Measurement and Analysis for COVID-19

Steffy Dias
Binghamton University
Binghamton, NY, USA
sdias1@binghamton.edu

Sagar Sahani
Binghamton University
Binghamton, NY, USA
ssahani1@binghamton.edu

Arun Reddy Gummala
Binghamton University
Binghamton, NY, USA
agummal1@binghamton.edu

Figure 1: Measurement and Analysis of COVID19

ABSTRACT

The hour of the pandemic in 2020 has brought intensified efforts and experimental analyses by researchers and scientists alike to analyze the dynamics of COVID-19. This has led to great importance to the related data that is readily available in huge amounts these days on social networking sites. The pouring amount of data available every week, every day and nearly every hour can be of great importance once brought to its correct use.Collecting this data to make meaningful insights can help a great deal. Data collected through these online social networks can lead to better evaluation of the pandemic situations and how quantified analysis from this data can be carried out to answer some relevant research questions. To get answers to such questions we have put forth some analysis algorithms to study the behavioural change experienced by mankind during the peaks and decline of the cases during this past year, and whether or not it has an associative effect on the number of online posts of how the new normal has treated everyone.

KEYWORDS

datasets, twitter api, reddit api, sentiment analysis,polarity

ACM Reference Format:

Steffy Dias, Sagar Sahani, and Arun Reddy Gummala. 2020. Measurement and Analysis for COVID-19. In *Binghamton '20: ACM Symposium on Data Collection for COVID-19, November 03–05, 2020, Binghamton, NY*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

One of the many severe pandemics that impacted the aspects of modern human society is undoubtedly the COVID-19. There is altogether a great difference between the world before and after this pandemic. This can be observed through the naked eye with the ongoing social distancing practices, the way our work environment got structured and the effect it brought to our economic activities. This pandemic has opened up several opportunities and challenges that require extensive scientific research and analysis. It has resulted in a rapid response and action from the scientific community.There is a rapid growth in online communities in these recent years. Enormous amount of data is generated by many social networking sites everyday and every single hour. This same data can go a long way and help us as a source for a variety of analysis purposes. Twitter and Reddit are one of the many popular networking sites with millions of users. This includes users with a variety of views, reviews and opinions that can be gathered with the help of tweets and Reddit posts.

The COVID-19 pandemic has led to many efforts being taken to

Binghamton '20, November 03–05, 2020, Binghamton, NY
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

analyze the data collected in many countries and cities to study and predict its growth so that it can help plan for healthcare resources and socio-economic decision making. The pandemic has caused huge loss in public health, affecting millions of people world wide throughout countries.The coronavirus pandemic created an urgent threat to global health causing disruption in daily livelihood all over the world. Despite government lock downs and public health responses aimed at containing the disease and delaying the spread, several countries have been crippled with a critical care crisis, and many more following in future. Dynamic projections play a key role in conditions like COVID-19 where daily cases change rapidly based on government restrictions, test conditions, infection rate, etc.

This for sure has brought about natural change to the way mankind has reacted, a change in their behavioural pattern is certainly seen. This new normal has certainly put forth many questions unanswered. Our aim in this project is to study this change in behaviour and to answer some of the many questions through analysis algorithms and to put forward a pictorial representation of the change that is observed[1] [2]. We intend to answer the following research questions through our study:

- RQ1: How has mankind accepted COVID19? What behavioral change can be seen when the cases rise to a peak and when the number of cases decrease?
- RQ2: Similarities in the number of posts during the peaks and declines in the COVID cases?
- RQ3: Does the analysis done by the data generated from these social networking sites correlate with the analysis done on data generated from other sources?

2 MOTIVATION

To put forth relevant analyses of the behavioural change that is observed in mankind and how he has been affected by the peaks and declines in the cases this pandemic has brought.

3 PROBLEM STATEMENT

To do insightful and impactful analysis of COVID19 data, collected from popular social media websites to answer some research questions.

4 LITERATURE SURVEY

He [4]through his paper explains the risk prevention and control decision-making model proposed due to the incompatible characteristics of the pandemics risk prediction. Further mentioning how the collection of data on the personnel flow helped analyze the attribute of epidemic risk.

S. Roy, M. N. Pal, S. Bhattacharyya and S. Lahiri [8] through their paper mentioned how they were successful in implementing an online platform for COVID-19 which is capable of disseminating accurate prediction of the confirmed, deceased and affected COVID-19 cases on the bases of data available in reliable online repositories.

S. Panja, A. K. Maan and A. P. James [7] through their paper present an implementation of a semantic search engine targeted at COVID19 research articles. Their system is lightweight and can be deployed as a stand alone system.

Koyel Chakraborty, Surbhi Bhatia, [5] through their paper have analysed that though there are tweets related to COVID-19 and WHO, they have been unsuccessful in guiding people. It shows how data collected through a particular time period during the pandemic, after analysis shows the maximum number of tweet portrayals as neutral or negative sentiments. Whereas tweets collected in a time period before the pandemic show positive or neutral sentiments.

Bishwo Prakash Pokharel [6] through his paper analyzes the data collected from users who shared their location as Nepal, through a particular time period resulted majority of them having a positive and hopeful approach, whereas there were instances of fear, sadness and disgust exhibited too. These papers represent how collected data through online sources can prove to be helpful in better analysis of the study and help a great deal to answer research questions.

5 DATASET

5.1 Twitter data

Initially we collected 18,750,000 unfiltered tweets from Nov 21st to Nov 27th and saved into the mongo in the json format. Then we filtered the data using the hashtag COIVD19 and got a total of 8680 tweets related to covid, which was stored in the json format and then converted to csv for further analysis. During the collection of the tweets we appended timestamps to each tweet as twitter does not give us the time and date parameter.

5.2 Reddit Data

For the data, we initially went with collecting all the params that came along with each post, that results in more than 100 subfields which is enough to go in depth analysis on how accurate the information we get from the post is. But for now as we are just focusing first on analyzing the sentiment of a post we will be using the title along with the timestamp.Here, the timestamp of the post is appended as Reddit does not give us the time and date parameter, but this should not be inaccurate as the collection code was configured to work with the API.All the data was collected, stored and retrieved by using MongoDB for easier access.

5.2.1 . **Preprocessing**
For training our classifier first we preprocess the post, in this we have done letter casing by changing every post to lowercase, then with tokenization separated all the words with spaces. Further, as it was observed that most of these posts included tagging punctuation and special characters, etc., everything as such was removed making the text in a normalized form. Next, for the processing of the words in the post, all stop words were removed as most of these words are of no use to the classifier. nltk's library was used for the removal. Further, We have used lemmatization here over stemming it is because if we do stemming a word may lose its actual meaning. Lemmatization replaces/changes the word to its corresponding vocabulary by using the dictionary itself known as lemma. And hence for our use this would be the best approach.[3]

5.2.2 . **Labelled Data**
For training the classifier, we prepared the data by labelling them into negative, positive and neutral. We made sure to include equal amount of labelled data for each case. For instance, we marked the sentence positive if the sentence indicated something that lead(s)

Table 1: Coivd Tweet count per day

Sr. No.	Date	Count
1	2020-11-23	1875
2	2020-11-24	2057
3	2020-11-25	1924
4	12020-11-26	1468
5	2020-11-27	1356

to decline in the number of cases or it could be a finding related to study of the virus(even if the finding is negative). Neutral when the sentence did not have any outcome for example 'Letter to the FDA About Chlorine Dioxide' which does not describe anything related. Negative is pretty straight forward.

6 TWITTER DATA ANALYSIS

The data collected from 21st November 2020 to 27th November 2020 from Twitter resulted to 18750000 tweets out of which when the filter to collect only COVID related tweets by applying the "COVID19" keyword as the filter resulted in 8690 tweets related to it. The table shows the COVID tweet counts per day.

6.1 Methodology

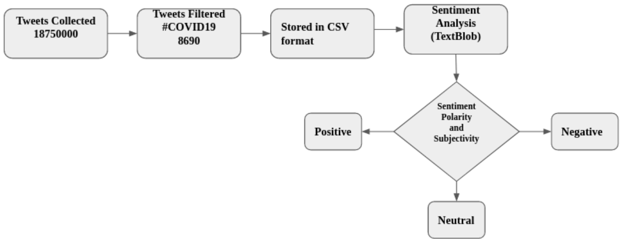

Figure 2: Workflow of Twitter Analysis

The tweets are filtered by using the hashtag COVID19. The gathered data has been stored in CSV format, and fed to the Sentiment Analysis library, namely, Textblob. After collecting the data to perform Sentiment analysis TextBlob library has been called. TextBlob is a Python library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation and more. The main focus on data collection is on the content and timestamps of the tweets. Sentiment analysis has a certain procedure that begins with grabbing the collected data, then identifying the data. The final decision is conducted in the phase of sentiment polarity as well as subjectivity. Then the textblob classifies the sentiments as "Positive", "Negative" and "Neutral", depending on the polarity values.

7 REDDIT DATA ANALYSIS

The data collected from 1st November 2020 to 7th November 2020 from Reddit and 500 posts were manually labelled.

Table 2: Polarity Percentage

Sr. No.	Polarity	Percentage
1	Positive	28.37
2	Neutral	10.49
3	Negative	61.13

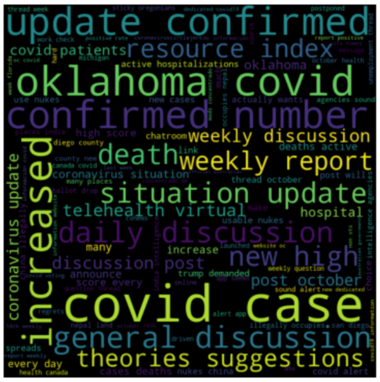

Figure 3: Reddit WordCloud

7.1 Methodology

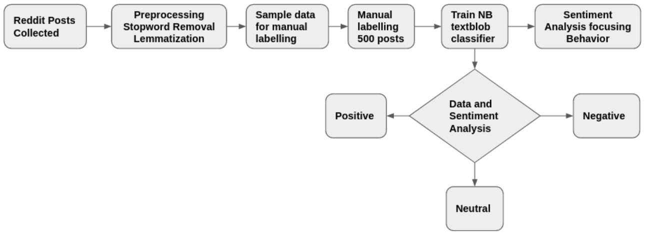

Figure 4: Workflow of Reddit analysis

The Reddit posts were first collected and preprocessed as described earlier. After which labelling was manually done for around 500 posts, which was then used for training the Naïve Bayes classifier from textblob, here after this we compared the polarities that we got from our trained classifier and the textblob, the difference in results was quite huge. As with the general textblob for all the days the positive values were higher than that of neutral and negative which was not possible with the epidemic still being a threat. This also confirmed our classifier to be much accurate than the general text classifier like textblob, vader, etc.
Results from textblob:
'neg': 40, 'neu': 379, 'pos': 381, 'timestamp': '01-nov'
Results from our trained classifier for the same day:
'neg': 225, 'neu': 567, 'pos': 8, 'timestamp': '01-nov'

Further, we have collected data from reddit for 8 days, and in that we have thousands of posts from each day, to get accurate insights we have taken 100 posts for each hour for around 7 hours from each day, this particular timeline was observed to have most active users to be fair. This accounted for 700posts for a day and 4900 posts overall.

8 RESULTS

8.1 Twitter data

In the positive sentimental analysis over the days we can see a gradual decrease in the positive tweet count. In the neutral analysis we can see a increase in the neutral tweet count on Nov 24th but after that, there is a decrease in the tweet count each day. And there is a sharp decrease from Nov 25th to Nov 26th. In the negative analysis we can see a peak in the negative tweet count on Nov 24th but after that, there is a decrease in the tweet count each day from Nov 25th to Nov 26th.

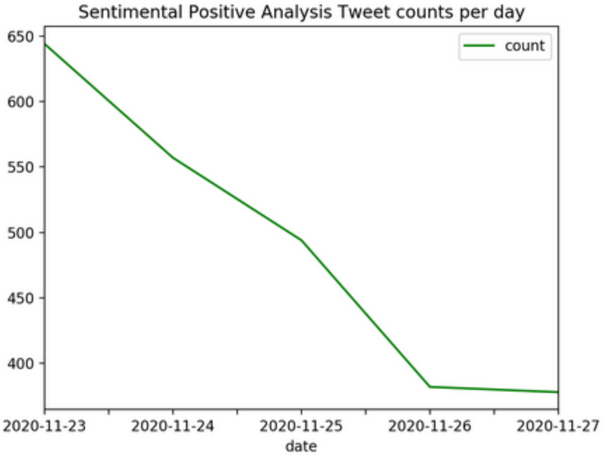

Figure 5: Sentimental Positive Analysis Tweet Count

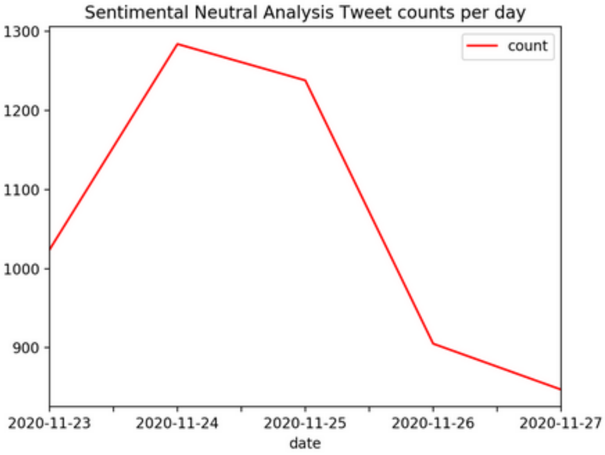

Figure 6: Sentimental Neutral Analysis Tweet Count

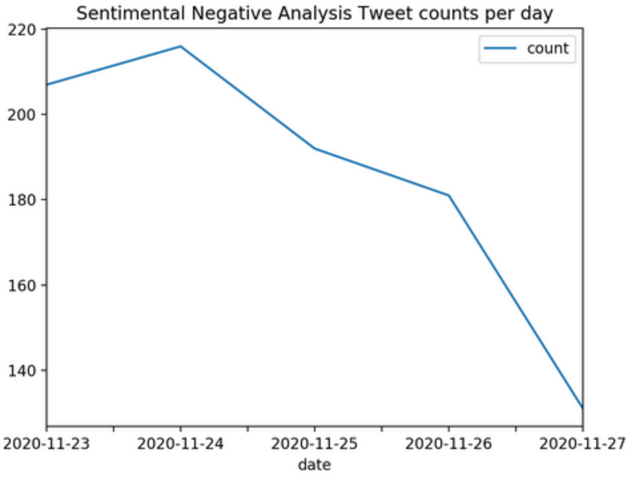

Figure 7: Sentimental Negative Analysis Tweet Count

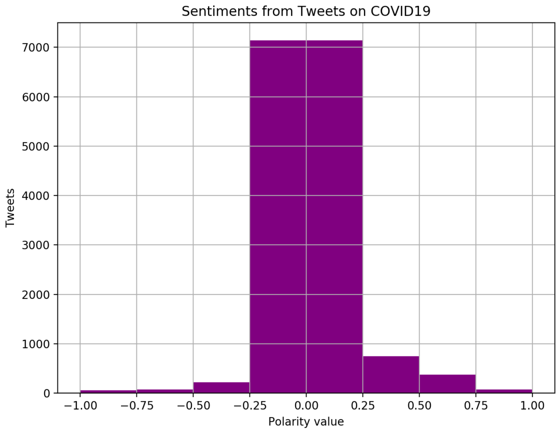

Figure 8: Sentiments on Tweets from COVID

8.2 Reddit data

Here, after applying sentiment analysis we get the following plots. The first one specifically highlights the polarities for one day of data, the x-axis represents the time where we see an absolute similar trend when it comes to all the polarities. But when we look at the wordcloud of the same sample data we can observe that most of the words correspond to the neutral and negative numbers. Also this in turn shows how the neutral posts involve discussions and awareness of the epidemic. Every day on same hour: Similar insights are drawn from the plot below showing the number of posts every hour for all the 7 days, though this sample includes posts for 6 hours of each day. The plot that shows data for all the days separately consists of average of 140 negative, 607 neutral and 3 positive posts making it identical.

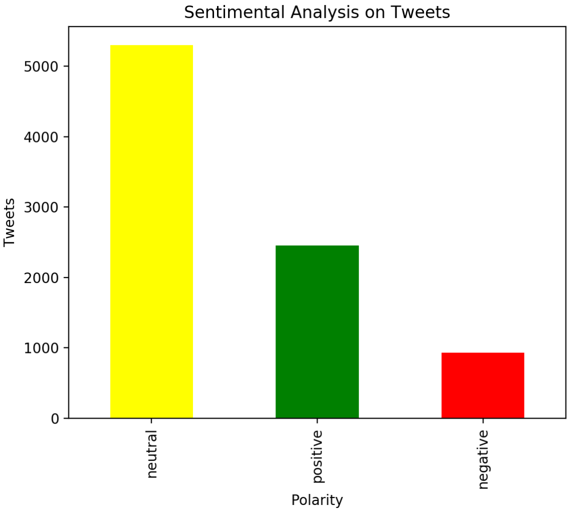

Figure 9: Sentiment Analysis on Tweets

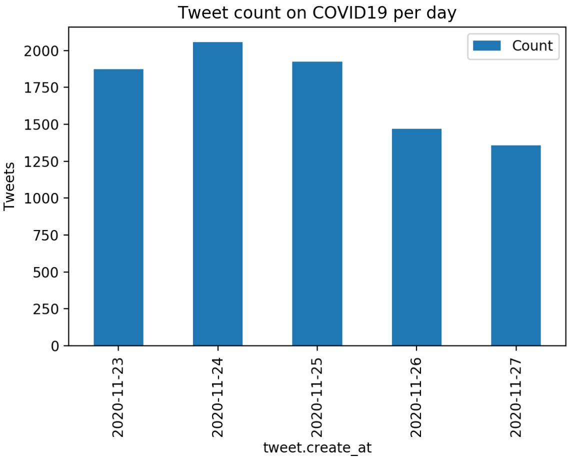

Figure 10: Tweet count on COVID19 per day

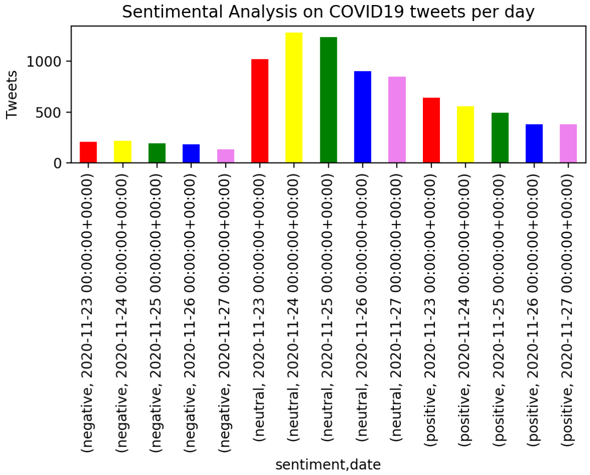

Figure 11: Sentiment Analysis on COVID19 tweets per day

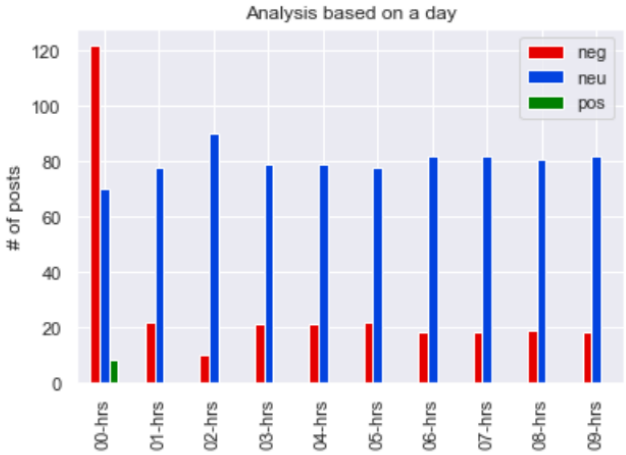

Figure 12: Sentiment Analysis on November 1

9 CONCLUSION

Finally, after seeing the analysis done on both the sides in different ways, it seems that both the social media websites correspond to the same results.

In both reddit and twitter, we observed a higher number of posts being neutral, which points in the direction of people contributing to awareness discussion of the epidemic. For the negatives, it is the second highest polarity for both where in depth analysis revealed that most of the posts in this had HTML-tags urls in these. For positives, in both the social medias we saw a very less number of posts in general and also it was also observed some of the positive posts were classified as neutral by the classifier.

Further, this analysis could be well extended to draw many other insights sentiments from the existing dataset. Also, with more manual labelling of data in hand we could have a more accurate classifier.

10 FUTURE WORK

Due to the lack of time and the computational process, many aspects have been left for the future work.

- We can choose a text blob using Naives Bayes model and the lexicon-based algorithms for the future work.
- This research application can be used to discover sentiment emotions for the future similar cases.

11 ACKNOWLEDGMENTS

We would like to thank Prof. Jeremy Blackburn for guiding and helping us come up with this project and providing his insights on the various steps that could help us plan this project.

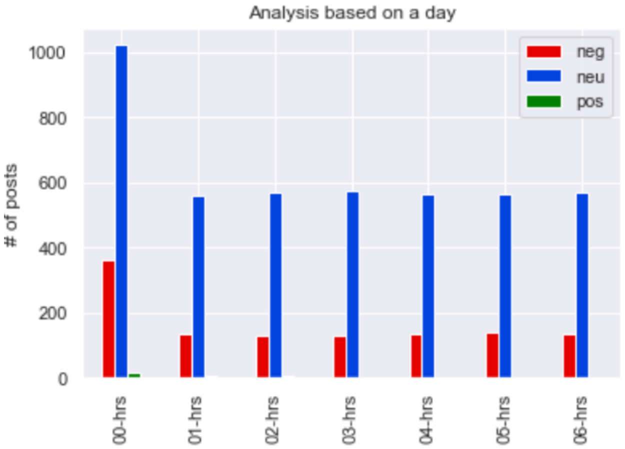

Figure 13: Sentiment Analysis at a particular hour everyday

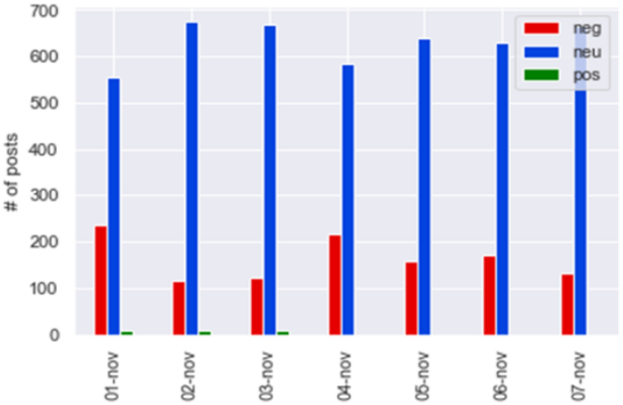

Figure 14: Sentiment Analysis on everyday basis

REFERENCES

[1] [n.d.]. https://www.reddit.com/dev/api/

[2] [n.d.]. https://developer.twitter.com/en/docs/twitter-api

[3] [n.d.]. https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/

[4] P. He. 2020. *Study on Epidemic Prevention and Control Strategy of COVID -19 Based on Personnel Flow Prediction.* IEEE, Zhuhai, China. 688–691 pages. https://doi.org/10.1109/ICUEMS50872.2020.00150

[5] Siddhartha Bhattacharyya Jan Platos Rajib Bag Aboul Ella Hassanien Koyel Chakraborty, Surbhi Bhatia. 2020. *Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media.* Springer. https://doi.org/10.1016/j.asoc.2020.106754

[6] Bishwo Prakash Pokharel. [n.d.]. *Twitter Sentiment Analysis During Covid-19 Outbreak in Nepal.* https://doi.org/10.2139/ssrn.3624719

[7] A. K. Maan S. Panja and A. P. James. 2020. *Vilokana - Lightweight COVID19 Document Analysis.* IEEE, Springfield, MA, USA. 500–504 pages. https://doi.org/10.1109/MWSCAS48704.2020.9184598

[8] S. Bhattacharyya S. Roy, M. N. Pal and S. Lahiri. 2020. *Implementation of an Informative Website – Covid19 Predictor Highlighting COVID-19 Pandemic Situation in India.* IEEE, Vancouver, BC, Canada. 1–6 pages. https://doi.org/10.1109/IEMTRONICS51293.2020.9216352.