IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 5, May 2020

World Wide Covid 19 Outbreak Data Analysis and Forcasting Using Fbprophet

Unnimaya Rajan

Department of Computer Science, Sree Narayana Guru Institute of Science and Technology, Ernakulam, India

Abstract: Nowadays, sustainable development is considered a key concept and solution in creating a promising and prosperous future for human societies. Nevertheless, there are some predicted and unpredicted problems that epidemic diseases are real and complex problems. Hence, in this research work, a serious challenge in the sustainable development process was investigated using the classification of confirmed, death, and recovery cases of COVID-19 as one of the epidemic diseases. The inception of the coronavirus was the fish market of Wuhan city, Hubei territory in China. The instances of somebody experiencing COVID-19 can be followed back to the finish of December 2019 in China. This is the most irresistible malady and spread worldwide inside a quarter of a year after the main case announced. The complete name of the coronavirus is serious intense respiratory disorder SARS-CoV. Thus, the data mining predictive modelling method of data handling and predictive or forecasting the spread of COVID-19 virus. The full name of the coronavirus is severe acute respiratory syndrome SARS-CoV. It spread on humans as well as animals and infected around 213 countries and territories with 6,399,977 confirm cases and 378,065 deaths till 2 June 2020. This study introduces the spreading pattern of COVID-19 in the top ten infected countries. After China, European countries are the most infected ones. In this study, data was analysed on the attributes confirmed, active, recovered and death cases and the next 14 days outbreak prediction. This research work mainly works on worldwide COVID 19 data analysis and forecasting by using fbprophet. Prophet it is a python library package used for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonally, plus holiday's effect. Being the COVID-19 pandemic during a very short time span, it is very important to analyze the trend of these spread and infected cases. The real-time datasets are used in this project and it plotted on the worldwide map. Firstly analyse the no. of infected people around the world, the no. of people who died, and the no. of recovered people in the world on the basis of real-time data. Secondly predicting and visualizing the Number of COVID-cases in India using the Fbprophet algorithm.

Keywords: Analysis and Visualization, Time Series, Fbprophet, COVID-19.

I. INTRODUCTION

Coronavirus is certainly not another infection yet the serious intense respiratory disorder coronavirus, SARS CoV is the new infection of the family Coronaviridae cases found in China. World Health association formally renames SARS-CoV (SARS-nCoV) or Novel coronavirus as COVID-19 on 11 February 2020 (WHO, 2020; COVID-19, 2020). Bats are the wellspring of COVID-19 infection and spread in people just as a warm-blooded creature. Brooding of coronavirus is 2 to 14 days, talk about the sign that the transmission of COVID-19 happens during the hatching time frame. Being the COVID-19 pandemic during a very short time span, it is very important to analyze the trend of these spread and infected cases. This research presents the medical perspective of COVID-19 towards the epidemiological triad and the study of state-of-the-art. The main aim of this research is to present different predictive analytics techniques available for trend analysis, different models and algorithms, and their comparison. The Core Data Science team at Facebook recently published a new method called Prophet, in this research using the Fbprophet algorithm which enables data analysts and developers alike to perform forecasting at scale in Python 3. The prediction of COVID-19 using the Prophet algorithm indicating more faster spread in the short term. These predictions will be useful to government and healthcare communities to initiate appropriate measures to control this outbreak in time. The calculation utilized Fbprophet to anticipate the flare-up of COVID-19 in India on an everyday base and discovered consistency results with recuperation, affirmed, and passing cases. Social separating and lockdown is the weapon to battle with COVID-19. At that point anticipate the affirmed, recuperation, and passing cases in the overall COVID-19 cases for the following 14 days to utilizing the Fbprophet calculation. The examination has been led on an example of information gathered from the constant informational collection and tests with considers the instances of coronavirus around the world, top tainted nations of the world date-wise by using the python language in Google colab. It is a research paper shown the countries wise confirmed, death, and recovery cases on the world wild Map. This spreading example of COVID-19 of top nations, for example, the United States of America, Italy, China, Spain, Germany, and Iran date-wise. It is broke down that the nations' astute and state savvy information for the better understanding flare-up of COVID-19. It is indicated the nations' savvy affirmed cases on the topical Map. Anticipate the future affirmed case in the USA, China, and Rest of the World, discover the precision of expectation. After all, using Machine Learning ALgorithms and the use of advanced techniques

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 5, May 2020

like FBProphet, predictive analysis is done and a trend regarding the forecasting of the outbreak of disease is achieved. Finally, this paper concludes with the prediction of COVID-19 using the Prophet algorithm indicating more faster spread in the short term. These real-time predictions will be useful to government and healthcare communities to initiate appropriate measures to control this outbreak in time.

II. PROPOSED SYSTEM

I propose a prediction and analysis model to predict the Outbreak of COVID-19 on the overall world on the three basic things. Those are recovery, confirmed and death cases using data mining techniques for the coming 14 days. For the prediction purposed I have proposed to the time series prediction data mining algorithm to forecast for the coming 14 days and analysis everything graphically and compare and contrast the current status of the COVID-19 and predicted values. To forecast the outbreak of the COVID-19 virus I will use the Fbprophet open-source python software which is developed by Facebook data scientists. Facebook data scientists purposely they have developed for such types of unseasonal disease or virus based on the given dataset to predict or forecast for the future and show them clearly mention the impacts.

III. DATA MINING TECHNIQUES

Data mining plays an important role in various fields such as artificial intelligence, machine learning, and database systems. Systems that help to discover hidden patterns and are very useful for disease prediction. Data mining is the technique in which useful information is extracted from the raw data. The data mining is applied to accomplish various tasks like clustering, prediction analysis, and association rule generation with the help of various Data Mining Tools and Techniques. In the approaches of data mining, clustering is the most efficient technique which can be applied to extract useful information from the raw data. These research works mainly focus on the design and implementation of a COVID-19 analysis and prediction of the confirmed, recovered and death cases for the coming 14 days on the whole world. The predictive data-mining model predicts future outcomes based on past records present in the d with known answers. Data mining will help figure out the future credit risk of the applicant and predict future credit history of the applicant by using past data. Classification is known as the procedure used to locate a model that best suits identified data sets or ideas. The model helps predict the class of objects when class labels are not available. The most widely used for data mining prediction is time series forecasting methods i. It is also one of the most popular models in traditional time series forecasting and is often used as a benchmark model for comparison with any other forecasting method. It is often difficult to identify a forecasting model because the underlying laws may not be clearly understood. In addition, hydrological time series may display signs of seasonality and nonlinearity in which traditional linear forecasting techniques are ill-equipped to handle, often producing unsatisfactory results.

IV. TIME-SERIES DATA MINING

A time series is a sequence of data points recorded at specific time points most often in regular time intervals. Every organization generates a high volume of data every single day is it sales figures, revenue, traffic, or operating cost. Time series data mining can generate valuable information for long-term business decisions, yet they are underutilized in most organizations. Forecasting future values using historical data is a common methodological approach from simple extrapolation to sophisticated stochastic methods. Predictive analytics: Advanced statistical analysis such as panel data models rely heavily on multi-variant longitudinal datasets. These types of analysis help in business forecasts, identify explanatory variables, or simply help understand associations between features in a dataset. In my context, this predictive analysis is in predicting the COVID-19 outbreak and shows the gap or impact the whole world. Prediction models can be useful in forecasting different future scenarios for time series data.

V. METHODOLOGY

The general methodology of the given study, to perform the spreading pattern of COVID-19, the real-time data collected from John Hopkins University and other sources, which contains 213 countries and territories with 34442 rows and 8 columns. These eight columns are renamed as states, country, longitude, latitude, date, confirmed, recover, and deaths. Few countries give their dataset wise i.e. United States, Brazil, Russia, United Kingdom, Spain, Italy, and others, but 9882 rows of state columns are NaN (Not a number or no value in cell), replace these NaN, with empty white space. This data undergoes various steps of pre-processing which makes it more sensible. Then, for data analysis calculate the active case worldwide by subtracting recover and death cases from the confirmed cases. Ranking changes with every upcoming day due to coronavirus infection. In this study, overall data is divided into 11 parts i.e. top ten infected countries till 2 June 2020 namely United States, Brazil, Russia, United Kingdom, Spain, Italy, France, Germany, India, and Turkey which internally contain 183 country data.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 5, May 2020

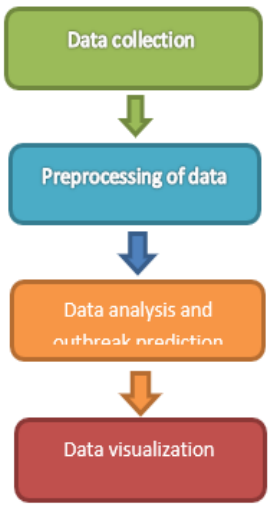

Fig. 1 General Methodology

VI. DATA SOURCE

Data is extracted from verified sources such as John Hopkins University, The sites reported real-time datasets confirmed COVID-19 cases, as well as recovered and deaths for affected countries and regions. The COVID-19 module allows importing and displaying of data related to the 2019 Novel Coronavirus COVID-19 (2019nCoV) from multiple sources. The intention is not only to make it easier to display and update data but to select which data source to use depending on the information being displayed. For example, one data source would be used when displaying country-level data, but another, perhaps more current or accurate data source would be used when displaying data for a specific state, province, or county. There are a lot of official and unofficial data sources on the web providing COVID-19 related data. One of the most widely used dataset today is the one provided by the John Hopkins University's Center for Systems Science and Engineering (JHU CSSE).

Table I Data Set

	Country_Region	Last_Update	Lat	Long_	Confirmed	Deaths	Recovered	Active	Incident_Rate	People_Tested	People_Hospitalized	Mortality_Rate	UID	ISO3
0	Australia	2020-06-02 16:33:06	-25.0000	133.0000	7221.0	102.0	6625.0	494.0	28.362471	NaN	NaN	1.412547	36	AUS
1	Austria	2020-06-02 16:33:06	47.5162	14.5501	16759.0	668.0	15625.0	461.0	186.078788	NaN	NaN	3.986885	40	AUT
2	Canada	2020-06-02 16:33:06	60.0010	-95.0010	93365.0	7404.0	50105.0	35856.0	246.533915	NaN	NaN	7.930167	124	CAN
3	China	2020-06-02 16:33:06	30.5928	114.3055	84160.0	4638.0	79400.0	122.0	5.991416	NaN	NaN	5.510692	156	CHN
4	Denmark	2020-06-02 16:33:06	56.2639	9.5018	11934.0	580.0	10667.0	687.0	206.035686	NaN	NaN	4.860064	208	DNK

VII. DATA VISUALIZATION

Data Mining is utilized to find patterns, peculiarities, and correlation in the huge dataset to make the forecasts utilizing a wide scope of techniques, this extricated data is utilized by the association to build their income, cost-cutting decreasing danger, improving client relationship, and so on though data visualization is the graphical portrayal of the information and data separated from data mining utilizing the visual components like a graph, chart, and maps, data visualization tool, and techniques help in analyzing the massive amount of information and make the decision on top of it. Machine learning makes it easier to conduct analyses such as predictive analysis, which can then serve as helpful visualizations to present. But data visualization is not only important for data scientists and data analysts, but it is also necessary to understand data visualization in any career.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 5, May 2020

Fig 2: Data visualization and analysis diagram

VIII. DATA ANALYSIS

The data gathered from various data sources particularly the above notice data sources will be put away in the data warehouse. At that point, the stored data is preprocessed, and analyzed by using the data mining displaying techniques and visualize it dependent on the given dataset. For the data analysis, python language with package NumPy, Pandas, and Plotly were used. Pandas is an extremely fast and flexible data analysis and manipulation tool and allows us to store and manipulate tabular data. Table 2: shows the top ten infected countries. After China, European countries are the most infected ones. In this study, data was analyzed on the attributes confirmed, active, and recovered and death cases and the next 14 days outbreak prediction.

Table 2 Most 10 affected countries

n 10

n	country	last_update	lat	long_	confirmed	deaths	recovered	active	incident_rate	people_
17	US	2020-06-02 16:33:06	40.000000	-100.000000	1817785.000000	105475.000000	458231.000000	1306985.000000	551.736276	nan
21	Brazil	2020-06-02 16:33:06	-14.235000	-51.925300	526447.000000	29937.000000	211080.000000	285430.000000	247.670523	nan
13	Russia	2020-06-02 16:33:06	61.524000	105.318800	423186.000000	5031.000000	186602.000000	231553.000000	289.983599	nan
16	United Kingdom	2020-06-02 16:33:06	55.000000	-3.000000	277738.000000	39127.000000	1221.000000	237390.000000	409.124096	nan
18	Spain	2020-06-02 16:33:06	40.463667	-3.749220	239932.000000	27127.000000	150376.000000	62429.000000	513.171027	nan
10	Italy	2020-06-02 16:33:06	41.871900	12.567400	233515.000000	33530.000000	160092.000000	39893.000000	386.218888	nan
93	India	2020-06-02 16:33:06	20.593684	78.962880	205147.000000	5782.000000	99350.000000	100015.000000	14.865677	nan
6	France	2020-06-02 16:33:06	46.227600	2.213700	189348.000000	28836.000000	68558.000000	91954.000000	290.083978	nan
7	Germany	2020-06-02 16:33:06	51.165691	10.451526	183879.000000	8563.000000	166570.000000	8746.000000	219.468062	nan
22	Peru	2020-06-02 16:33:06	-9.190000	-75.015200	170039.000000	4634.000000	68507.000000	96898.000000	515.709675	nan

Table 2 shows the worldwide confirmed recovered, death and active cases which depicts on 2 June 2020.

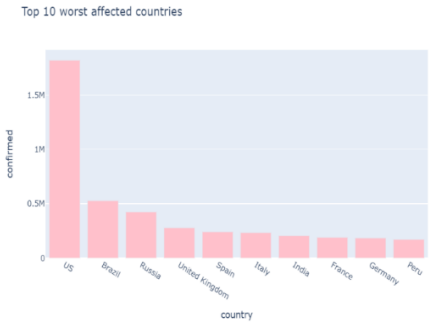

Fig 3: Worldwide confirmed cases

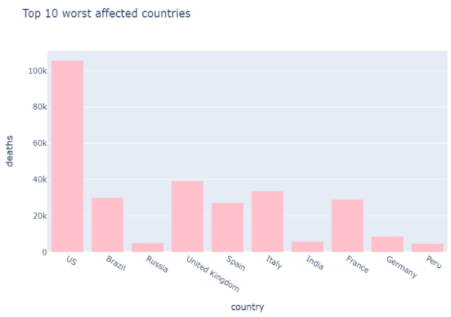

Fig 4: Worldwide death cases

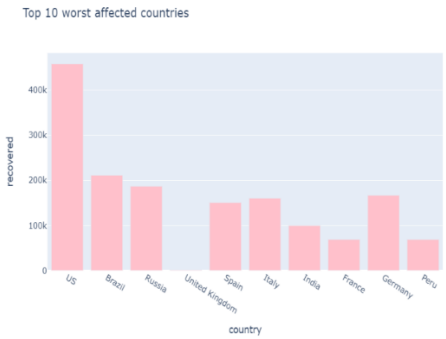

Fig 5: Worldwide recovery cases

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 5, May 2020

The figs 3, 4 and 5 are shows the worldwide confirmed, death and recovery cases. Coronavirus is continuing its spread across the world, with more than six million confirmed cases in 188 countries. More than 375000 people have lost their lives.

Fig 6: Countries wise COVID 19 cases on World map

Fig 6 shows the highlighted area represents the pandemic affected regions and the data of regions where the cursor is placed. This result is gained with the Fbprophet algorithm. Fig 6 shows the real-time map of the world with confirmed, death, and recovery cases in the whole world. The red big circle on the map depicts that these are the COVID 19 more affected countries, and the small red circle depicts, with fewer cases. The interactive thematic map shows the country name, confirmed case, deaths cases, and death rate also. The confirmed and death cases are highest in us, but active and death cases are highest in Italy worldwide.

IX. PREDICTIVE MODELING

The predictive model does the analysis for identifying the patterns observed in historical and transactional data to predictive analytics comprises of several statistical and analytical techniques for developing strategies for the future possibilities of prediction. Predictive analytics and data mining use algorithms to discover knowledge and find the best solutions. Data mining is a process based on algorithms to analyze and extract useful information and automatically discover hidden patterns and relationships from data. In this paper I will predict the future impacts of the COVID-19 virus, the confirmed, recovered and death cases overall India based on the current dataset. There some Data mining perdition and analysis algorithms are there. For this paper, I have proposed the time series prediction algorithm by using fbprophet python library to forecast the estimation of affected people, recovered and deaths for the coming 14 days, or the future assumption of the COVID-19 virus. The prediction models can help in health resources management and planning for prevention purposes. Data mining algorithms and techniques are well-known tools for predictive model development and data analysis.

X. FORECASTING TOTAL NUMBER OF CASES IN INDIA USING FBPROPHET

Machine learning is the subpart of artificial intelligence. It is not a computer programming but a set of rules by using statics function predicts the better output for given data in limited time. The prophet is the time series forecasting algorithm for future prediction and Implemented in Python. Python is the programing language that is used for machine learning and data analysis. It is an open-source software developed by Facebook. It is an adaptive model which uses nonlinear data to predict for yearly, monthly, and daily excluding holiday. Prophet easily handles missing data and outliers of the trend. The prophet is accurate and fast because it used a state-of-the-art platform for statistical modeling that provides forecast in very quickly. COVID-19 is not a seasonal issue because of this thing I have used this algorithm to implement this research. this algorithm was used for prediction of the confirmed case for the India for next 14 days shown in Table 3.

Table 3 CONFIRMED CASES

	ds	y
127	2020-05-28	165386
128	2020-05-29	173491
129	2020-05-30	181827
130	2020-05-31	190609
131	2020-06-01	198370

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 5, May 2020

As we have seen the table-3 this confirmed dataset starts from May 28/2020 to June 01/2020 and ds indicates Date Time stamp and y indicates the values in numeric form.

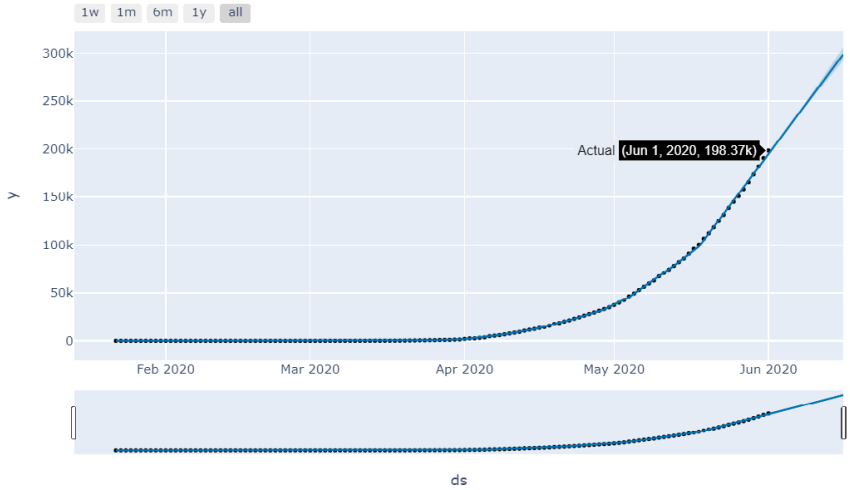

Fig 7: The predicted and original values

The fig-7 indicates the relationship of the original values the blue dotted line and the solid line is indicates the predicted values of the COVID-19 confirmed cases overall affected countries.

Fig 8: The predicted values of confirmed cases

Fig 7 and 8 shows the prediction of the spread of COVID-19 confirmed cases in India. The numbers of confirmed cases of COIVD-19 within respective duration are presented in the graph, X-axis presents the duration and Y-axis shows the number of CVOID-19 confirmed cases. ML model is trained for prediction based on the worldwide real-time dataset. Fig 8 shows India's current COVID-19 confirmed (Jun1,2020,198.37k) cases. Using the Fbprophet Model, the predicted number of confirmed (Jun 15, 2020,291.6754k) cases could be around 1800 in the next 14 days hence it can be concluded that with increasing duration spread of COVID-19 increasing and government should initiate appropriate control measures in time to regulate this pandemic.

XI. RESULTS AND DISCUSSION

The coronavirus disease has terrifically affected the lives of people around the globe. Many people have lost their loved ones with the number of deaths worldwide currently goes beyond 65 lakhs keeps increasing exponentially. While Different technologies have penetrated into our daily lives with many successes, they have also contributed to helping humans in the extremely tough fight against COVID-19. This paper has predicted a survey of COVID-91 spreading so far in the literature relevant to the COVID-19 crisis's responses and control strategies. This paper basically analysis the COVID-19 outbreak prediction and analysis based on the confirmed, recovered, and death cases on the given dataset and predict the Indias COVID 19 confirmed cases for the last 14 days by using time series data mining techniques. The Confirmed case rate on June 1 is 198.37k. Fig 8 shows the confirmed case predicted values start from 1/6/2020 to

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 5, May 2020

15/6/2020. Then the predicted rate after 14 days is above the 2 lakhs the exactly the rate is 291.6754k. The COVID 19 positive cases in India continue to rise on a daily basis. Fig 9 shows the COVID 19 status on 2 June 2020.

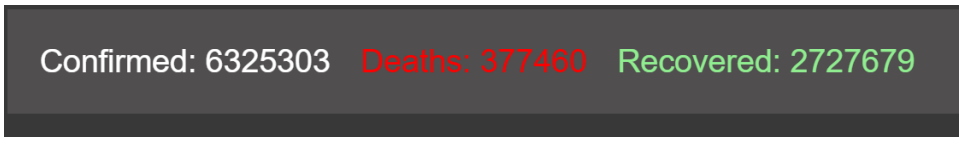

Fig 9: World Wide COVID 19 status

XII. CONCLUSION

Due to the pandemic of Coronavirus and COVID-19, all countries are looking towards mitigation plans to control the spread with the help of some modeling techniques. In this work, I have shown the prediction of covid19 by using a time series data mining technique based on the current dataset on the proposed combination of three major pillars to analyze the outbreak of the COVID-19 virus. The coronavirus disease has terrifically affected the lives of people around the globe. While Data mining techniques and related technologies have penetrated into our daily lives with many successes, they have also contributed to helping humans in the extremely tough fight against COVID-19. This paper has presented a predicted and analysis of confirmed, recovered, and death cases of COVID-19 and forecasting based on the number of cases time series based on the current data. This is partly due to the limited availability of data about COVID-19 whilst Data mining methods normally require large amounts of data for computational models to learn and acquire knowledge. This paper mainly predicted the COVID-19 outbreak in India for the last 14 days and analyzed graphically by using the data mining time series technique for both confirmed, recovered, and death cases. In this study, a machine learning data-driven Prophet Time series forecast algorithm has been used to predict the outbreak analysis of COVID-19 in India for the next 14 days i.e. 1/6/2020 to 15/6/2020.

XIII. ACKNOWLEDGEMENT

I need to offer my thanks to Almighty God who gave me power and patience in each attempt of my life. Next, I might want to thank the World Health Organization for persistently distributing COVID-19 related information which assumes a very important role in the present study.

REFERENCES

[1]Coronavirus latest: pandemic could have killed 40 million without any action. https://www.nature.com/articles/d41586-020-00154-w. Accessed March 27, 2020.
[2]Zhu W, Xie K, Lu H, Xu L, Zhou S, Fang S. Initial clinical features of suspected Coronavirus Disease 2019 in two emergency departments outside of Hubei, China. J Med Virol. 2020;(PG10.1002/jmv.25763-10.1002/jmv.25763):10.1002/jmv.25763-10.1002/jmv.25763. doi:10.1002/jmv.25763.
[3]The Novel Corona virus Pneumonia Emergency Response Epidemiology Team. The Epidemiological Characteristics of an Outbreak of 2019 Novel Corona virus Diseases (COVID-19)—China, 2020. China CDC Weekly. 2020;2:113-22.
[4]World Health Organization (WHO), "Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCov)," WHO, 2020.
[5]CoronaTracker Community, "CoronaTracker," CoronaTracker, 2020. [Online]. Available: https://www.coronatracker.com/. [Accessed 28 February 2020].