COVID-19 Outbreak Forecast Model using FbProphet

Nihla Iqbal
Department of Information and Communication Technology
South Eastern University of Sri Lanka
Oluvil, Sri Lanka
mifnihla@gmail.com

Fanoon Raheem
Department of Information and Communication Technology
South Eastern University of Sri Lanka
Oluvil, Sri Lanka
fanoonarfs@gmail.com

***Abstract* — The novel Corona virus, COVID-19 is a threat to the whole world at present. It has impacted economies, and the livelihood of countries and the people in them. Despite health precaution measures implemented around the globe the spread of this virus is still on the rise. This study aims to predict the number of confirmed individuals and deaths by the corona virus. The research is mainly focused on predicting time-series data using FbProphet forecasting model. The result helps the health institutions and the public on the understanding of the impacts and influence of the corona virus in their life and the importance to control its spread.**

Keywords — COVID-19, Corona, FbProphet

I. INTRODUCTION

The recent threat to global health is the outbreak of the novel Coronavirus. The global pandemic Coronavirus disease 2019 (COVID-19), was reported in December 2019, in Wuhan, Hubei Province in China. Since then the disease has been spreading throughout the globe driving the whole world under risk. From December 2019 to till September 30th 2020, the virus has affected around 213 countries or territories [1] with nearly 33, 844, 926 confirmed cases identified and more than 1, 012, 677 deaths all around the world have been reported [2]. The number of individuals being tested positive for COVID-19 is laddering up which has put medical doctors into a struggle to immediate invention of a vaccine that would control the spread of virus. The governments, World Health Organization along with the medical experts all around the world are on the run to minimize the spread of the virus by adopting several surveillance methods, PCR tests, tracking the individuals who have been tested positive and their movements, thereby sending for quarantine and self- isolation for a certain period of time or by predicting the impacts that would be caused by the virus.

This research studies time-series forecasting of the number of corona virus individuals who will be tested positive as well as the number of deaths that would occur due to the spread of the disease all around the globe, which would help the medical experts as well as the public to understand the severity and impacts that would be caused by the virus.

The methodology in section 3 clearly describes the method employed for the research work. It is mainly of four steps as: data collection, data visualization, building the forecasting model and making the predictions. The results and discussion in section 4 presents the results from the study conducted. The results are given as a comparison between the actual and predicted values of confirmed and death cases on a selected date.

II. OBJECTIVES

The proposed study on forecasting the COVID-19 outbreak aims to predict the number of confirmed cases and fatalities that would occur in the next 365 days by the spread of the Corona virus using FbProphet model. The study also aims to make the public aware of the consequences of not following the health safety measures.

III. METHODOLOGY

The methodology adopted forecasts the outbreak of COVID-19 globally based on two aspects: global confirmed cases and global death cases for the next 365 days starting from 2020-08-24. FbProphet model is used for forecasting, which is a time series prediction algorithm.

The flow diagram in Fig. 1. illustrates the proposed methodology. The process consists of four significant phases such as:

i. Data Collection
ii. Data Visualization
iii. Build the Forecasting Model
iv. Make Predictions

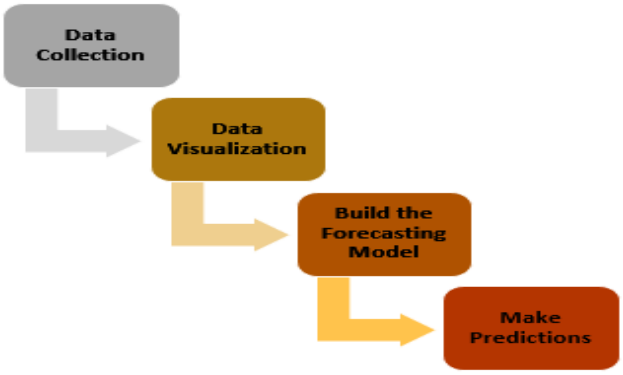

Fig. 1. The flow diagram of the proposed methodolgy

i. Data Collection

The most recent COVID-19 outbreak epidemical data on a daily basis are collected from the data repository operated by Johns Hopkins University center for Systems Science and Engineering (JHU CSSE) [3]. The total of globally confirmed and death COVID-19 cases from 2020-01-22 to 2020-08-24 were collected.

ii. Data Visualization

To have a clear understanding on the dataset used, it has been transformed into a visual context. The graphical representations are generated to identify the current trends for the total confirmed and death cases of COVID-19, globally. The dataset consists a total of 192 countries around the world. A total number of 23,457,652 people have been infected by the virus and nearly 808,892 people have died because of the virus during the timeline considered.

As shown in Fig. 2. the topmost infected country by COVID-19 is US with the overall of 24.31%. Similarly, the pie chart visualized, illustrates the countries in order with the corresponding percentages of the confirmed cases they have. Meanwhile, Fig. 3. shows the visualized pie chart based on the global death cases reported because of the virus from January to August 2020. Based on the analysis the highest number of deaths are reported in United States (US) with the

ISSN 2756-9160 / November 2020.
International Conference on Advances in Computing and Technology (ICACT–2020) Proceedings

percentage of 21.86%. Brazil takes the second place with 14.19%. Mexico (7.48%), India (7.11%) and United Kingdom (5.12%) are the counties next in line taking up the top most 5 countries with increased fatality rates from the global pandemic.

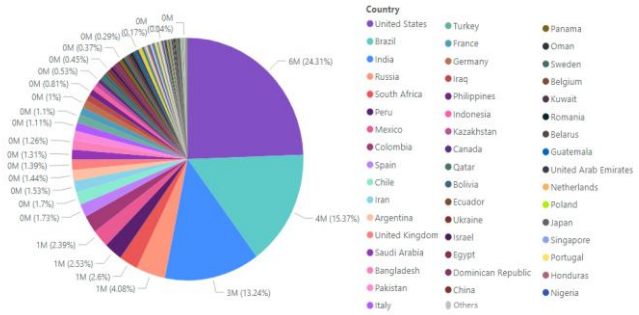

Fig. 2. The total number of COVID-19 confirmed cases with respective countries

.

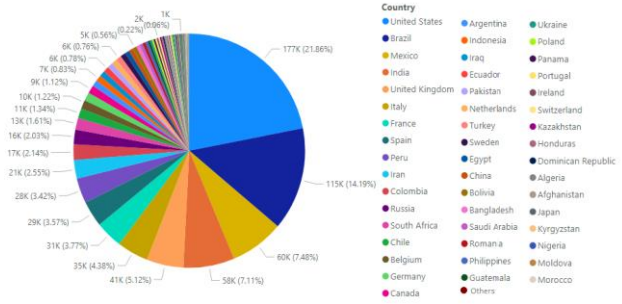

Fig. 3. The total number of COVID-19 death cases with respective countries

Further, the visulaizations confirm that the number of infections along with death are increasing globally. And from one country to another, the degree of infection varies in count. Therfore, it is understood that there is an exceptionally increasing demand for forecasting the trends and the growth pattern of this pandemic globally.

iii. Build the Forecasting Model

Based on the dataset the future impacts of COVID-19 will be forecasted considering the total confirmed and death cases from January to August worldwide. Among the prevailing algorithms that are used for data mining prediction and analysis, the adopted methodology uses FbProphet; a forecasting model developed by data scientists of Facebook for an enhanced time series prediction and unseasonal situation.

Sklearn model API is followed by FbProphet and an instance of Prophet Class is created where its fit and predict methods are called. The ds (datestamp) and y are the two dataframe columns of input to Prophet. Then the Prophet object is instantiated to fit the model. As a result of model fitting, the model will be able to learn the data and it can later be generalized to similar kind of data.

Table 1. Confirmed cases (ds,y)

Index	ds	y
0	22/01/2020	555
1	23/01/2020	654
2	24/01/2020	941
3	25/01/2020	1434
4	26/01/2020	2118
5	27/01/2020	2927
...	...	...
211	20/08/2020	22670483
212	21/08/2020	22949234
213	22/08/2020	23203532

Table 2. Death cases (ds,y)

Index	ds	y
0	22/01/2020	17
1	23/01/2020	18
2	24/01/2020	26
3	25/01/2020	42
4	26/01/2020	56
5	27/01/2020	82
...	...	...
211	20/08/2020	793698
212	21/08/2020	799252
213	22/08/2020	804416
214	23/08/2020	808676
215	24/08/2020	813022

Table 1. and Table 2. respectively show the COVID-19 confirmed and death dataset for the duration from 2020-01-22 to 2020-08-24. The two columns ds and y specify the date and the values considered, in a numeric form. In simple, it is a type of data conversion to a format preferred by Prophet.

iv. Make Predictions

To make predictions, a dataframe need to be created with the future dates. In context of FbProphet these are identified as periods. The periods are parameters indicating how many days to create. In the methodology adopted, the forecasting consists of 365 days as the period. Table. 3 shows some of the selected forecast dates (ds) after 2020-08-24.

Table 3. The selected forecast dates (ds)

Index	ds
188	2020-09-02 00:00:00
189	2020-09-03 00:00:00
190	2020-09-04 00:00:00
191	2020-09-05 00:00:00
192	2020-09-06 00:00:00
193	2020-09-07 00:00:00
194	2020-09-08 00:00:00

Table 4. shows the prediction of COVID-19 for the dates selected from the forecasting model. Here for each future row, a predicted value is assigned. It is given by the name yhat and is calculated based on yhat_lower and yhat_upper. The exact prediction is shown as yhat and yhat_lower, yhat_upper represents the minimum prediction and maximum prediction respectively.

Table 4. Predictions for confirmed cases

Index	ds	yhat	yhat_lower	yhat_upper
188	2020-09-02 00:00:00	1.05068e+07	2.45479e+06	1.83765e+07
189	2020-09-03 00:00:00	9.89288e+06	2.43156e+06	1.76665e+07
190	2020-09-04 00:00:00	1.02092e+07	2.82213e+06	1.77639e+07
191	2020-09-05 00:00:00	1.0618e+07	2.50816e+06	1.86283e+07
192	2020-09-06 00:00:00	1.06996e+07	2.37792e+06	1.86511e+07
193	2020-09-07 00:00:00	1.05705e+07	1.96485e+06	1.89826e+07
194	2020-09-08 00:00:00	1.07965e+07	2.64086e+06	1.85042e+07

As stated above, the forecasting model is built to predict for 365 days and these are some of the selected sample dates to demonstrate the predictions. Similarly, the predictions for death on dates selected are shown in Table 5. below. According to both predictions the number of confirmed cases along with the fatalities are increasing exponentially.

The plotting indicated in Fig. 4. and Fig. 5. illustrate the relationship between the original values and predicted for both confirmed and death cases all over the world. The actual values are shown in black dotted lines while the predicted values are shown in blue solid like lines.

ISSN 2756-9160 / November 2020.
International Conference on Advances in Computing and Technology (ICACT–2020) Proceedings

Table 5. Predictions for death cases

Index	ds	yhat	yhat_lower	yhat_upper
188	2020-09-02 00:00:00	453279	160006	739040
189	2020-09-03 00:00:00	428459	153452	718179
190	2020-09-04 00:00:00	438882	149182	731260
191	2020-09-05 00:00:00	457788	166662	734695
192	2020-09-06 00:00:00	465110	164528	739444
193	2020-09-07 00:00:00	457107	155454	740350
194	2020-09-08 00:00:00	464766	175482	742936

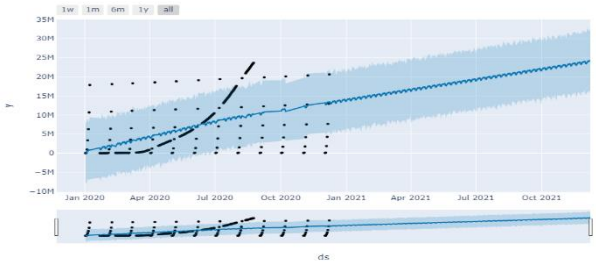

Fig. 4. The actual and predicted values for confirmed cases

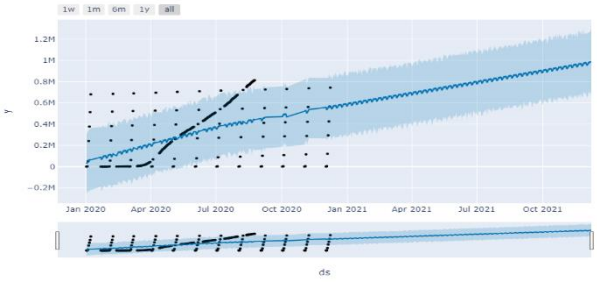

Fig. 5. The actual and predicted values for death cases

IV. Results and Discussion

Confirmed and death cases from COVID-19 are the time series that are constructed from the data collected. The experiments for forecasting the trends of COVID-19 are carried out based on a selected date (2020-10-02). While taking the confirmed cases, 32.30% is accomplished on the predicted values for 2020-10-02. And on the same date 46.31% is accomplished on the predicted values for death cases. Table 6. and Table 7. respectively show the COVID-19 confirmed prediction status and death prediction status on the date chosen.

Table 6. COVID-19 confirmed prediction status on 2020-10-02

ds	yhat	yhat_lower	yhat_upper
2020-10-02 00:00:00	1.10365e+07	3.31913e+06	1.91345e+07

Table 7. COVID-19 death prediction status on 2020-10-02

ds	yhat	yhat_lower	yhat_upper
2020-10-02 00:00:00	471064	192522	772306

The actual value with predicted value comparison is done based on WHO's Coronavirus Disease Dashboard as given in Fig. 6. R-Squared and Mean Squared Error metrics were used to compare the predicted values to the actual values, which values to 0.6532 and 2.342.

Fig. 6. WHO COVID-19 dashboard on 2020-10-02

V. Conclusion

In this work a forecasting model for COVID-19 has been proposed to predict the confirmed and death cases of COVID-19. The obtained data is visualized as pie charts to show the countries along with the number of cases reported; confirmed and death. The time series prediction is done for the upcoming 365 days from 2020-08-24. From the percentages achieved on the predicted values, it is observed that the future works must focus on increasing the predict percentage so that the predictions become much closer to the actual values for the forthcoming days. Since data mining algorithms require a larger amount of data for model generalizations and information extraction, time series data for COVID-19 is limited. Thus, this sort of constraint is experienced while using the model to predict 365 days. As a suggestion, an effort can be given to develop a hybrid quality model for better forecasting by combining the model implemented during the research, with the other existing time series forecasting models.

The usage of FbProphet model to evaluate and predict COVID-19 pandemic is a significant challenge. As stated above, with a larger data set the actual advantage of using this model can be achieved. More importantly, the model can be used to target the peak values of the pandemic which can assist the public health strategy of flattening the curve in controlling the COVID-19.

References

[1] Worldometer, "Countries where Coronavirus has spread - Worldometer," 2020. [Online]. Available: https://www.worldometers.info/coronavirus/countries-where-coronavirus-has-spread/. [Accessed: 30-Sep-2020].

[2] Worldometer, "Coronavirus Cases," *Worldometer*, 2020. [Online]. Available: https://www.worldometers.info/coronavirus/? [Accessed: 30-Sep-2020].

[3] CSSE, "GitHub - CSSEGISandData/COVID-19: Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE," *Johns Hopkins Whiting School of Engineering*, 2020. [Online]. Available: https://github.com/CSSEGISandData/COVID-19. [Accessed: 04-Oct-2020].