Epidemic Outbreak Prediction Using AI

Mohammed Mehran[1], Austin George[2], Umesh Yadav[3], R. Logeshwari[4]

[1,2,3]Student, SRMIST Chennai
[4]Assistant Professor, Dept. of CSE, SRMIST Chennai, Tamil Nadu, India

---***---

Abstract - *An epidemic outbreak happens when a disease spreads rapidly in a short interval of time in a particular region. An outbreak may occur in one community or even extend to several countries. It can last from days to years. Therefore, it is imperative to contain the epidemic as quickly as possible. There is an ever increasing need for intelligent models for predicting diseases and preventing its spread in an area globally. Here we present an approach to predict the epidemic prone area using the potential of Prophet model and Data Visualization. The purpose of our model is to contain the further spread of a particular ongoing epidemic. There are two approaches towards our model : societal approach and computational epidemiology approach. Societal approach consists of gathering data from social media platforms and analyzing the public awareness regarding the epidemic. Computational Epidemiology approach consists of analyzing and predicting the future trends based on medical data sets.*

Key Words: Twitter API, COVID-19, Prophet Model, Sentiment Analysis, Data Visualization.

1.INTRODUCTION

Data analysis and Artificial Intelligence have proved their importance in the areas of research when taken into account individually, but when combined together it opens new possibilities and enhanced outcomes in the field of predictive modelling. Researchers are taking the advantage of the power of analytics to analyse the data and obtain an outcome out of it. History stands witness to deadly Epidemics which wiped out million of human lives and also caused worldwide economic depression. Therefore, it is of vital importance that an epidemic be contained as soon as it begins. A lot of time and resources are spent on collecting and maintaining about the ongoing epidemic, a part of of this problem can be overcome by using social media platforms to collect most recent data.

An epidemic can be eliminated by two possible ways 1: find the cure for ongoing epidemic , 2: containment of further spread of the epidemic until a cure is found or it is eliminated by itself. The methodologies that currently exists are inefficient because they require large amount of continuous data and require a lot of time to compute results, which may exceed the incubation period of the epidemic. Till date there are no studies that combines both societal and computational epidemiology. Existing systems are inefficient in working with time series data. Also, they make use of twitter data only for prediction which requires numerous algorithms and techniques like SVM, LSTM, Naive Bayes and NLP. The results obtained are not reliable enough to be taken into consideration.

This paper focuses on containing the further spread of the epidemic. Twitter has been used as a social media platform for this paper. Tweets are fetched from the twitter API using specific keywords pertaining to the epidemic, followed by text cleaning and sentiment analysis. Furthermore, this data is stored in the form of data-sets and used for data visualization later on. Another effective method to contain the epidemic is by predicting future trends using predictive models. Data-sets from certified bodies such as WHO are used to analyse the current trends and predict future outcomes. The rest of the paper is structured to highlight Related Work, Proposed System, Data Extraction, Data Visualization, Predictive Modeling, Conclusion and Future Scopes.

2 Related Works

Social media platforms like twitter can be used to obtain real time data and aids in analyzing specific trends in different fields like health care, businesses, market trends. Devin Francis Gaffney [1] proposed the idea that in addition to being a versatile communications platform to users around the globe, Twitter is also an excellent source of current information. He explains about the different twitter APIs available to researchers. Our study uses one of the specified twitter API.

Polarity and Subjectivity are two important factors in performing sentiment analysis using textblob in python. It focuses on some common areas like parts of speech, Noun and phrase from text, text classification, sentiment analysis etc. [2] as proposed by Arpana Alka. Tokens passed to textblob can be treated as stings in python that can perform natural language processing. The sentiment analyzer returns a tuple of the form sentiment (polarity, subjectivity), where polarity ranges from [-1.0, 1.0] and the range of subjectivity is from [0.0, 1.0].

Interactive approach enhances the outcome of data analysis significantly. It also improves the comparative analysis drastically[3].Andreas BUJA, Dianne COOK, and Deborah F. SWAYNE proposed an approach of interactive data visualization based on certain analytic tasks such as making comparisons. This study makes use of plotly which is an open source and interactive python graphing library. Different types of charts can be plotted such as statistical charts, financial charts, scientific charts etc.

It is difficult to manage time component in traditional prediction approaches, whereas time series forecasting techniques also takes time component into account along with other factors.

The outcomes produced by time series forecasting are much more accurate than the outcomes produced by traditional prediction methodologies. The results concluded by James W. Taylor[4] in his study indicated that strong potential for the use of time series forecasting for predicting the future outcomes.

3. PROPOSED SYSTEM

Our proposed model eliminates existing drawbacks and is more reliable and accurate as it uses twitted data for measuring public awareness only and for prediction, only dataset from certified bodies is used. It uses two approaches: societal approach and computational epidemiology approach. Societal approach consists of gathering data from social media platforms and analyzing the public awareness regarding the epidemic. First data is collected from twitter API, preprocessed and finally sentiment analysis is performed. The results obtained are stored in a dataset. The obtained dataset contains fields such as used_id, original text, cleaned text, retweet_count, location, hashtags, sentiment, polarity, subjectivity and had 1000 entries. This dataset can be used for further analysis. Computational Epidemiology approach consists of analyzing and predicting the future trends based on medical datasets. Dataset provided by John Hopkins University is used for data visualization as well as for predictive analysis. This dataset consists of 10984 entries dating from January 22 2020 to April 1 2020 which accounts for a total of 71 days. The dataset contains fields such as Observation date, State, Country, Last update date, Confirmed deaths, Recovered deaths. Prophet model is used for training, testing and prediction.

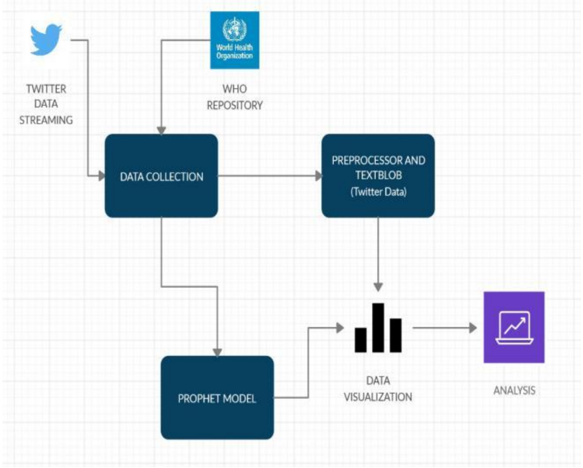


Fig 1. Architecture Diagram

3.1. Data Extraction

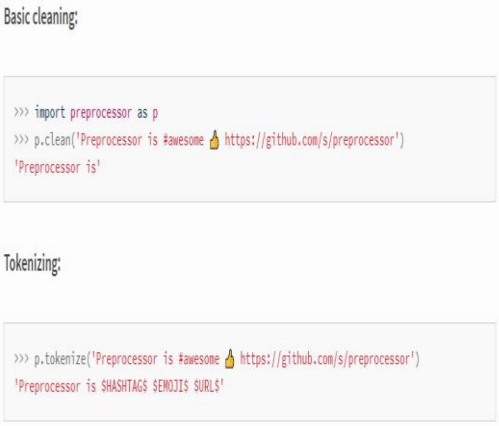
In this section, we discuss about the data collection and data extraction. Data collection basically deals with gathering all the possible information, whereas data extraction deals with retrieving relevant details form the collected data.

Twitter is a micro blogging system that allows users to send and receive short posts called tweets,it is an information network and communication mechanism that produces more than 200 million tweets a day and thus can also be treated as a reliable source for collecting data on specific areas which can further be used for data analysis and prediction. Twitter API allows users to use different features of twitter platform like posting a tweet or finding tweets that contain a word, etc. without the need of visiting the website interface. The working of twitter API can be fully automated.

The World Health Organization is a particular office of the United Nations liable for worldwide general well being. It closely monitors all epidemics and pandemics and issues advisories and collect related data from around the world. This data is organized by John Hopkins University and is publicly available and can be used for research purposes. In the epidemiological approach of this study we make use of this dataset. Data extraction module of this study uses both twitter data as well as data provided by John Hopkins University.

3.1.1. Twitter Data Extraction

Tweets collected from twitter contains a lot of noise such as symbols, emoticons, hash-tags, words other than English. These are unwanted parts of tweet which can hamper the data analysis part, therefore they need to be removed before data extraction.[5] For this purpose inbuilt pre-processor python library is used. It supports cleaning, tokenizing and parsing of strings passed to its function. Its main features are removing URLs, hash-tags, mentions, reserved words, emoticons, smileys etc.



Basic cleaning:

```
>>> import preprocessor as p
>>> p.clean('Preprocessor is #awesome 👍 https://github.com/s/preprocessor')
'Preprocessor is'
```

Tokenizing:

```
>>> p.tokenize('Preprocessor is #awesome 👍 https://github.com/s/preprocessor')
'Preprocessor is $HASHTAG$ $EMOJI$ $URL$'
```

Fig -2: Functioning of Pre-Processor

The input data including tweet id, tweet content, created at, hash-tags, locations, re-tweet counts, cleaned tweet, source are all obtained using Twitter API. Twitter API cursor is called along with parameters such as keyword, count, include_retweets, duration, number of pages.

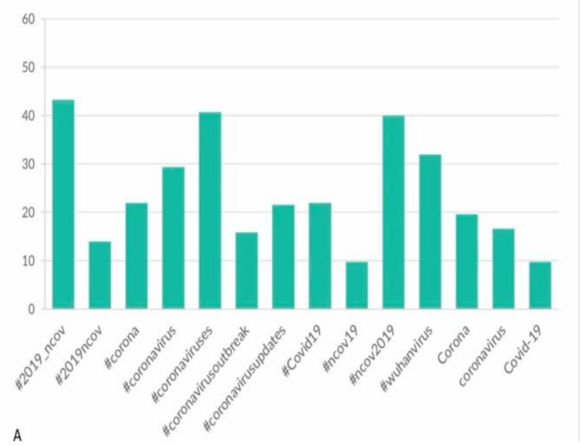
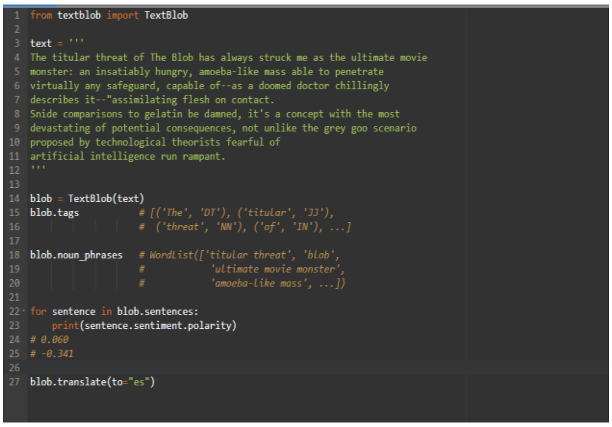


Fig-3: Graph summarizing top hash-tags used in tweets for COVID-19

3.1.1.1. Sentiment Analysis

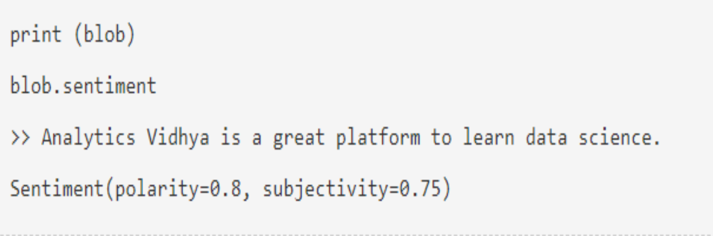
Sentiment analysis is procedure of computationally recognizing and classifying sentiments communicated in a bit of content, particularly so as to decide if the author's behaviour towards a specific theme, item, and so on is positive, negative, or impartial.[6]Singh, Rameshwer & Singh, Rajeshwar & Bhatia, Ajay (2018) Sentiment. This study makes use of textblob, which is a python library for processing textual data. It gives a basic API for jumping into regular natural language processing assignments, for example, grammatical feature labeling, noun phrase extraction, sentiment analysis, order, interpretation, etc.



```
from textblob import TextBlob

text = '''
The titular threat of The Blob has always struck me as the ultimate movie
monster: an insatiably hungry, amoeba-like mass able to penetrate
virtually any safeguard, capable of--as a doomed doctor chillingly
describes it--"assimilating flesh on contact.
Snide comparisons to gelatin be damned, it's a concept with the most
devastating of potential consequences, not unlike the grey goo scenario
proposed by technological theorists fearful of
artificial intelligence run rampant.
'''

blob = TextBlob(text)
blob.tags              # [('The', 'DT'), ('titular', 'JJ'),
                       #  ('threat', 'NN'), ('of', 'IN'), ...]

blob.noun_phrases      # WordList(['titular threat', 'blob',
                       #            'ultimate movie monster',
                       #            'amoeba-like mass', ...])

for sentence in blob.sentences:
    print(sentence.sentiment.polarity)
# 0.060
# -0.341

blob.translate(to="es")
```

Fig-4: Functioning of textblob

Two components of textblob sentiment analysis are taken into account i.e., Polarity and Subjectivity. Polarity consists of floating point values which lies between the range [-1, 1], where 1 refers to a positive statement and -1 refers to a negative statement. The range of Subjectivity lies between [0, 1] , where 0 refers to objective statements which are factual in nature and 1 refers to statements containing personal opinion, emotion or judgement.



```
print (blob)

blob.sentiment

>> Analytics Vidhya is a great platform to learn data science.

Sentiment(polarity=0.8, subjectivity=0.75)
```

Fig-5: Sentiment analysis function

Here the polarity is 0.8, which means that the statement is positive and subjectivity is 0.75 which means that it is mostly a public opinion and not factual information.

3.1.2. Covid-19 Data Extraction

Dataset is provided by John Hopkins University is used for epidemiological approach of this study, certain factors have been taken into consideration form this data such as number of deaths, number of active cases, dates, location, number of recovered cases. The data has also been categorized into global and mainland china. John Hopkins University organize the data collected by WHO periodically hence latest trends can be analyzed. From the dataset we extracted the total number of confirmed cases corresponding to their observation dates. The dataframe thus created contains a total number of 71 days stretching from January 22, 2020 to April 1, 2020.

3.2. Data Visualization

Data Visualization is the graphical portrayal of data and information. By utilizing visual components like diagrams, charts, and maps, it is an instruments that gives an open method to see and get patterns, anomalies, and examples in information. Visualization is an inexorably key device to understand the trillions of columns of information produced each day. Data Visualization assists with recounting stories by curating information into a structure more clear, featuring the patterns and anomalies. A decent visualization recounts to a story, expelling the commotion from information and extracting the helpful data.

3.2.1.Visual Representation

Visual representation is any technique for creating images, diagrams, or animations to communicate a message. Visualization through visual imagery has been an effective way to communicate both abstract and concrete ideas since the dawn of humanity.

Plotly (the Python library) uses declarative programming which means writing code to describe what to make rather than how to make it. Basic framework and end goals are provided by the users and plotly figures out the implementation details. In practice, this means less effort spent building up a figure, allowing user to focus on what to present and how to interpret it. A number of different trends are visually plotted such as polarity vs location plot, number of confirmed cases vs locations and date, number of deaths vs location and date.

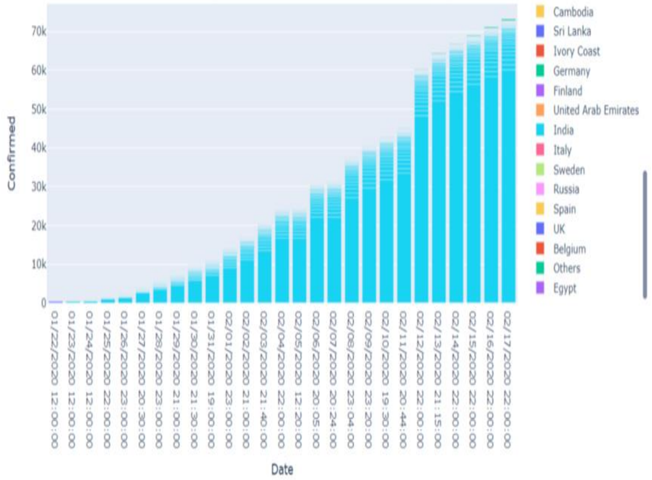


Fig-6: Confirmed bar plot for each country

3.2.2. Data Analysis

Data analysis is the way of gathering and arranging information so as to make accommodating inferences from it. The procedure of data analysis utilizes diagnostic and intelligent thinking to pick up data from the information. The fundamental reason for data analysis is to discover significance in information with the goal that the inferred information can be utilized to settle on educated choices. Data analysis has been performed throughout our study including societal and epidemiological approaches.

In societal approach data analysis has been performed in observing sentiment analysis trends based on specific locations. From the graph obtained by plotting the twitter data set the variations in polarity and subjectivity of different locations can be inferred, which further tells about the level of public health awareness in the corresponding locations[7].

In Epidemiological approach the relevant data from covid-19 data sets are taken apart and plotted visually. This gives an insight into the various trends of how the epidemic is progressing. Studying these trends allows us to come up with actions to be undertaken to effectively contain the epidemic.

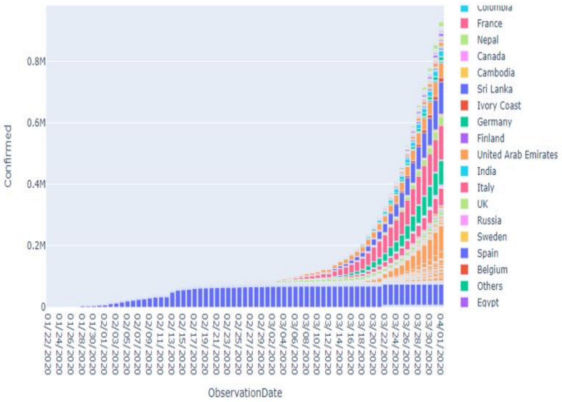


Fig-7: Confirmed bar plot in mainland china

Form the aforementioned bar plot a number of different trends can be inferred such as, name of province, date and time, number of confirmed cases, number of deaths, number of recovered cases. The same set of trends can be observed for various nations also.

3.3. Predictive Modeling

Predictive modeling is a process that uses data analysis and probability to forecast outcomes. Each model is comprised of various indicators, which are factors that are probably going to impact future outcomes. When information has been gathered for important indicators, a measurable model is defined.

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles out liners well.

Prophet uses a decomposable time series with three main model components: trends, seasonality, and holidays [8].

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

Here,

g(t) = trends

s(t) = seasonality

h(t) = holidays

εt = outliers

From the extracted dataframe which consists of total number of confirmed cases corresponding to its dates, it can be inferred that their is a increasing trend in the number of confirmed. A custom seasonality of 14 days has been used in this model since the incubation period of COVID-19 is 2 weeks.

The obtained dataframe has been split into training data and testing data. Out of 71 days 61 continuous days from January 22, 2020 to March 22, 2020 which is almost 85% of total data is used for training prophet model, whereas remaining 10 days from March 23, 2020 to April 1, 2020 which is almost 15% of the total data is used for testing prophet model. The model is then fitted for predicting the number of confirmed cases for next 30 days.

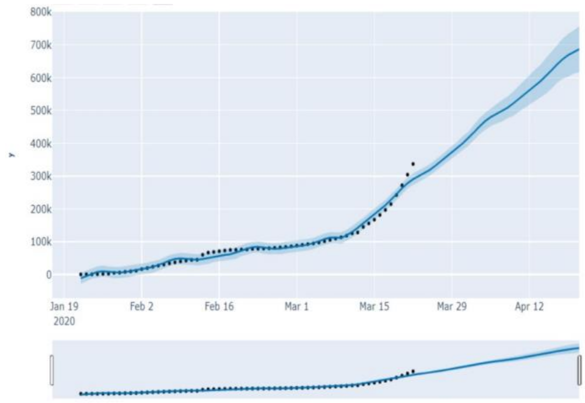


Fig-8: prediction of future confirmed cases

Majority of time series forecasting models requires an exhaustive comprehension of how the basic time arrangement models work.[9] Some of the required parameters are: maximum orders of differencing, moving average components, and the auto-regressive components, adjusting these components is hard to acquire and scale. whereas, the Prophet package provides intuitive parameters which are easy to tune.

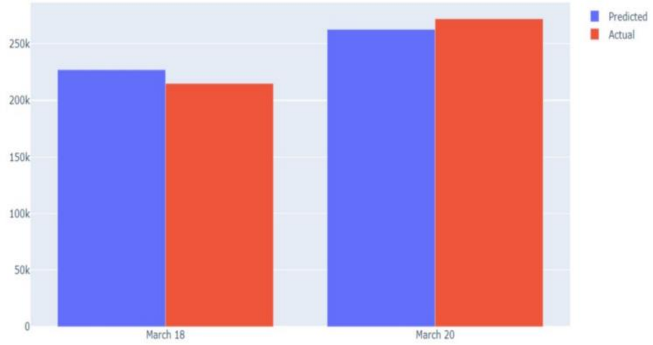
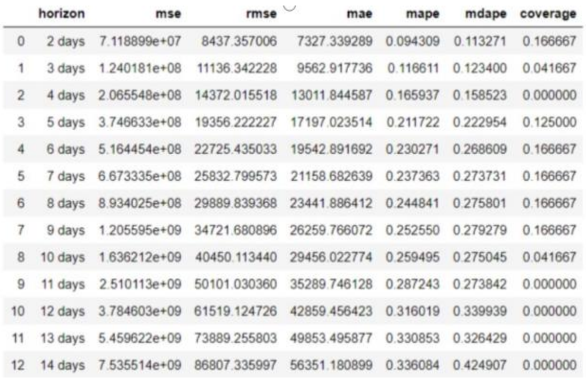


Fig-9: Predicted data and actual data

The acquired outcomes shows trends over a period of two days, it can be inferred from the graph that the predicted trend closely follows the actual trend with a high level of accuracy.

4. RESULTS

The data has been plotted using data visualization techniques, which helps us in understanding the varying trends better. We can also perform a comparative study across different time lines and locations as the graphs plotted are interactive in nature. A dataframe is created by extracting confirmed cases data from COVID-19 dataset to train and test the prophet model. Following this, prediction was performed on various dates across a 2 week time line and the performance matrix for the same is shown in Fig 10. Different error rate parameters can studied from the performance matrix. We can see that for a time period of 2 days the mean absolute percentage error is 0.094 whereas, for a time period of 7 days it increases to 0.237. Furthermore, we see that for a time period of 14 days the error rate increases to 0.336. This suggests that the predictions are more reliable for a shorter time duration. The model became more accurate when more data is fed to the model.



	horizon	mse	rmse	mae	mape	mdape	coverage
0	2 days	7.118899e+07	8437.357006	7327.339289	0.094309	0.113271	0.166667
1	3 days	1.240181e+08	11136.342228	9562.917736	0.116611	0.123400	0.041667
2	4 days	2.065548e+08	14372.015518	13011.844587	0.165937	0.158523	0.000000
3	5 days	3.746633e+08	19356.222227	17197.023514	0.211722	0.222954	0.125000
4	6 days	5.164454e+08	22725.435033	19542.891692	0.230271	0.268609	0.166667
5	7 days	6.673335e+08	25832.799573	21158.682639	0.237363	0.273731	0.166667
6	8 days	8.934025e+08	29889.839368	23441.886412	0.244841	0.275801	0.166667
7	9 days	1.205595e+09	34721.680896	26259.766072	0.252550	0.279279	0.166667
8	10 days	1.636212e+09	40450.113440	29456.022774	0.259495	0.275045	0.041667
9	11 days	2.510113e+09	50101.030360	35289.746128	0.287243	0.273842	0.000000
10	12 days	3.784603e+09	61519.124726	42859.456423	0.316019	0.339939	0.000000
11	13 days	5.459622e+09	73889.255803	49853.495877	0.330853	0.326429	0.000000
12	14 days	7.535514e+09	86807.335997	56351.180899	0.336084	0.424907	0.000000

Fig-10: Performance Matrix for prophet model

5. CONCLUSION AND FUTURE SCOPE

Epidemic is an unanticipated surge in the number of cases of an infectious disease over a geographic region, which cause huge damage to life and the global economic infrastructure. It takes decades to recover from an epidemic and requires enormous amount of time and resources. The foremost measure in tackling an epidemic outbreak is its containment. In such scenarios time is of vital importance as any delay may lead to exponential destruction of life and economy. Therefore, it is necessary that governments and health ministries around the world have to stay one step ahead in determining the possible rate of escalation of the epidemic. Most nations across the world are underprepared to tackle such sudden outbreaks. Predictive modelling brings a revolutionary change as it can be first line of defence in containing an epidemic in its early stages.

In this study we have made use of prophet model to predict future trends of an epidemic. A comparative study suggests that prophet model is one of the most accurate, reliable and easy to use as compared to other time series based forecasting models. Our study concluded that for a period of 2 days the mean absolute percentage error was 0.094 and the accuracy was almost 91%, whereas for a period of 14 days the mean absolute percentage error increased to 0.336 and the accuracy decreased to 77%. This suggests that prophet modelling is better suited for short periods of time and changing trends. The results obtained from prophet model can be used by governments and health ministries around the globe and can be analyzed further to come up with effective containment strategies.

This proposed model can further be extended and made more accurate by including other factors and modules, such as monitoring local traffic and global airline data. Extensive twitter data can be used along with better NLP techniques and machine learning algorithms to gain deeper insight into the ongoing epidemic. The whole models can be fully automated with features like periodic extraction, analysis and prediction.

REFERENCES

[1] Devin Gaffney and Cornelius Puschmann (2015). Data collection on Twitter.
[2] Nimai Chand Das Adhikari, Jitendra kumar Kushwaha, Arpana Alka, Ashish kumar Nayak(2018).Sentiment classifier and analysis for epidemic prediction.
[3] Andreas Buja, Dianne Cook and Deborah F. Swayne(1996). Journal of Computational and Graphical Statistics.
[4] Taylor, James. (2008). A Comparison of Univariate Time Series Methods for Forecasting Intraday Arrivals at a Call Center. Management Science. 54. 253-265. 10.1287/mnsc.1070.0786.
[5] Chand, Nimai & Das Adhikari, Nimai & Alka, Arpana & Kurva, Vamshi Kumar & Nayak, Hitesh & Kushwaha, Jitendra & Nayak, Ashish & Nayak, Sankalp & Shaj, Vaisakh & Rishav, Kumar. (2018). Epidemic Outbreak Prediction Using Artificial Intelligence. International Journal of Computer Science and Information Technology.10. 10.5121/ijcsit.2018.10405.
[6] Singh, Rameshwer & Singh, Rajeshwar & Bhatia, Ajay. (2018). Sentiment analysis using machine learning techniques to predict outbreaks and epidemics.
[7] Healey, Christopher. (1996). Choosing Effective Colours for Data Visualization. 10.1109/VISUAL.1996.568118.
[8] https://www.analyticsvidhya.com/blog/2018/05/generate-accurate-forecasts-facebook-pro phet-python-r/
[9] Tyralis, Hristos and Georgia Papacharalampous. "Large-scale assessment of Prophet for multi- step ahead forecasting of monthly streamflow." (2018).
[10] K. Krishna Rani Samal, Korra Sathya Babu, Santosh Kumar Das, and Abhirup Acharaya. 2019. Time Series based Air Pollution Forecasting using SARIMA and Prophet Model. In Proceedings of the 2019 International Conference on Information Technology and Computer Communications (ITCC 2019). Association for Computing Machinery, New York, NY, USA, 80–85. DOI:https://doi.org/10.1145/3355402.3355417.