the tweets under each emotion. The pie chart shows the percentage of each emotion.



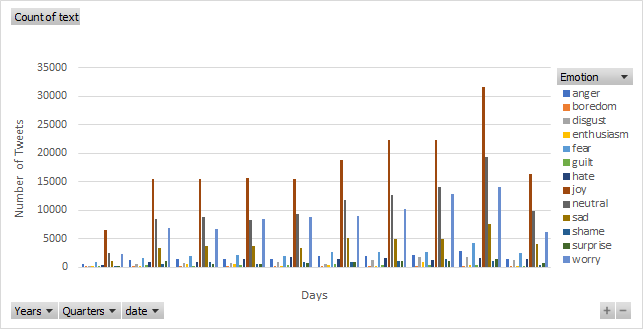
**Fig. 1.1** *(a) Column chart of Tweets according to their emotion*

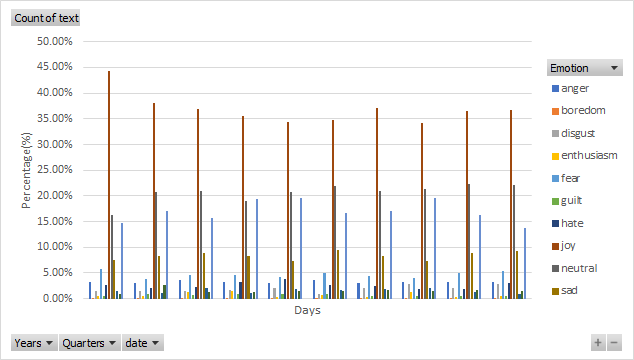
**Fig. 1.1** *(b) Pie chart of Tweets according to their emotion*

***Analysis***: Count of Tweets under each emotion tells us which one of them is the most prevalent.

**1.2** We were given data belonging to a 10 day period. So to compare which emotion turns out to be the most prevalent on a given day , we created the following column chart.

The rows of the chart show different days from 13th Sep to 22nd Sep 2020, which have the values in the form of count of the tweets under each emotion. This chart helps us to compare data given to us on different days.

**Fig 1.2** *Column chart of Tweets on different days according to their emotion*

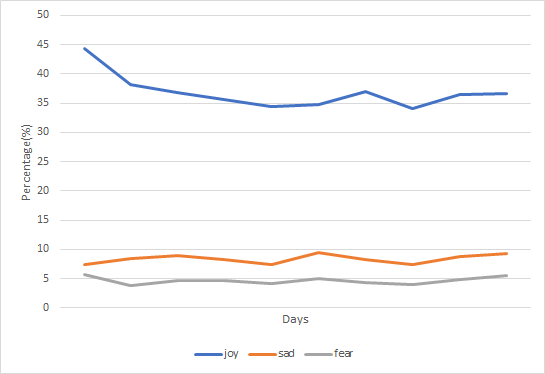
This chart is a good representation of the data on a daywise basis but it does not provide for our purpose of data analysis. So we created a modified version of this chart where the ordinate is the ratio of the number of tweets for an individual emotion for a day to the number of total tweets for that day.

**Fig 1.2** *Column chart of Tweets on different days according to their emotion*

***Analysis***: Here we can observe the trendline of different emotions throughout the given period.

**1.3** We chose a Line chart to compare the three basic emotions namely Joy, Sad and Fear which cover a great portion of how people feel on a given day.

The abscissa shows days ranging from 13th Sep to 22nd Sep 2020, which have the values in the form of percentage of tweets for an emotion(which are Joy, Sad and Fear).

**

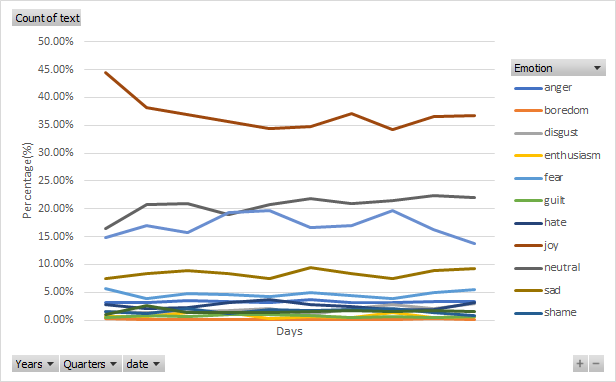
**Fig 1.3** *Line chart of Tweets on different days according to three given emotions*

***Analysis***: Trendline showing the three dominating emotions in the given dataset.

**1.4** This graph is used to analyse the same data as chart 1.2. However, this helps to visualize the correlation even better that is there between the different fields.

Given below is the image of the line chart used to compare which emotion turns out to be the most prevalent on a given day.

The rows of the chart show different days from 13th Sep to 22nd Sep 2020, which have the values in the form of percentage of tweets under each emotion.

**

**Fig 1.4** *Line chart of Tweets on different days according to their emotion*

***Analysis***: Trendline showing all the emotions in the given dataset over a period of ten days.

**Conclusion**

*From the graphs***→**

**1.1** From Count of Tweets under each emotion, we come to know Joy is the emotion with the highest number of Tweets(1,79,880/4,96,448)

**1.2** The trend in the plot for different days is almost similar. This shows that it takes time and effort to change the public opinion and an interval of 10 days is not enough for that.

**1.3** The plot basically shows us the public opinion over the given period of time. Here, Trend for joy has declined a bit while for fear and sadness, rend has remained mostly constant.

**1.4** This plot shows us the correlation between different emotions for the overall span.

In this fast growing digital world, social networking services like Twitter, Facebook, Instagram are becoming a very popular medium of communication and a medium to express oneself.

We were provided with the twitter data of recent times. Compared to the time when the outbreak occurred, the public opinion has changed a lot.

From the Analysis of the data given, we come to know that the mental health of society has improved a lot. Emotions like worry and sadness are still quite dominating but the times are getting better. As it can be seen from the charts above, Joy has been the most prevalent emotion on twitter for the given dataset.

However, if we were to sum up the count of tweets under emotions such as sad, fear, anger, hate, shame, worry, etc., (***Negative Emotions***) we would find that they combinely surpass those related to (***Positive Emotions***) joy.

So, if we were to do classification based on positive, neutral and negative tweets rather than what we actually did, we would find that there is still a greater number of negative emotions on twitter or public opinion in general than the positive during the pandemic.

During this pandemic situation managing mental health and psychological well-being is as important as physical health. Understanding public opinion and emotions can help the Ministry of Health and family Affairs to take necessary actions.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

A Report By:-

Team Ravi Kiran

(Ravi Kiran.B,

Alakh Verma,

Harsh Surya)