5 IEEE International Conference on Recent Advances and Innovations in Engineering- ICRAIE 2020 (IEEE Record#51050)

Sentiment Analysis on COVID-19 Twitter Data

Tanmay Vijay
Computer Science and Engineering
Rajasthan Technical University
Kota,India
tanvijay27994@gmail.com

Ayan Chawla
Computer Science and Engineering
Dav University
Jalandhar,India
ayanchawlae@gmail.com

Balan Dhanka
Centre for Converging Technologies
University of Rajasthan
Jaipur,India
dhankabalan@gmail.com

Purnendu Karmakar
Electronics and Communications Engineering
The LNM Institute of Information Technology
Jaipur,India
purnendu.karmakar@lnmiit.ac.in

2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE) | 978-1-7281-8867-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICRAIE51050.2020.9358301

***Abstract*—Corona Virus or COVID-19 first appeared in December, 2019 in Wuhan, China. People tweeted aggressively on twitter at that time. This paper analysed the tweets regarding COVID-19 from November, 2019 to May, 2020 in India and its affect. All tweets are categorized into 3 categories(Positive, Negative and Neutral). Multiple datasets are created State-wise, Month-wise, combined of all states to analyze the people's reactions towards Lockdown in June, 2020 and about everything related to COVID-19. Most people started having Negative tweets but with increasing time people shifted towards positive and neutral comments. In April, 2020 most comments were Positive and about winning against Corona virus.**

***Index Terms*—Sentiment Analysis, COVID-19, Polarity.**

I. Introduction

Corona viruses are a gathering of related RNA infections that cause illnesses in warm-blooded animals and feathered creatures. In people, these infections cause respiratory plot contamination that can go from mellow to deadly. Mellow diseases incorporate a few instances of the basic cold (which is additionally brought about by different infections, overwhelmingly rhinoviruses), while progressively deadly assortments can cause SARS, MERS, and COVID-19. Side effects in different species change: in chickens, they cause an upper respiratory plot illness, while in dairy animals and pigs they cause loose bowels. Antibodies or antiviral medications to forestall or treat human coronavirus diseases are not found yet. Coronavirus spreads from humans to humans in several manners. It causes respiratory problems and causes breathing problems. It has a low fatality rate compared to SARS or MERS [1].However, as there is no vaccine available and due to the nature of transmission and spread of virus, countries around the world taken lockdown and isolation as the only preventive steps.As a result people's movement being restricted they spent a good amount of time in home or place of stay. This creates a ideal scenario for people to express their views on social networking sites including Twitter. Like in other parts of the world people of India also express their views about COVID-19 on twitter. In this paper, those tweets are used to analyze the behavior and sentiment of people from December 2019 to May 2020. People were very negative about COVID in the early days. Many were not confident to win the fight against COVID. But as the Lockdown happened, people became optimistic about it. Positive tweets are tremendously high in the amount in the month of April and May. This work provides new insights on COVID-19 and people's thoughts about it. Social media is today's way to read people.There is a good number of papers appeared in the literature since the outbreak causes countries to go in Lockdown state. Sentiment analysis and data visualization of World wide Twitter data on Covid 19 has been presented in [2]

A. Contribution

The trends of positive, negative, and neutral tweets state-wise and month-wise in India are captured and presented in this paper. In this paper we have presented the sentiment of people from the Indian state Maharastra as a case study. It shows that people were worried about the COVID-19 situation before Lockdown and after that, though COVID-19 impacted the moral as well as socio-economic life of people still there is a positive sentiment dominates in the opinion of the people. The twitter feeds were full of positive and hopeful tweets of winning the war against the Coronavirus. A dataset with minimal cleaning needed is created having tweets from November 2019 to May 2020 with their Polarity and Sentiments of each State in India. We have specifically presented sentiments around the dates of announcements about lockdown in figure 4. In our paper, we have considered the emotional level as positive, negative and neutral. For

better understanding of emotion level of people, this model can be further extended to more complex multi-emotion levels such as joy, panic, happiness, sorrow.

II. Data Collection and Preprocessing

This section explores how data is collected using APIs and what Pre-processing steps were followed. For analysis, the tweet-data was collected from Nov 2019 to May 2020. All the tweets collected in that window were separated based on the States(Indian) they belong to and Months they were collected from. The second phase of data collection included a collection of tweets separated by each day of each month.

A. Data Collection

Data Collection was one of the lengthy and major tasks to perform.

1) Phase - 1: Month-wise Collection

Data was collected using GetOldTweets3 [4] API (Python) and was stored in multiple files separated state-wise and month-wise. Three keywords (corona, COVID, COVID-19) were searched through twitter [5] for those months for collecting tweets shared among different states. For instance, a file named "Assam_01_tweets.csv" contains the tweets shared in Assam during January.

We created a list of all Indian states and their approximate radius to find the data required within a restricted region at a time. We looped through all the months for all the states while collecting the month-wise data and saved the collected tweets in separate CSV files. This phase of data collection resulted in a collection of approx 140,000 raw tweets over the duration of December 2019 to May 2020.

2) Phase - 2: Day-wise Collection

The second phase of data collection was performed to collect all tweets from a given region (Indian State), but this time separated by individual dates. This phase was separated from the month-wise collection because month-wise analysis of tweets required fast data collection to enable easy and rapid verification of ideas, while day-wise analysis [6] required more data points and computation-intensive graphical calculations to be done on large data points.

Automated query generation using $GetOldTweets3$ [4] API in Python eased the task of collecting approx 250,000 data points and took about a day of scraping time. The scraping process was broken down in threads running on Google Collaboratory and local machines simultaneously. It took around 10 hours for scraping of the data.

B. Pre-processing

The pre-processing phase was comparatively easier and more enlightening than other phases. This phase included both data Pre-processing and part of Exploratory Data Analysis.

1) Phase - 1: Data Cleaning

As we delved into the process of cleaning the data to find useful features, we realized that the raw tweets were incapable of generating unbiased outputs in Sentiment prediction. The main hurdles were #tags (hashtags), @mentions, web links (URLs), and stop words present in tweets. We used regular expression based substitution to remove the #tags, @mentions, and URLs from the tweets. Stop words were handled by NLTK library in python under the hood of Textblob (python library).

2) Phase - 2: Finding Polarity

This phase was the most essential Pre-processing step through the process. With the help of the TextBlob module in Python, we estimated the polarity scores of each tweet in the dataset. The cleaned tweets from the previous phase were subjected to multiple evaluation models using TextBlob and a generalized score of polarity was found for each tweet. This score was directly correlated with the group of words present in the text, i.e., Unigrams, Bigrams, Trigrams, etc.

TexBlob API returns a value in the range [-1,+1] where +1 implies Extreme Positive Polarity, -1 implies Extreme Negative Polarity and 0 implies neutral. This score was added as a separate attribute in the dataset.

3) Phase - 3: Finding Sentiments

This phase was an extension over the last phase to categorize the polarity scores of tweets into 3 classes (namely: Positive, Negative and Neutral). Positive Sentiments are those having range (0,1]. Negative sentiments range is [-1,0), and neutral sentiments are having 0.0 polarity. A simple looping through the dataset and applying filters concluded this phase. These 3 classes were stored as a separate feature in the dataset called "Sentiments".

4) Phase - 4: Combining the Dataset

In this section, we combined the various datasets created during Data Collection and Finding Polarity phase into more manageable and workable datasets. We combined the polarity datasets state-wise, i.e., we create a common dataset with the "Month" column for each state. After that, we also combined all the newly formed state-wise datasets to form a large combined dataset with a new attribute named "State".

The whole process of Pre-processing took just over a day of processing and coding time. This process revealed a lot of interesting properties of data, like Positive polarity was more common, surprisingly than neutral and negative ones, even during the most hyper-active months of Corona. Neutral polarity took the second spot. The reasons for such occurrences are discussed in the Analysis and Results section.

III. Frameworks

We have collected and analyzed data in a system of i5 8th generation with 8 GB of RAM in windows operating system. TextBlob API was used to get the polarities of Tweets that use Natural Language Processing(NLP) to get the polarities of each tweet [3]. TextBlob uses multiple NLP techniques to get the sentiments [7]. Techniques used in this process were-:

A. Parts Of Speech Tagging (POS)

POS tagging is assigning tags to each word in the sentence which is used in Lemmatization [8].
For example-: "Corona is making economy of World down". In this sentence, it makes a list of tuples in Python.
[('Corona', 'NNP'), ('is', 'VBZ'), ('making', 'VBG'), ('economy', 'NN'), ('of', 'IN'), ('World', 'NNP'), ('down', 'IN')] NNP-Proper Noun, VBZ-Verb, NN-Common Noun. This tagging is used in the Lemmatization of words.

B. Lemmatization

It is the process of making words in their first form of the verb [9]. This done because for instance make and made gives the same meaning but if Lemmatization is not used it will treat all these words separately and it will increase the features and will create redundancy in the analysis. After Lemmatization made, make are converted to make and hence the count will become 2 of make. in the sentence.

C. Stemming

Stemming creates the root form of inflected words. For instance, making and make will be counted as 2 different words in a sentence affecting our accuracy if analysis but after using Stemming will be removed from the word making and it is converted to make [9]. suffixes like *es, ies, ing* and many more are removed from words making our final vector of low dimension.

D. Stopwords Removal

Stopwords are words like is, have, has, etc are diminishing the significance of analysis [8]. Hence, these were also removed. It reduces the redundancy of words and making analysis better as the meaning of the sentence is not changed. The sentence "Corona is making economy of World down" will be converted to "Corona make economy world down". Now this sentence is passed as a parameter to the already trained model of TextBlob and it will predict its polarity based on Unigram Bag of Words model and a Floating point number is returned.

IV. Algorithms

The whole analysis was done in multiple parts. Each time algorithm was followed.

A. Data Collection

The Data Collection Algorithm is shown in Algorithm 1. Location(Each state's name), Radius(Each state's approximate radius, Time(Starting Date and Last Date) Dataset of state Month-wise trytry: catchcatch:end

Required Libraries Imported A list of all State's name, Time and Radius is created List not Empty Get Tweets Save the Tweets in statename_month_csv file Exception **Print**(This state is empty) Data Collection Algorithm

B. Getting Polarities

The Polarities and Sentiments are generated by Algorithm is shown in Algorithm 2. Each statename_month_csv Dataset Each statename_month_csv Dataset with Polarities and Sentiments of each Tweet trytry: catchcatch:end

Required Libraries Imported A list of all statename_month_csv file is created List not Empty Get Polarity of Each Tweet in each file Polarity is greater than 0 Sentiment of that tweet is saved Positive Polarity is Less than 0 The sentiment of that tweet is saved Negative The sentiment of that tweet is saved Neutral Save the Polarity and Sentiment in statename_month_Polarity_csv file Exception **Print**(This state is empty) Polarity and Sentiments Generation Algorithm

V. Analysis

In this paper, TextBlob is used to find the polarity of scraped tweets and Natural Language Toolkit (NLTK) for word frequency [10]. In this paper firstly state-wise analysis is done and then the frequency of Positive, Negative, and Neutral tweets are calculated. Each state is analyzed month-wise separately. Certain insights were discovered about the data using Visualizations [11]

As per the frequency plot in Fig. 1 it is observed that there

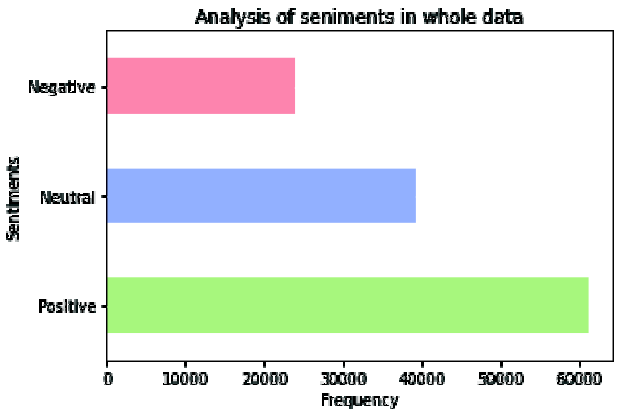

Fig. 1. Analysis of Polarity in India from Dec 2019 to May 2020.

were 61,102 tweets with positive sentiments, 39,195 were having neutral sentiment and 23,987 tweets with negative sentiments in India from Dec 2019 to May 2020. People in India were expressing more positive sentiments as compared to negative and neutral.

Fig. 2 shows the totals no. of tweets in India from different states. Madhya Pradesh was having more no. of tweets which is 44,252 as compared to any other states in India from Dec 2019 to May 2020. Jammu and Kashmir were on 2nd place with 31,769 tweets. Madhya Pradesh was in 3rd place with 18,577 tweets. Rajasthan was on 4th place with 13,732 tweets from Dec 2019 to May 2019. Mizoram was having only 1 tweet during this period.

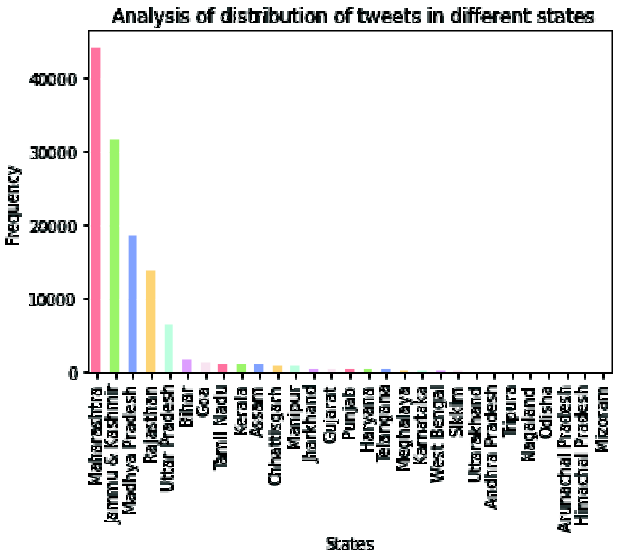

Fig. 2. Analysis of the frequency of tweets in different states of India from Dec 2019 to May 2020 .

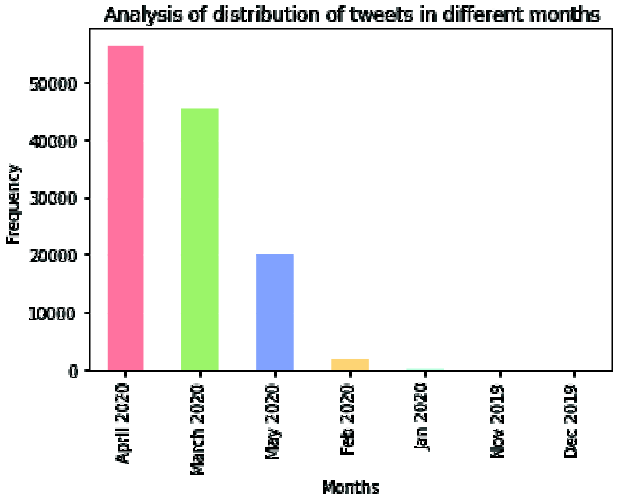

Fig. 3. Month wise Analysis of the frequency of tweets in India from Dec 2019 to May 2020.

Fig. 3 shows that in India there were 56,392 tweets in April 2020 which was much larger than all the months from Dec 2019 to May 2020. and in March there were 45,476 5 which was also greater than any other months except April because at the end of March 2020 Lockdown was announced by the government and people started sharing their views by tweets. So there was a hike in the no. of tweets in May and April 2020.

Fig. 4 shows the frequency of positive, negative and neutral tweets in India on lockdown announcement dates. As in the graph, we can see that frequency of tweets on 22nd day is very high which is fair enough because Lockdown in India was also announced in March 2020 so people started sharing their

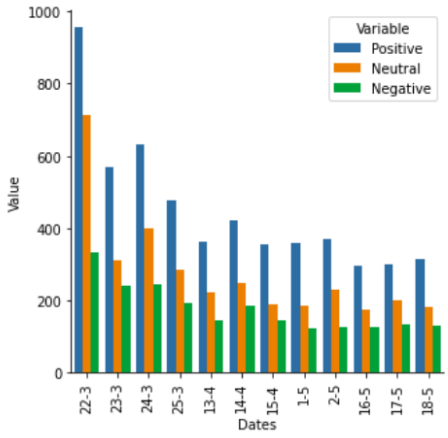

Fig. 4. Overall sentiments on Lockdown and nearby dates.

views on this decision of Lockdown taken by the government. And after 22 March 2020, it is observed that the frequency of the tweets was very high than that of the normal days.

Fig. 5 is the bar graph is of Maharashtra which shows that

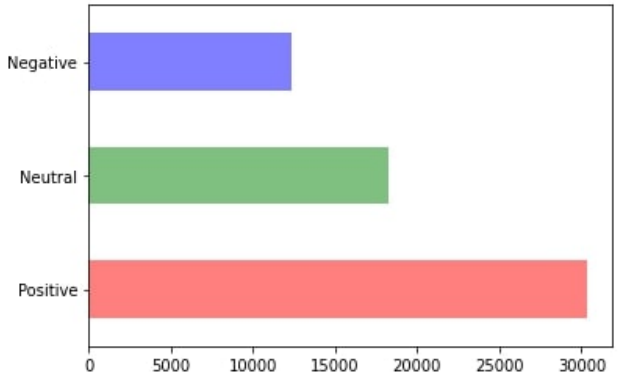

Fig. 5. Sentiment Frequency plot of Maharashtra.

there more no. of positive sentiment tweets as compared to neutral and negative sentiment tweets. People started accepting Lockdown and happy with spending their time with their families.

TABLE I
OVERALL SENTIMENTS WITH POLARITY IN MAHARASTRA.

Month	Positive	Neutral	Negative
Nov, 2019	0	0	2
Dec, 2019	0	0	0
Jan, 2020	20-25	25-30	15-20
Feb, 2020	155-160	140-145	115-120
Mar, 2020	$>$ 9000	$>$ 5500	$>$ 3800
Apr, 2020	$>$ 11500	$>$ 7000	$>$ 4000
May, 2020	$>$ 1150	$>$ 600	$>$ 500

The data in table 1 shows that people in Maharashtra were completely ignorant of the Corona Situation until

December. They picked up some concern about Corona during January-February, but remained mostly unconcerned.
In March, people started to realize the impact of the viral disease and the increase in no. of tweets reflect that.
People remained mostly positive about the situation but no. of negative tweets confirm that situation had started taking a serious front in the state and people's psychology.
In April, the effect magnified itself to a great extent but more people kept hope, maybe as a result of constant efforts of the government to keep the population away from anxiety.
May data tends to show that people were now lesser concerned about the pandemic and more hopeful about the future.

Fig. 6 represents the line plot of overall sentiment with

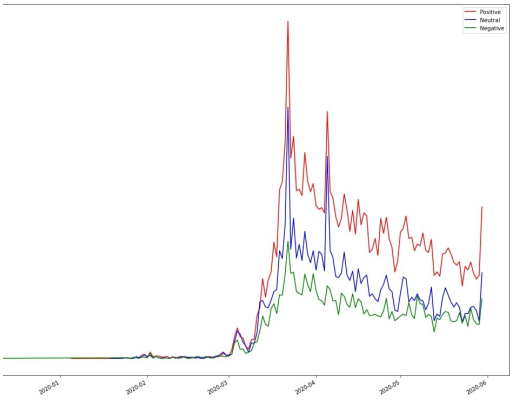

Fig. 6. Overall Sentiments with Polarity in Maharashtra.

polarity in Maharashtra. It is observed that there were more positive sentiment tweets as compared to negative and neutral sentiment tweets.
And there is a peak in the frequency on 22 March which is because of the announcement of the Lockdown.

In Fig 7 word $corona$, and $COVID$ was used most

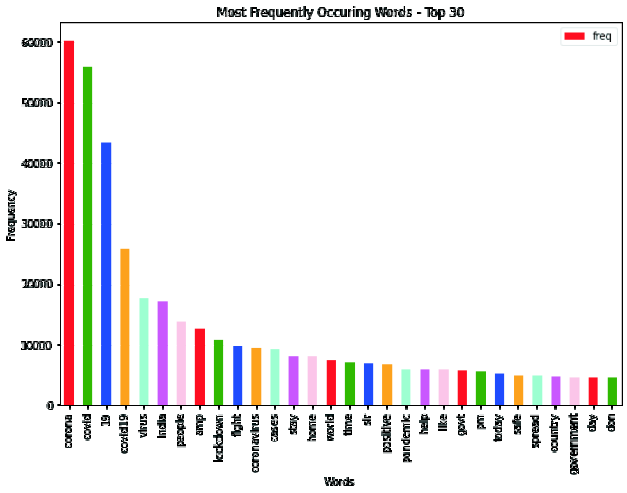

Fig. 7. Word Frequency plot in Tweets.

frequently used with a frequency count of more than 50,000 in tweets from November 2019 to May 2020. The words $country, government, day, anddon$ was used less than 6000 times. The top 10 words which were used frequently are:-
$corona, COVID, 19, covid19, virus, India, people, amp, lockdown, fight$
This shows that people's sentiments varied from moderate to highly negative emotions. Overall India spent the months in Lockdown mostly in fear and partly in anguish.
The lower end of the graph contains words like
$like, country, day, today, government, spread, safe$
The right part of the above graph states a different story. It shows people's emotion in that region varied from health worries and outcomes of the pandemic to political scenarios.

A. Overall Analysis of each State

Every state has overall polarity combining all months. It shows how people of different states reacted towards the COVID-19 situation and Lockdown. Each polarity has a percentage in each state. November and December months doesn't contain any tweet in Maharashtra in English.
In Maharashtra, Positive Tweets Percentage is shown in the Table(number)

1) Positive Tweets in Maharashtra

Positive Tweets increases as time increases and reached up to 50% in May 2020. Many states show an alike trend.

2) Negative Tweets in Maharashtra

Negative Tweets started from 0% in January 2020 then, a sudden jump to 31% in February, and then reduced to 19% in May 2020. Most states started from a high quantity of negative tweets and get reduced as time increases.

3) Neutral Tweets in Maharashtra

Neutral Tweets started from 66% in January 2020 and reduced to 30% in May 2020. It is a very unlikely trend as Neutral was high in this state. Most states contained an average of the negative and positive amounts of neutral tweets.
This analysis helped a lot in understanding the mindset of people about this whole pandemic situation. Some states don't have a fair amount of tweets for analysis to be done on them fairly. Indians have found a way to be positive in each situation. Most positive tweets are supporting Lockdown and are about having a Vaccine for Corona Virus. Neutral tweets are mostly sarcastic in nature as NLP is not still that great in differentiating between sarcasm and positive and negative sentiment. Negative tweets were at peak when COVID destroyed countries like Italy from February to April. Lockdown was a necessary evil.

VI. Results

The analysis phase of the process gave us great insights into the emotional state of people in India, and also, how sentiments varied from state to state on daily basis.A similar

study was carried out in [12] for Nepal.In that initial steps for data collection and tools used are same.However, number of tweets collected and range of dates are lesser.Further, authors in [12] more focused in word cloud variation and sentiment variation presented in the form of pie chart. Authors in [13] also taken same approach and used map and bar charts to visualize the sentiment. In our study we have taken different approach and presented the sentiment variation over time in the form of graph. Also, we have pr resented tweet frequency of different states.Then as a case study Maharashtra has been taken. This approach of analysis allowed us to realize that emotional versatility in the Indian population can be very much affected by the sate government and measures they take to control a pandemic like a corona, by the amount of information available to people of a certain region and how actively people are willing to accept the negative side of a situation. In this work, our aim was to show that physical happenings in the real world are also reflected in the virtual social network. The positive tweets peaks are coinciding with the lockdown announcements as can be seen in Fig. 4. Further, we can see that corona and COVID-19 dominate the tweeter space in the observed duration. Our results will help Govt. or policy making agencies to decide upon the framework in the event of a future pandemic or similar events.

A. Polarity Analysis Results

From the analysis of Fig. 1, it is clear that even in such harsh situations of strict Lockdowns and increasing corona cases, people were mostly Positive. Negative tweets are mainly caused by trigger events that are more political in nature.

B. Most Affected States

As seen in Fig. 2, it is an interesting observation that some states were more affected by the pandemic that others. The tweet frequency in a state directly correlates to how many cases of the corona were realized in that state. People's emotions varied the same way. The no. of tweets from a state also reflects the vulnerability of people of that state to the pandemic, hence they were more active and more informed about the situation. Maharashtra, Rajasthan, and Madhya Pradesh were greatly affected but remained more hopeful and Positive compared to states like Jammu and Kashmir and Uttar Pradesh.

C. Most Active Months

Fig. 3 shows that India was most active during the months of March 2020 to May 2020. This is directly correlated with the fact that India was introduced to the virus in March and the no. of active cases increased vastly in March. In India, April was mostly active Month because a lot of trigger events like Lockdown announcements and no. of deaths were also at the peak during that time.

D. Looking at the Individual States

From Fig. 4, we can conclude that 22 March was the day of Lockdown starting in India. This graph is for the state of Maharashtra in the Month of March, but it speaks for all states of India. So does the Fig.5, in most of the states of India, Positivity remained intact compared to Negativity.

E. Variation of Sentiments in Each State

The sentiments predicted using the tweets, when categorized into classes like Positive, Neutral, and Negative, and plotted on a date-vs-frequency curve, Fig. 7 is what was resulted. Fig. 7 shows not only the active dates from Nov 2019 to May 2020 but also how people were bothered by the situation at any given date. The highest peak in March is when the government first announced the Lockdown. The word frequency curve can also be analyzed to predict the sentimental state of the people. The top 30 words show that people were indeed fearful and anxious about the effects of the corona, but remained hopeful and trustful with governmental measures.

VII. Conclusions

From the analysis in this paper, it is observed that people in India were mostly expressing their thoughts with positive sentiments. This paper concludes that there was a sudden hike in the tweets on every date when there is an announcement of Lockdown. The states like Madhya Pradesh, Maharashtra, Jammu and Kashmir, and Rajasthan was having higher no. of tweets as compared to other states because there were more no. of positive cases of coronavirus as compared to other states. While the people were posting tweets with negative sentiments in India, the twitter audience of India seems to reward positive sentiment much more than negative sentiment reversing the overall polarity to positive. It seems although we have faced with fear and anticipation about the Coronavirus and the future, the trust in the Government of India to address the Corona crisis supersedes all such fears and anticipation emotions. This analysis was unique in its way as Lockdown shows peaks and were positive. Most people were criticizing Lockdown on the face but twitter was full of positive tweets. This analysis can be further taken to new possibilities of Emotion analysis. Rather than having Positive, Negative, and Neutral tweets, we can analyze based on emotions. Text can also represent the emotions of a person writing. Tweets can have emotions like Hate, Respect, Agreement, Anger, Happiness. Each tweet can have multiple emotions and we can have the emotion having a maximum score. Only a few states like Maharashtra are more affected by Corona and states like Jharkhand are having very few cases in comparison to others. In these states, people can have multiple emotions and several amazing insights can be generated. In the future, it holds the potential to discover the mindset of people. In this work, we have only used tweets in English language. Tweets in other Indian languages if collected will have a better representation of people's sentiments.

References

[1] Medical News Today,https://www.medicalnewstoday.com/articles/how-do-sars-and-mers-compare-with-

covid-19,last accessed on 20-10-2020
[2] Manguri, Kamaran H, Ramadhan Rebaz N Amin,and Pshko R Mohammed,"Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks," Kurdistan Journal of Applied Research,54–65,2020
[3] C. Kaur and A. Sharma, "Twitter Sentiment Analysis on Coronavirus using Textblob," EasyChair2516-2314, 2020.
[4] Release, GetOldTweets3 API in Python Release v3 0.0.11, November 2019.
[5] Twitter sentiment analysis: The good the bad and the omg!Kouloumpis, Efthymios, and Wilson, Theresa, and Moore, Johanna, Fifth International AAAI Conference on weblogs and social media,2011.
[6] A. D. J. A. a. S. Dubey, "Twitter Sentiment Analysis during COVID19 Outbreak," 2020.
[7] P. Tyagi and R. J. A. a. S. Tripathi, "A Review towards the Sentiment Analysis Techniques for the Analysis of Twitter Data," 2019.
[8] Varsha Sahayak, Vijaya Shete and Apashabi Pathan, "Sentiment Analysis on Twitter Data," in International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 1, Volume 2 (January 2015).
[9] Namrata Godbole, Manjunath Srinivasaiah, Steven Skiena, "Large-Scale Sentiment Analysis for News and Blogs," Google Inc., New York NY, USA and Dept. of Computer Science, Stony Brook University, Stony Brook, NY 11794-4400, USA.
[10] N. K. Rajput, B. A. Grover, and V. K. J. a. p. a. Rathi, "Word frequency and sentiment analysis of twitter messages during Coronavirus pandemic," 2020.
[11] M. Ra, B. Ab, and S. Kc, "COVID-19 Outbreak: Tweet based Analysis and Visualization towards the Influence of Coronavirus in the World,"2020
[12] Twitter Sentiment Analysis During Covid-19 Outbreak in Nepal, Pokharel, Bishwo Prakash, Available at SSRN 3624719,2020.
[13] Guntaka, V. S. P. R., Gupta, A. K., Somisetty, S. 2020. Twitter sentiment analysis and visualization – In Proceedings: 16th Annual Symposium on Graduate Research and Scholarly Projects. Wichita, KS: Wichita State University, p.31