FORECASTING AND ANALYSIS OF COVID-19

Mr. Korivi Vamshee Krishna [1], Dr. Pulime Satyanarayana [2], Mrs. Palakollu Divya [3]
Department of Computer Science Engineering, Samskruti College of Engineering and Technology

.

ABSTRACT: Millions of people have been infected and lakhs of people have lost their lives due to the worldwide ongoing novel Coronavirus (COVID-19) pandemic. It is of utmost importance to identify the future infected cases and the virus spread rate for advance preparation in the healthcare services to avoid deaths. Accurately forecasting the spread of COVID-19 is an analytical and challenging real-world problem to the research community. Therefore, we use day level information of COVID-19 spread for cumulative cases from whole world and 10 mostly affected countries; US, Spain, Italy, France, Germany, Russia, Iran, United Kingdom, Turkey, and India. We utilize the temporal data of coronavirus[1-5] spread from January 22, 2020 to May 20, 2020. We model the evolution of the COVID-19 outbreak, and perform prediction using ARIMA and Prophet time series forecasting models. Effectiveness of the models are evaluated based on the mean absolute error, root mean square error, root relative squared error, and mean absolute percentage error. Our analysis can help in understanding the trends of the disease outbreak, and provide epidemiological stage information of adopted countries. Our investigations show that ARIMA model[7-11] is more effectivefor forecasting[6] COVID-19 prevalence.The forecasting results have potential to assist governments to plan policies to contain the spread of the virus.

KEYWORDS- Time Series Analysis, ARIMA model[7-11] , COVID 19 , Future forecasting , Dataset.

I.INTRODUCTION

The novel Coronavirus (COVID-19) has infected millions of people worldwide since it emerged from China in December 2019. COVID-19 has very high mutating capability, and it can spread very easily. Infected people from this virus suffer from severe respiratory problems, and may develop serious illness if suffering from chronic diseases like cardiovascular disease or diabetes or having weak immune system or being older in age. World health organization (WHO) declared on 11th March, 2020, the outbreak of COVID-19 as a pandemic. There are challenges to contain the disease because an infected person showssymptom after a long time or no sign of the disease. At present, no vaccination has been discovered for COVID-19. In this situation, social distancing, identifying the positive cases using testing at large scale, and containment of infected person is the only option to prevent the spreading of the virus .

The spread of COVID-19[2-5] can be classified under three major stages- 1. Local outbreak: at this stage, spreading chain of the virus among the people can be tracked, and the source of infection can be found out. The cases in this stage mostly relate to within family or friends, or the local exposure. 2. Community transmission: at this stage, source of the chain of infected people cannot be found out. The infected cases grow through cluster transmission in the communities. 3. Large scale transmission: at this stage, the virus spreads rapidly to other regions of a country due to uncontrolled mobility of people at large scale.

Due to high scale community impact and easy spreading worldwide, national governments imposed lockdown to control the spread of corona virus. As of 20th May, 2020,4996472 cases have been confirmed, 1897466 cases have recovered, 2328115 deaths have been reported, and 2770891 active cases have been identified worldwide. The statistical data is collected from , and the number of COVID-19 cases is calculated between 22 Jan, 2020 to 20 May 2020.

As no vaccine has been discovered of the disease, so motivation behind this paper is to model spreading of the corona virus, and predict the impact to optimize the planning to manage the various services and resources for the public by the governments. Some showing statistical analysis, modeling, and artificial intelligence to contain the spread of the virus, and highlight impacts in coming days. These early studies are carried out using very limited information available at early stage of the outbreak. Now, the virus has spread at large scale, and much information is available for the analysis. Predictive analysis of COVID-19[1-5] has become a hot research area to support health services and governments to plan and contain the spread of the infectious disease . Modeling and forecasting the daily spread behavior of the virus can assist the health systems to be ready to accommodate the upcoming number of patients. Accurate forecasting of the disease is a matter of concern because it may impact governments policy, containment rules, health system, and social life. Regarding this context, we explore the predictive capability of the ARIMA forecasting models. The models are widely used and accepted due to their more accurate forecasting capability. We use the day level cumulative cases of COVID-19 worldwide and 10 mostly affected countries; US, Spain, Italy, France, Germany, Russia, Iran, United Kingdom, Turkey, and India for our analysis study.

The objective of this paper is to provide evaluative study of prediction models using COVID-19 cases, and forecasting[6] the impact of the virus in the affected countries, and worldwide.We present trend analysis of COVID-19 cases, and compared the performance of the models using the metrics such as the mean absolute error (MAE), root mean square error (RMSE), root relative squared error (RRSE), andmean absolute percentage error (MAPE). We generate forecasting results for COVID-19 confirmed, active,recovered, and death cases.

II. LITERATURE SURVEY

Intensive research work is going on to evaluate and contain the worldwide disaster of COVID-19 on the human race. Research studies include predictions about the future cases , and analysis of the variables responsible for spread of the coronavirus .

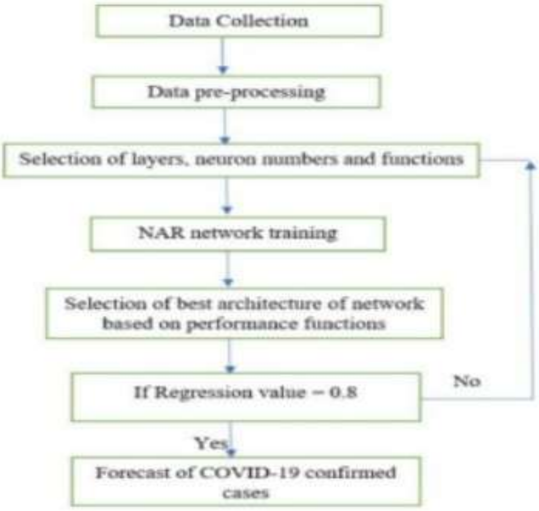
In the literature, time series forecasting problems have been studied widely in which COVID-19 forecasting is an emerging problem. Forecasting models[6] can be used to forecast the impact of the disease on the community which can help to control the epidemic. Performed forecasting evaluation study of the models using COVID-19 day level cases from 10 mostly affected states from Brazil. According to the authors, the stacking ensemble and SVR performed better as compared to ARIMA, CUBIST, RIDGE, and RF models for the adopted criteria.

III EXISTING SYSTEM

Arora et al. reported the Covid-19 forecasting of all states of India by applying long short-term memory (LSTM) models and predicted the next day and one-week Covid-19 cases with error of 3%. Hariri et al. used COVD-19 Essential Supplies Forecasting Tool (COVID-ESFT) to forecast Covid-19 severe, critical and death cases ofnorthwest Syria. It also identified the time points when health care system capacity will be worsened.

IV. FLOW CHART

Time Series Analysis of Covid 19


Data Collection
Data pre-processing
Selection of layers, neuron numbers and functions
NAR network training
Selection of best architecture of network based on performance functions
If Regression value – 0.8
No
Yes
Forecast of COVID-19 confirmed cases

V. SERIES FORECASTING MODELS

Time series forecasting models are used to predict the futuristic outcomes based on historical information. We have adopted ARIMA and Facebook Prophet (FBProphet) model in our evaluative and forecasting study. An overview of the models is given in the following sections.

AUTOREGRESSIVEINTEGRATEDMOVING AVERAGE (ARIMA) :

ARIMA[7-11] is composite of Autoregressive (AR) model, MovingAverage (MA) model, and 'I' stands for integration; where p is order of autoregression, d is order of differencing, q is order of moving average.

Dataset and Forecasting Framework ;

This section describes about the dataset we have used to forecast COVID-19 cases[1-5] of adopted countries, and worldwide. It also describes the modeling framework which we have followed.

MODELING DATASET

It is observed from the trends in section-III that rate of the reported COVID-19 cases in each country increases with time and flattens after sometime if large scale testing is performed, lockdown is imposed, and containment is followed. For our study, we disaggregated the available day level data of the adopted 10 countries. We discarded the initial 5 days data for each country in our study. The reason to discarding the initial samples is that testing of the samples grows slowly in starting phase which does not depict the actual rate of the spread. The utilized samples detail is given in. The end date of the collected samples is 20 May 2020. In this study, we consider 80 percentage samples for training and 20 percentage samples for testing the models for each country and worldwide data.

FORECASTING FRAMEWORK

The adopted framework for prediction and analysis of the COVID-19 cases using ARIMA[7-11], and FBProphet models. For the analysis, we have split the datasets of confirmed, active, recovered, and death cases into training and testing. We performed prediction after removing trends wherever applicable, and used statistical measures to evaluate the performance.

VI. SYSTEM TESTING

Framework Testing is a kind of programming testing that is performed on a total incorporated framework to assess the consistency of the framework with the relating necessities. Framework testing recognizes absconds inside both the coordinated units and the entire framework.

The consequence of framework testing is the noticed conductof a part of a framework when it is tried. Framework Testing is fundamentally performed by a testing group that is autonomous of the improvement group that assists with testing the nature of the framework unbiased. The steps of testing are involved for the proposed system as follow:

- Unit Testing
- Integration Testing
- Validation TestingUNIT TESTING

Unit Testing, otherwise called Component Testing, is a degree of programming testing where individual units or parts of programming are tried. The reason for existing is to approve that every unit of the product proceeds as planned.

A unit is the littlest testable piece of any product. It for the most part has one or a couple of data sources and normally a solitary yield. Unit testing expands trust in changing/keeping up with code. It is worried about practical rightness of the independent modules.

INTEGRATION TESTING

Joining testing takes as its feedback modules that have been unit tried, bunches them in bigger totals, applies tests characterized in a coordination test plan to those totals, and conveys as its yield the incorporated framework prepared for framework testing. It includes the mix of numerous modules which are firmly combined with one another. The primary capacity or objective of this testing is to test the interfaces between the units/modules. Blend testing can be started once the modules to be attempted are open. It doesn't need the other module to be finished for testing to be finished.

VALIDATION TESTING

An approval test is utilized to test and check the ultimate result before conveying it to the client.

VII. IMPLEMENTATION

1. EXPONENTIAL SMOOTHING

It is a strategy to constantly change a statistic, generally average, in the light of later experience. It relegates dramatically diminishing loads as the perceptions get more seasoned. As such, ongoing perceptions are given moderately more weight in determining than the more seasoned perceptions.

a) Single exponential smoothing

It is otherwise called simple exponential smoothing. It is utilized for small-range anticipating, typically only a few weeks or months into what's to come. The model accepts that the information varies around a sensibly steady mean (no pattern or reliable example of development).

b) Double Exponential Smoothing

It is the technique that is utilized when there is some sort of trend in the information. Double smoothing with a trend works a lot like Simple smoothing aside from that here, instead of one, the two segments i.e. level and trend, should be refreshed every period. The level is defined as the smoothen value of the approximation of the information toward the finish of every period. The trend can be defined as the smoothened estimate of mean development at the end of each window.

c) Triple Exponential Smoothing

This technique is utilized when the time series has trend and seasonality. Seasonality can be handled using a third boundary. We currently acquaint a third condition to deal with seasonality. The subsequent arrangement of conditions is known as the "Holt-Winters" (HW) strategy, which is named after its creators. There are two primary HW models, based upon the seasonality present in series.

- Multiplicative Seasonal Model
- Additive Seasonal Model

VIII. COMPONENTS OF TIME SERIES

Most of the time series have trend, seasonality, and irregularity associated with them. Moreover, some of these do have a cyclic order also. However, it is not compulsory to have a pattern in the time series model. So, let us discuss each one of them in detail. These components help find suitable forecasting methods for the short term analysis .

TREND

The Trend is a movement of higher values and lower values over a long time. So, when the values are directing upward ina time series is known as an upward trend. Also, the trend exhibits lower patterns to the downward is known as downward trends. Moreover, if it does not show any trend, it will be called a horizontal or stationary trend.

SEASONALITY

Seasonality generally has upward or downward swings. Nevertheless, it is quite a different form of a trend that showsa repeated pattern for a fixed period.

IRREGULARITY

Irregularity is also known as noise. It is the iritic nature of dataand is also called residual. So, it happens for only a short duration and not repeated like the other.

CYCLIC

Cyclic is the repeating up and down movement of a set of data in a graph. It means it can happen over more than a year and have no fixed pattern. They can repeat in one year, two years, or half of a year, and it is harder to predict.For predicting a time series analysis, it is crucial to consider the stationarity of the dataset. Time series requires data to be stationary, and it is a must for the analysis. The stationary has three components: the constant mean, constant variance, and autocovariance that does not depend on time. To check whether the dataset is stationary or not, two popular tests exist in python. The one is the rolling statistics and the Augmented Dickey-Fuller test (ADCF).

IX. TIME SERIES MODELS

Time series models are used to forecast events based on verified historical data. Common types include ARIMA[7-11], smooth-based, and moving average. Not all models will yield the same results for the same dataset, so it's critical to determine which one works best based on the individual time series.When forecasting, it is important to narrow down the specifics of your, ask questions about:

1. VOLUME OF DATA AVAILABLE— more data is often more helpful, offering greater opportunity for exploratory data analysis, model testing and tuning, and model fidelity.

2. REQUIRED TIME HORIZON OF PREDICTIONS — shorter time horizons are often easier to predict with higher confidence — than longer ones.

FORECAST UPDATE FREQUENCY — Forecasts might need to be updated frequently over time or might need to be made once and remain static (updating forecasts as new information becomes available often results in more accurate predictions).
FORECAST TEMPORAL FREQUENCY — Often forecasts can be made at lower or higher frequencies, which allows harnessing and up-sampling of data (this in turn can offer benefits while model.

X. RESULTS

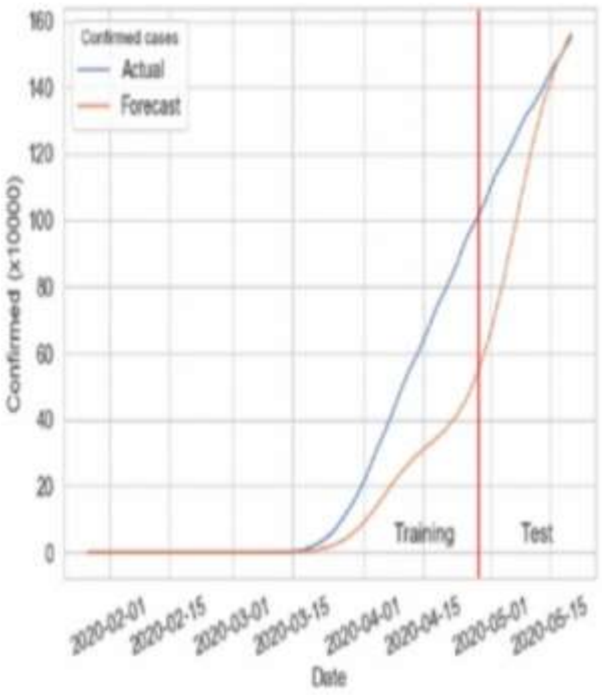
The adopted framework is implemented in Python 3.8, and we have used ARIMA and Prophet models[7-11] from openly available packages statsmodels and fbProphet respectively. We have performed our experiments in Intel Core i5 processor clocked at 2.40 GHz, 8 GB RAM, and 4GB NVIDIA GTX-1650 GPU. In this section, we will discuss about forecasting accuracy of adopted models for active, recovered, deaths, and confirmed cases.

FORECASTING OF ACTIVE CASES

Active cases are the number of infected people who are under medical supervision. Active cases are derived as shown in the following equation.
Active=Confirmed−Recovered−Deaths

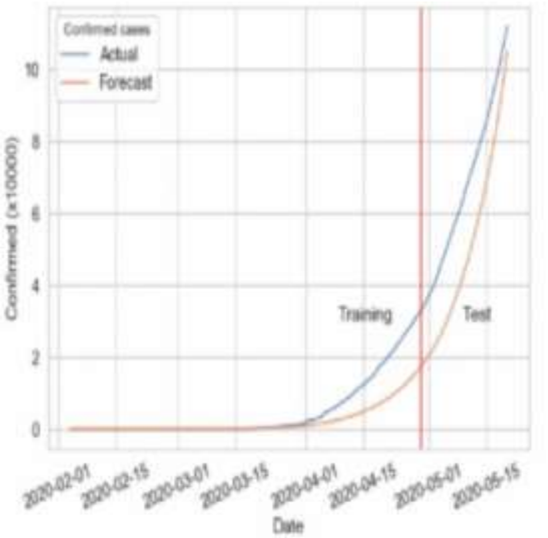
FORECASTING OF DEATH CASES

Coronavirus[1-5] has taken many lives. So, it is necessary to analyze the fatality rate of the virus, and forecasting to highlight future cases which can guide governments to act in advance. In this section, we have evaluated the forecasting models for death cases of the adopted countries, and worldwide. We have converted the non-stationary fatality data into stationary form to fit the ARIMA models[7-11]. FBProphet model is applied on the actual data to forecast the prediction results. Prediction accuracy of the models for the fatality cases.We can see that prediction errors of ARIMA are very less whereas FBProphet prediction have high error factor in the results. The results suggest that ARIMA can be used for actual forecasting of the cases to plan the services accordingly.



FORECASTING OF CONFIRMED CASES

In this section, we have highlighted the fitted accuracy of the models using confirmed cases. For this analysis, we have chosen only two countries; US and India. The results of US and India by both the models ARIMA[7-11] and FBProphet are shown in Fig. 6 and Fig. 7 respectively. We have shown training and testing data split using vertical line in the figures. Forecasted and actual data are plotted together to visualize the fitting accuracy of the models. We can see that FBProphet model is able to fit well in case of US data as shown whereas ARIMA is able to perform well in case of India data as shown . FBProphet adopts successively progression, and avoid outliers during modeling and forecasting. The results also depicts that FBProphet can fit well in case of less data Whereas ARIMA[7-11] requires sufficient data to model andpredict the results.



ARIMA Forecasting for India Confirmed Cases

DISCUSSION

In the current world, there are still frightening increases in the number of cases of new coronavirus[1-5]. The overall number of people affected is over 199 million, spreading across 222 countries, about 0.64% of Bangladesh. The primary goal of this study is to employ a commonly-used statistical analytic model, known as an ARIMA model[7-11], to observe and predict the epidemiology of COVID-19 in Bangladesh based on the data of daily confirmed cases publicly published by the Bangladesh Ministry of Health.

We believe that theoretical studies based on statistical modeling are essential for understanding the pandemic features of the epidemic to forecast the COVID-19 pandemic's possible trend. The results of such statistical models provide a complete picture of the present pandemic condition, allowing authorities to be active by developing plans and effective decisions to battle the pandemic and thereby limiting its impact on the economy, healthcare institutions, and society.

Building a reliable model to predict COVID-19 daily cases has constraints that begin with the uniqueness of the virus. It does not seem to follow any trend. After the first wave, the second wave arrived faster, and the third wave came in a flash. This work was completed on July 28, 2021, and we did not use the entire dataset for prediction. The infected rate in Bangladesh was very low at the start of 2020. All death and confirmed case records have surpassed in the third wave, which starts in April 2021. That is why we only used that dataset to predict for the upcoming days.

XI. CONCLUSION

This study investigates the applicability of different time series models to analyze COVID-19[1-5] data for India lockdown and unlock. An ARIMA-based prediction model[7-11] is developed that makes prediction taking into account the number of positive cases, number of tests conducted and the average positivity rate. In the present circumstances, the proposed model could be valuable in anticipating future cases of infection if the pattern of virus spread did not change abnormally. The analysis showed that the increase in number of cases per day that shoot up after lockdown was not an abnormal trend. The predicted graph based on the lockdown data had showed a significant increase in number of COVID-cases. With the ARIMA model, the forecasts are produced on the prior values of the time series and the error lags which actually helps the model to adjust its prediction values from sudden change in trends. A short-term forecasting was made in the time series and the outcome in the very near months.

FUTURE SCOPE

XII. REFERENCES

4. Kumar N., Susan S. 2020 11th international conference on computing, communication and networking technologies (ICCCNT) IEEE; 2020. Covid-19 pandemic prediction using time series forecasting models.

5.2021.Coronaviruscases:worldometer.URL https://www.worldometers.info/coronavirus/country/banglades h/, [Accessed on 02.08.2021] .

The future modifications to further improve the predictive accuracy of the models will include the creation of ensembles of the presented models that would combine the best of many worlds in order to reduce the overall error as

1. 2021. Bangladesh: WHO coronavirus disease (COVID-19) dashboard with vaccination data. [Online]. URL https://covid19.who.int/region/searo/country/bd/, [Accessedon 02.08.2021]

2. Surveillances V. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) China, 2020. *China CDC Weekly*. 2020;2(8):113–122.

3. Haghani M., Bliemer M.C., Goerlandt F., Li J. The scientific literature on coronaviruses, COVID-19 and its associated safety-related research dimensions: A scientometricanalysis and scoping review.

6. Mahalle P., Kalamkar A., Dey N., Chaki J., Shinde G. 2020. Forecasting models for coronavirus (COVID-19): a survey of the state-of-the-art. TechRxiv.

7. Alzahrani S.I., Aljamaan I.A., Al-Fakih E.A. Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. *J Infection Public Health.* 2020;13(7):914–919.

8. Sahai A.K., Rath N., Sood V., Singh M.P. Arima modelling & forecasting of COVID-19 in top five affected countries. *Diabetes Metab Syndr: Clin Res Rev.* 2020;14(5):1419–1427.

9. Khan F.M., Gupta R. ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India. *J Safety Sci Resilience.* 2020;1(1):12–18.

10. Kufel T., et al. Arima-based forecasting of the dynamics of confirmed Covid-19 cases for selected European countries. *Equilib Q J Econ Econ Policy.* 2020;15(2):181–204.

11. Dehesh T., Mardani-Fard H., Dehesh P. 2020. Forecasting of covid-19 confirmed cases in different countries with arimamodels MedRxi.