International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Covid-19 Case Analysis Using Deep Learning

Priyam Gupta

Department of Electronics and Computers, Thapar institute of Engineering and Technology, Patiala, Punjab, India

ABSTRACT

This paper aims to analyzing and then implementing predictive models for the purpose of forecasting the direction covid 19 cases is likely to take in India with a purely quantitative analysis. After proper analysis of each predictive deep learning model, I am going to use the model which gives the best result to give predictions on infection rate, death rate and recovery rate based on the entire population till Jul 31, 2021 then we will be further seeing how our model did with respect to the actual data till that day. The accuracy of each model understandably decreases as I increase the time frame of predictions and after careful literature survey, I used the 50 day window for the model to make fairly accurate mid-term forecasts. The models analyzed in this paper will be different layer LSTMs and ARIMA and its sub models.

Keywords: LSTM (Long short-term memory), ARIMA (auto regressive integrated moving average)

1. Introduction

On March, 2020 India lost its first life due to the virus Sars-Cov2 which causes coronavirus (Covid-19). The outbreak that started in early December 2019 in the Hubei province of the People’s Republic of China and spread worldwide as of September, 2021, has claimed more than 400 thousand lives in India alone.

The pandemic continues to challenge the government to arrange vaccines for all and make sure there isn’t any shortage of hospital beds or medical facilities in case we have another spike. Governments need access to some educated forecasts on the spread of the disease so they have a mock direction according to which they can plan their policies.

In this paper we propose a model which gives very good result on back testing and has been selected after experimenting with various models and referring current literature.

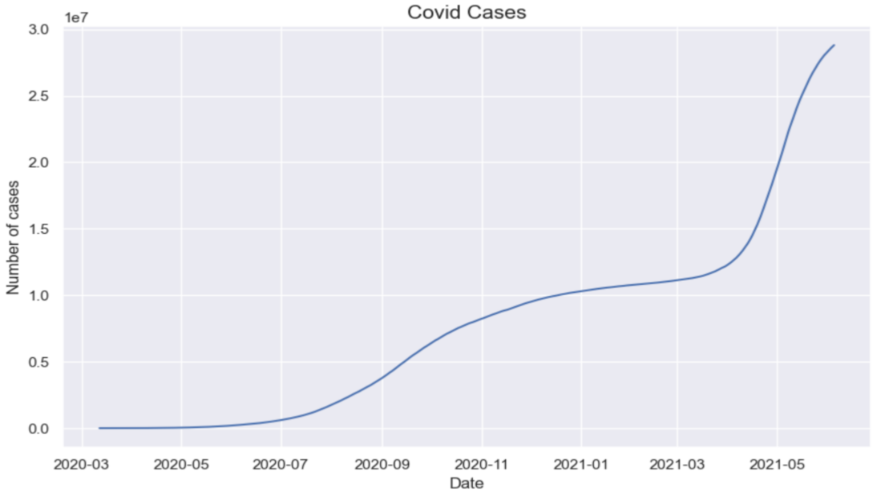
2. Literature survey:

Ref.	Forecasting method (Learning Algorithm)	Forecasting horizon	Type of data and Sample size	Data source	Accuracy	Purpose of prediction
Kırbaş et al. [1]	ARIMA, Nonlinear Autoregression Neural Network (NARNN) and Long-Short Term Memory (LSTM)	14 day ahead forecast	Cumulative confirmed cases data of 8 different European countries and the dataset is considered till 3, May 2020	European Center for Disease Prevention and Control	MAPE values of LSTM model are better than the other models.	To model and predict the cumulative confirmed cases and total increase rate of the countries was analyzed and compared. LSTM outperforms other models.
Arora et al.[2]	Deep LSTM/Stacked LSTM, Convolutional LSTM and Bidirectional LSTM	Daily and weekly predictions	Confirmed cases in India. March 14, 2020 to May 14, 2020	Ministry of Health and Family Welfare	Bi-directional LSTM provides better results than the other models with less error.	Daily and weekly predictions of all states are done to explore the increase of positive cases.
Zeroual et al.[3]	RNN (Recurrent Neural Network), LSTM, Bi-LSTM(Bi-directional), VAE (VariationalAutoEncoder)	17 days ahead forecast	Daily confirmed and recovered cases for six countries. Data from 22, January 2020 till 17, June 2020	Center for Systems Science and Engineering (CSSE) at Johns Hopkins University	Based on the performance metrics, VAE outperformed other models in forecasting the pandemic.	To forecast the number of new COVID-19 cases and recovered cases.
Shahid et al.[4]	ARIMA, support vector regression (SVR), long short-term memory (LSTM), Bi-LSTM	48 days ahead forecast	22 January 2020 to 27 June 2020. 158 samples of the number of confirmed cases, deaths and recovered cases.	Dataset is taken from the Harvard University	Bi-LSTM outperforms other models with lower R2 score values.	To predict the number of confirmed, death and recovered cases in ten countries for better planning and management.
Chimmula and Zhang[5]	LSTM	14 days ahead forecast	confirmed cases of Canada and Italy till 31, March 2020	Johns Hopkins University and Canadian Health authority	92% accuracy	To predict the number of confirmed cases of Canada and Italy and to compare the growth rate.
Alzahrani et al[6]	ARIMA, Autoregressive Moving Average (ARMA)	1 month ahead forecast	Cumulative daily cases from March 2, 2020, to April 20, 2020	Daily and cumulative confirmed COVID-19 cases in Saudi Arabia were collected from Saudi Arabia Government website.	ARIMA performs well than ARMA, MA and AR.	To predict the daily reproduction of confirmed cases one month ahead.
Ogundokun et al[7]	Linear regression model	8 days ahead forecast	March 31, 2020 to May 29, 2020	NCDC website	95% confidence interval	To predict the COVID-19 confirmed cases in Nigeria.

Ribeiro et al.[8]	ARIMA, cubist regression (CUBIST), random forest (RF), ridge regression (RIDGE), support vector regression (SVR), and stacking-ensemble learning	1,3 and 6 days ahead forecast	Cumulative confirmed cases in Brazil until April, 18 or 19 of 2020	The dataset was collected from an application programming interface that retrieves the daily data about COVID-19 cases which are publicly available	Based on the performance metrics, SVR, and stacking-ensemble learning outperformed other models	To predict the cumulative confirmed cases in Brazil
Tomar and Gupta[9]	LSTM	30 days ahead forecast	Cumulative and daily dataset of COVID-19 cases in India	Center for Systems Science and Engineering (CSSE) at Johns Hopkins University	LSTM has got 90% accuracy in predicting COVID cases	To predict the number of confirmed and recovered cases using data-driven estimation method.
Car et al[10]	Multilayer Perceptron (MLP) artificial neural network (ANN)	30 days ahead forecast	22nd January 2020 to 12th March 2020. Infected, recovered and deceased data	Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) and supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL)	Higher accuracy for confirmed cases with 0.986R2 Value	To predict the spread of pandemic world-wide.
Shastri et al[11]	LSTM, Stacked LSTM, Bi-directional LSTM and Convolutional LSTM	30 days ahead forecast	India and USA-Confirmed cases data from 7th Feb to 7th July 2020. Death cases data from 12th March to 7th July 2020.	Datasets of India and USA are taken from the Ministry of Health and Family Welfare, Government of India and Centers for Disease Control and Prevention, U.S Department of Health and Human Services.	ConvLSTM outperforms stacked and bi-directional LSTM in confirmed and death cases.	To predict the COVID-19 confirmed and death cases one month ahead and to compare the accuracy of deep learning models
Hawas[12]	Recurrent Neural Network (RNN)	30 days and 40 days ahead forecast	Daily confirmed cases in Brazil, 54 to 84 days, 7th April to 29th June 2020	Center for Systems Science and Engineering (CSSE) at Johns Hopkins University	Achieved 60.17% accuracy.	To predict one month ahead confirmed cases and to take preventive measures.
Papastefanopoulos et al[13]	Six different forecasting methods are presented. ARIMA, the Holt-Winters additive model (HWAAS), TBAT, Facebook's Prophet, Deep AR	7 days ahead for the ten countries	Jan 2020 to April 2020 and the population of countries.	Novel Corona Virus 2019 Dataset and population-by-country dataset from kaggle.com	ARIMA and TBAT outperformed other models in forecasting the pandemic	To predict the future COVID-19 confirmed, death and recovered cases by considering the country population.

3. Dataset:

The dataset used is the case_time_series.csv file from the website api.covid19india.org. It contains the following columns [“Date”, “Date_YMD”, “Daily confirmed cases”, “Total Confirmed cases”, “Daily Recovered”, “Total Recovered”, “Daily Deceased”, “Total Deceased”] and 451 entries beginning from 12th March from the time when then first death due to Covid-19 was reported to June 5th, 2021. For model performance analysis we are only using the date and “Total cases confirmed daily” statistic visualized below.



4. Model selection:

From the analysis of the current literature, it is clear that deep learning models for time forecasting like LSTM and ARIMA are the most popular and best working choices for our paper. So I am going to analyze these models and the ones closely related to them with respect to the data to give a complete overview.

5. Model parameters

I am only going to use the data of “total cases confirmed daily” since for now I am only analyzing which is the best model to use. The analysis will mainly focus on the metrics around the longer time frame close to 50 days as such a longer time frame hasn’t been covered in any research paper before and could give useful insights into the overall trend of covid-19.

For ease of analysis, I have taken both the look back period and size of prediction class as same. Also, for analysis I am more focused on how good is a model for identifying trends at this time frame instead of optimizing the actual predictions

- First, we analyze few LSTM model of type Vanilla and Stacked (1 layer, 3 layer) .
- Second is implementing ARIMA model but before doing that I will first analyze the Auto regressive model and the moving averages model and then analyze if ARIMA is actually necessary or not.

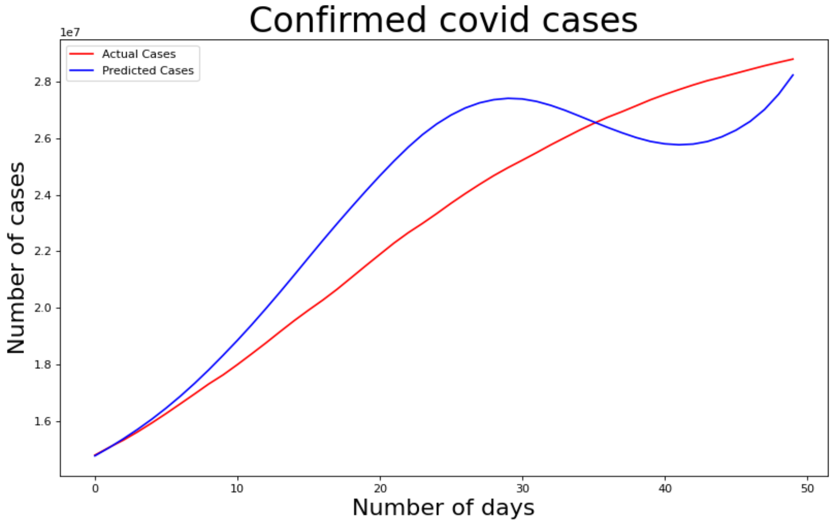
6. Model Performance:

In this section we calculate and compare the performance of various model on the total cases confirmed daily dataset.

6.1 Long Short-Term Memory (LSTM) model:

6.1.1. Vanilla LSTM:

A Vanilla LSTM is an LSTM model that has a single hidden layer of LSTM units, and an output layer used to make a prediction.



It has a mean absolute percentage error of 6.56%

6.1.2. Stacked LSTM

Multiple hidden LSTM layers can be stacked one on top of another in what is referred to as a Stacked LSTM model.
A 3-layer LSTM only marginally reduces the mean absolute percentage error to 6.5% which is improvement of only 0.06%.

6.2 AR, MA, ARIMA

Augmented Dicky Fuller test:

The Augmented Dicky Fuller test is a type of statistical test called a unit root test. The intuition behind a unit root test is that it determines how strongly a time series is defined by a trend. There are no. of unit root tests and ADF is one of the most widely used.

1. Null Hypothesis (H0): Null hypothesis of the test is that the time series can be represented by a unit root that is not stationary.

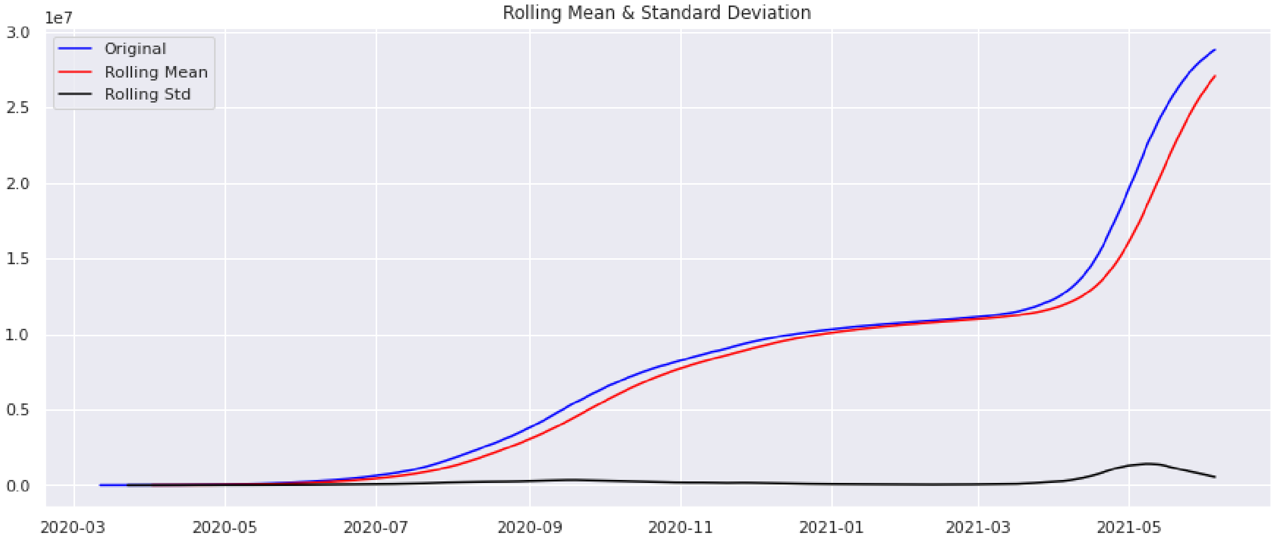
2. **Alternative Hypothesis (H1):** Alternative Hypothesis of the test is that the time series is stationary.

Interpretation of p value

1. p value > **0.05:** Accepts the Null Hypothesis (H0), the data has a unit root and is non-stationary.

2. p value < = **0.05:** Rejects the Null Hypothesis (H0), the data is stationary.

Result:



ADF Statistic: 2.0098

p-value: 0.998686

the graph is non stationary

critical value:

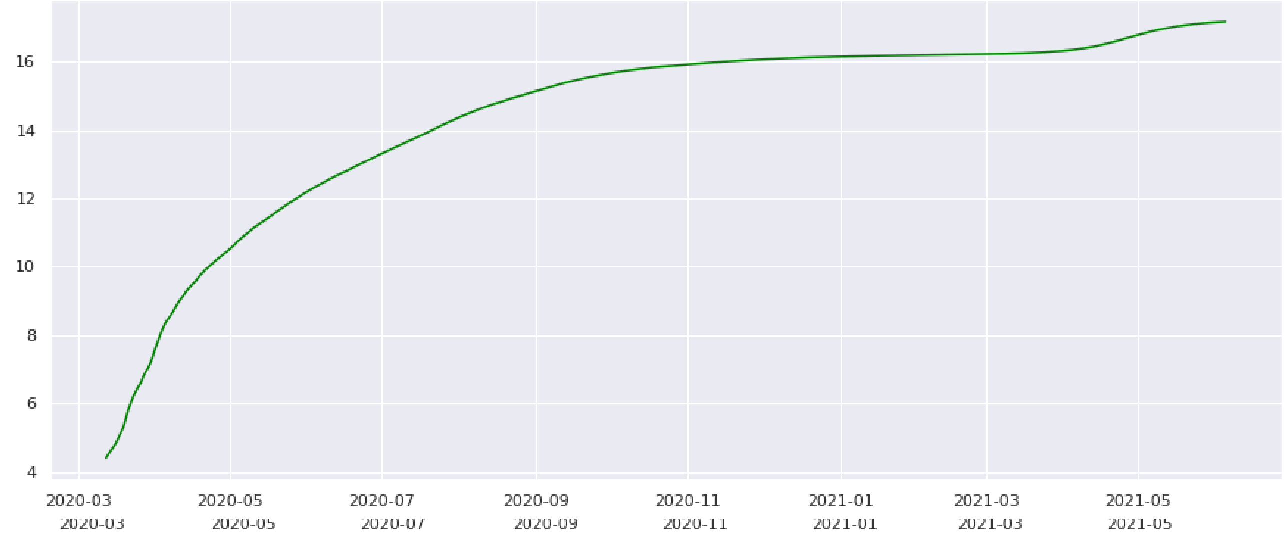
1%: -3.445

5%: -2.868

10%: -2.570

Now since p > 0.05 the time_series is non stationary. Thus, we need to do some transformations to make the series stationary.

Performing log transformation: often used to unskew highly skewed data



ADF Statistic: -3.441751

p-value: 0.009615

the graph is non stationary

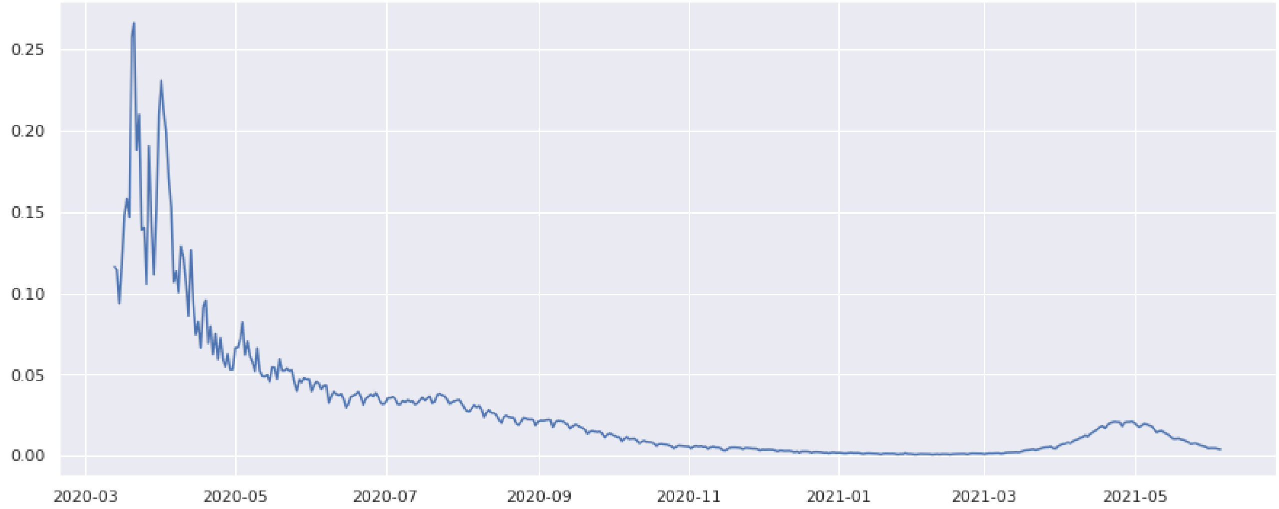
critical value:

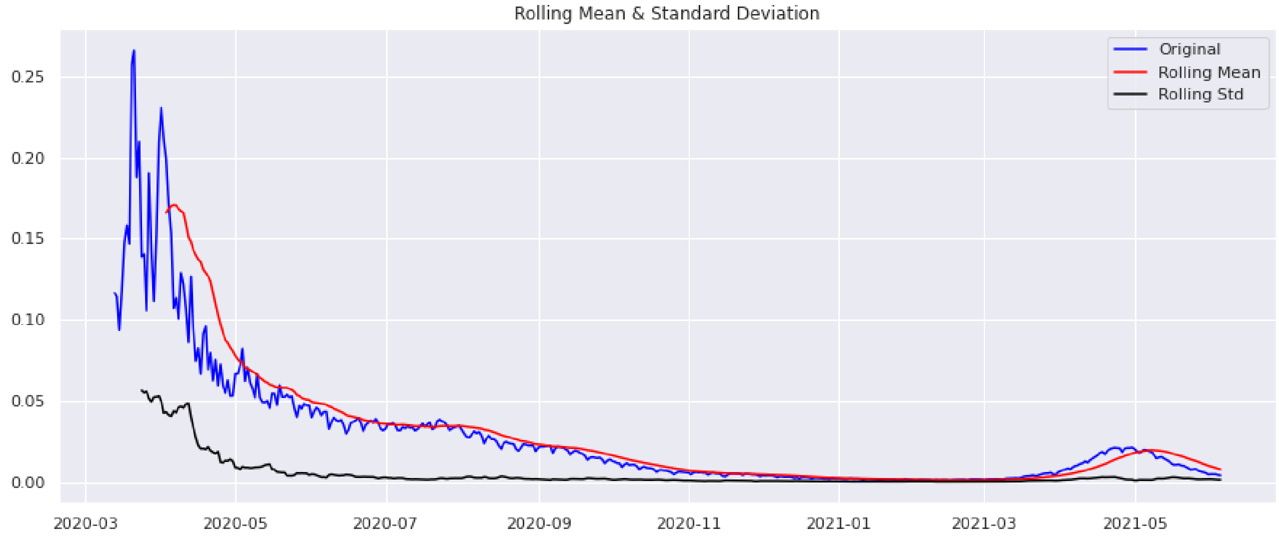
1%: -3.446

5%: -2.868

10%: -2.570

Removing trend and seasonality using difference: In case of differencing to make the time series stationary the current value is subtracted with the previous values. Due to this the mean is stabilized and hence the chances of stationarity of time series are increased.



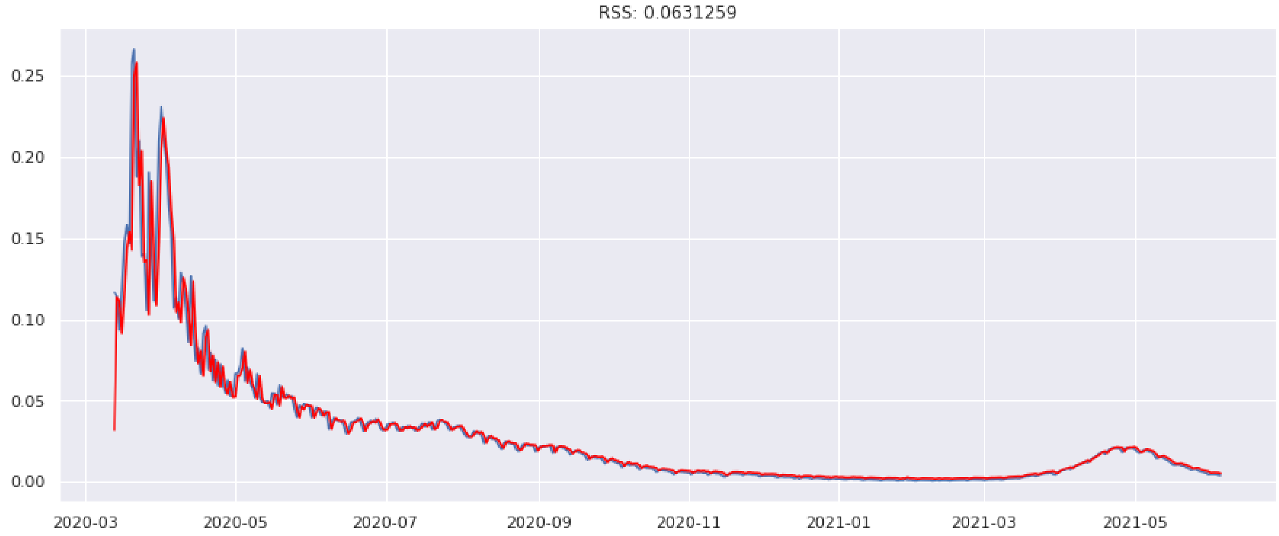


ADF Statistic: -8.425934
p-value: 0.000000
the graph is stationary
critical value:
1%: -3.446
5%: -2.868
10%: -2.570

Now our time series is stationary as p is equal to 0 and we can now apply time series forecasting model. We compare the RSS (Residual Sum of Squares) for all the models to decide which is better.

6.2.1 Autoregressive model:

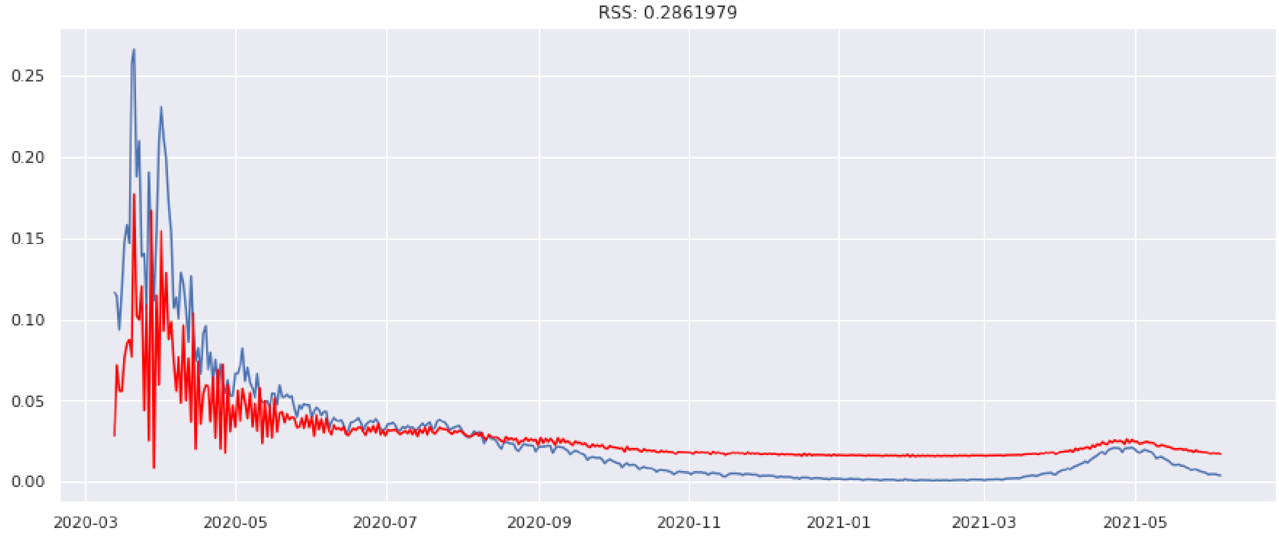
An autoregressive (AR) model *predicts future behaviour based on past behaviour*. You *only* use past data to model the behaviour, hence the name *auto*regressive (the Greek prefix *auto*– means “self.”). The process is basically a linear regression of the data in the current series against one or more past values in the same series.AR models are also called conditional models, Markov models, or transition models.



Result: RSS = 0.0631259

6.2.2 Moving Average model:

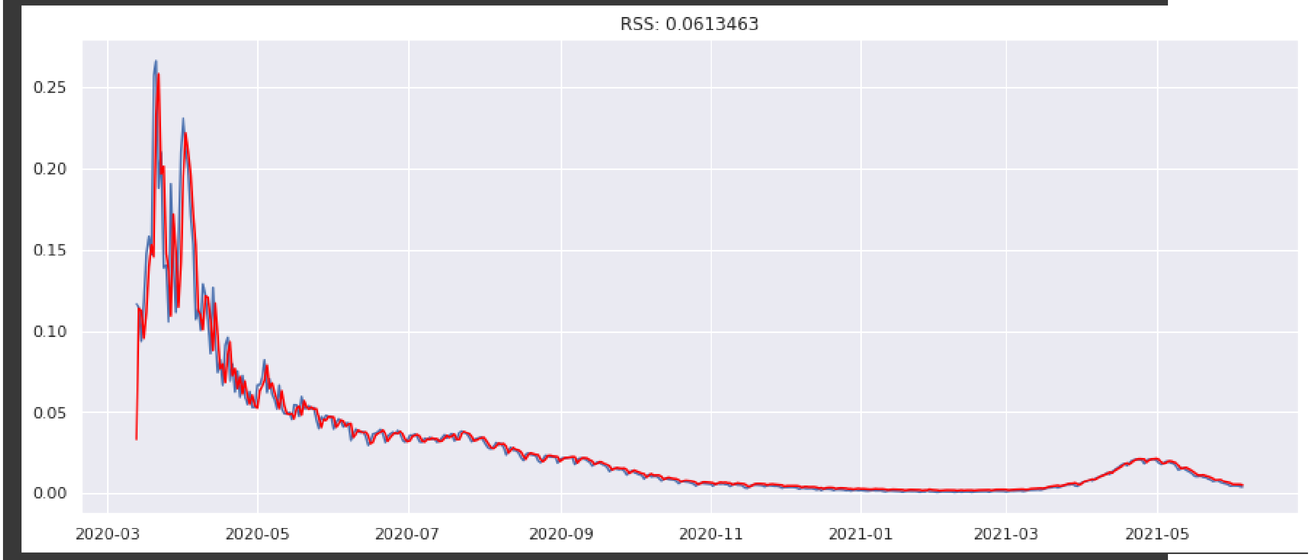
The moving average model is a time series model that accounts for very short-run autocorrelation. It basically states that the next observation is the mean of every past observation.



Result: RSS = 0.2861979

d structures in

n AR and MA



Result: Accuracy of the model for the selected parameters for a 50day time frame is around 84%

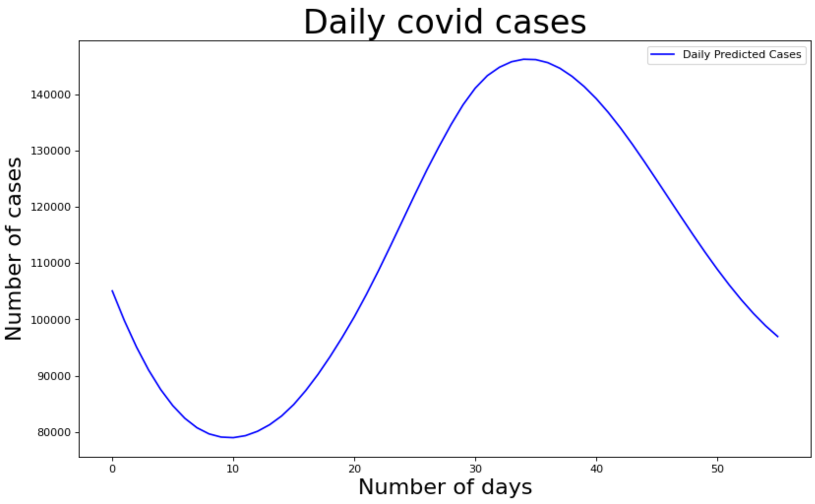
Analysis Result: After Analysis it was found that Vanilla LSTM model performed better for longer time frames than ARIMA and since the focus of this paper is on 50-day time frame we don't use the ARIMA model. Thus, for our project to predict the spread o covid-19 we would use Vanilla LSTM.

7. Predictions

Since 31st July is nearly 50 days ahead of our training data it will provide an insight how are model compared in results to the real-world data predicting around 2 months ahead.

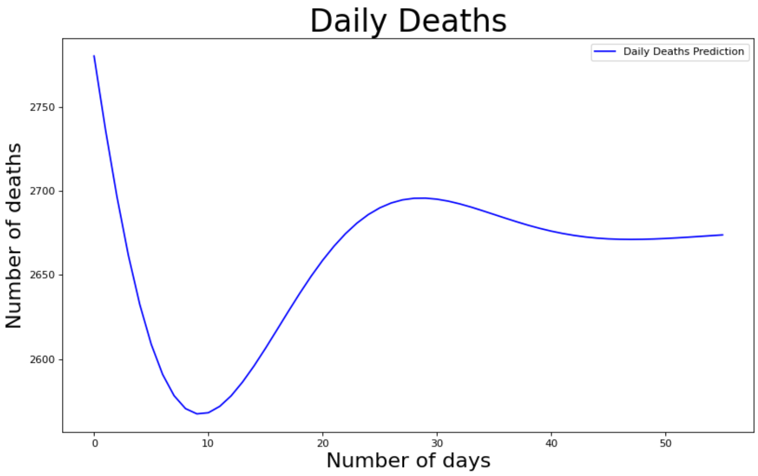
One point to note is that we can't use the same *"total confirmed"* or any other total table that we have been using for analysis because the way LSTM works is that it takes previous inputs (50 in our case) and tries to predict the next one and thus basically acts as a trend predictor so if it detects a sort of negative trend then that sentiment compounds in the next inputs and sometimes it might go below current value and we know that the total daily case tally can't go down (since new cases added every day) than the current number so it gives us an extremely unreliable prediction. Thus, we use the total cases daily dataset to predict and add all the daily predictions to give the final predictions.

Total cases prediction:



Result: According to our model the tally of the total confirmed cases on July 31st will reach 3,53,93,090and the actual result number of cases reached 3,16,13,993.

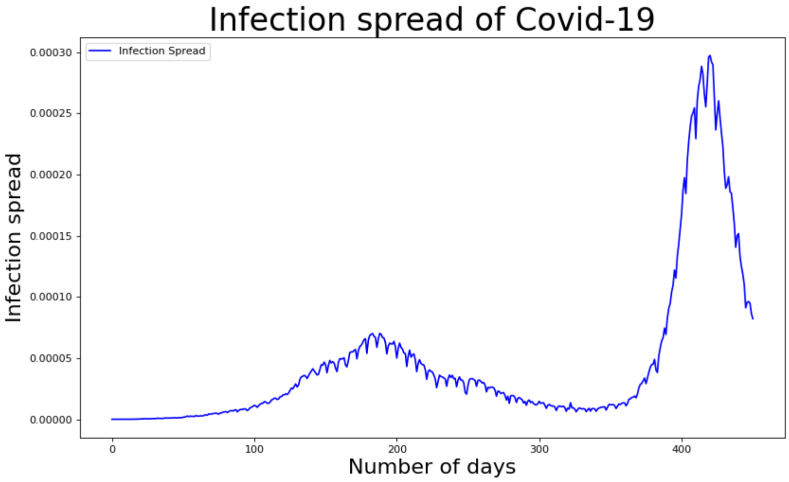
Total deceased prediction:



Result: According to our model tally of total deaths on July 31st will reach to 495122.40625 so we round that off to 4,95,123and the actual result number of deceased reached 4,23,810.

Predicting when will the infection reach 10% of its peak:

Using the formula (Infection spread = daily confirmed cases / Total population) we can visualise the spread of the virus.



Result: We generated values using our model to see when is the value we want to predict which is 10% of the all-time high infection spread is reached and our model predicts that the required infection spread will reach on June 19ththe actual result shows we reached it on June 21st.

Conclusion

Although such analysis doesn't take into account the developments in the medical field and the ongoing news on the matter, For Governments, It is still a good way to gage the direction of the trend the pandemic is taking and having a quantitively backed framework to base their policies on.

The models prediction for 2 months ahead that is October 31st is:

Total Cases: 3,45,03,788

Total Deceased: 4,54,679

References:

1. İ. Kırbaş, A. Sözen, A.D. Tuncer, F.Ş. Kazancıoğlu *Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches* Chaos Solitons Fractals, 138 (2020), p. 110015, 10.1016/j.chaos.2020.110015
2. P. Arora, H. Kumar, B.K. Panigrahi *Prediction and analysis of COVID-19 positive cases using deep learning models: a descriptive case study of India* Chaos Solitons Fractals, 139 (2020), p. 110017, 10.1016/j.chaos.2020.110017
3. A. Zeroual, F. Harrou, A. Dairi, Y. Sun *Deep learning methods for forecasting COVID-19 time-Series data: a Comparative study* Chaos Solitons Fractals, 140 (2020), p. 110121, 10.1016/j.chaos.2020.110121
4. F. Shahid, A. Zameer, M. Muneeb *Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM* Chaos Solitons Fractals, 140 (2020), p. 110212, 10.1016/j.chaos.2020.110212
5. Vinay Kumar Reddy Chimmula, Lei Zhang. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos Solitons Fractals 135; 2020: 109864. doi:10.1016/j.chaos.2020.109864.
6. Saleh I. Alzahrani, Ibrahim A. Aljamaan, Ebrahim A. Al-Fakih *Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions* J Infect Public Health, 13 (7) (2020), pp. 914-919, 10.1016/j.jiph.2020.06.001
7. Roseline O. Ogundokun, Adewale F. Lukman, Golam B.M. Kibria, Joseph B. Awotunde, Benedita B. Aladeitan *Predictive modelling of COVID-19*

confirmed cases in Nigeria Infect Disease Model, 5 (2020), pp. 543-548, 10.1016/j.idm.2020.08.003

8. M.H.D.M. Ribeiro, R.G. da Silva, V.C. Mariani, L.D.S. Coelho, *Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil* Chaos Solitons Fractals, 135 (2020), p. 109853, 10.1016/j.chaos.2020.109853
9. A. Tomar, N. Gupta, *Prediction for the spread of COVID-19 in India and effectiveness of preventive measures*, Sci Total Environ, 728 (2020), p. 138762, 10.1016/j.scitotenv.2020.138762
10. Zlatan Car, Sandi BaressiŠegota, Nikola Anđelić, Ivan Lorencin, Vedran Mrzljak, *Modeling the spread of COVID-19 infection using a multilayer perceptron*, Comput Math Methods Med, 2020 (2020), pp. 1-10, 10.1155/2020/5714714
11. Sourabh Shastri, Kuljeet Singh, Sachin Kumar, Paramjit Kour, Vibhakar Mansotra, *Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study*, Chaos Solitons Fractals, 140 (2020), p. 110227, 10.1016/j.chaos.2020.110227
12. M. Hawas, *Generated time-series prediction data of COVID-19's daily infections in Brazil by using recurrent neural networks*, Data Brief, 32 (2020), p. 106175, 10.1016/j.dib.2020.106175
13. VasilisPapastefanopoulos, PantelisLinardatos, Sotiris Kotsiantis. COVID-19: a comparison of time series methods to forecast percentage of active cases per population. ApplSci 10; 2020: 3880. doi:10.3390/app10113880.