COVID-19 Prediction and Modeling using Machine Learning and Natural Language Processing

Yuvraj Shivtare Meetkumar Patel Raghav Daga

I. INTRODUCTION

The current ongoing pandemic has affected all countries worldwide. While we wait for the medical professionals to come up with medicine or a vaccine, Data scientists and Machine Learning engineers all over the world have found a new challenge. With the immense amount of data resources accessible currently, the biggest and the most interesting problem for us has been the prediction and modeling for what the future will look like 5 days from now, 10 days from now, and months from now. In this project, we first try to learn what the Corona-virus is. We try to ascertain what kind of a virus it is by understanding the keyword from published papers. Once we know what the challenge we are facing, we can then examine if the steps we are taking to alleviate the spread of the virus, are those steps A) Necessary, B) Sufficient and C) Whether they are working or not. The first one is simple and can be answered by a Google search. Its the next two that require proof.

The proof lies in the analysis of the data. We need to visualize the data and understand the changes in the curve for every country. Only then can we say whether the actions taken by the country is sufficient to contain the virus. Once we understand the behavior of this data that we have on hand, we can then build a model around the data for prediction.

II. RELATED WORK

Ever since news of this pandemic broke out, everyone has been trying to come up with prediction models for their own countries and for the overall outbreak worldwide. The modeling and analysis of this has been one of the biggest challenges. Global competitions are being held on Kaggle and other data analysis websites, World Organizations are coming up with their own models for different conditions (like easing of the lock-down by a certain percentage) and some major Universities have made their own prediction models for this too.

We saw a lot of models that were used for the forecasting like SEIRD and Holts models. We were particularly interested in the SEIRD model and we dug a little deeper to find out why infectious disease outbreak models are usually loosely designed around it. SEIRD stands for exactly what a disease outbreak is made up of. There is a certain number of people who are counted as *Susceptible*. Some of them are *Exposed* with the progression of time. They then move on to the *Infected* phase after the virus incubation period. Out of those infected people, some *Recover* and some of them *Die*.

But this seemed like a very idealistic model to us. Granted that the infectious diseases have had a pattern in the past,

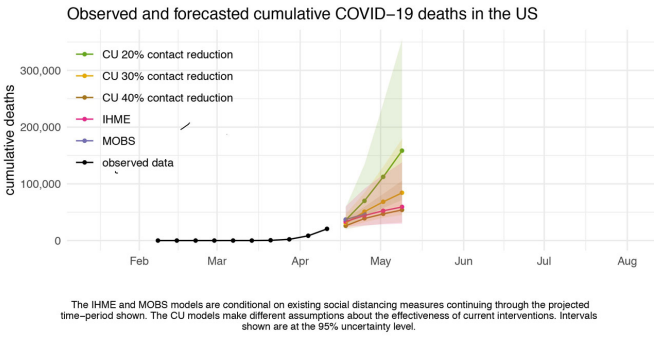

Fig. 1. Various Models forecasting cumulative COVID − 19 deaths in the United States of America

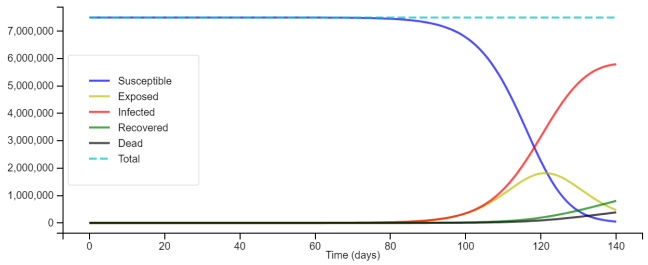

Fig. 2. SEIRD Model Output

but in the highly data driven and documented world, none of those diseases required drastic measures like enforcing social distancing, imposing lock-downs, or even caused shortage of Hospital supplies or PPE. We plot the model for the disease by ourselves and it seemed very hard-coded and inaccurate to us. So, instead we decided to come up with our own analysis for the disease in our project.

III. OUR SOLUTIONS

A. Description of the Data-set

In response to the COVID − 19 pandemic, the White House and a coalition of leading research groups have prepared the COVID − 19 Open Research Data-set (CORD − 19). CORD − 19 is a resource of over 44,000 scholarly articles, including over 29,000 with full text, about COVID − 19, SARS − CoV − 2, and related Corona-viruses. This freely available data-set is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. There is a growing urgency for these approaches because of the rapid acceleration in new

Corona-virus literature, making it difficult for the medical research community to keep up. This data set will be used to extract keywords related to COVID − 19[1].

Also, we will be using the data repository for the 2019 Novel Corona-virus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) to monitor and access the real − time data on a count of the number of deaths, recovered cases, country/region[2].

B. Data Visualization

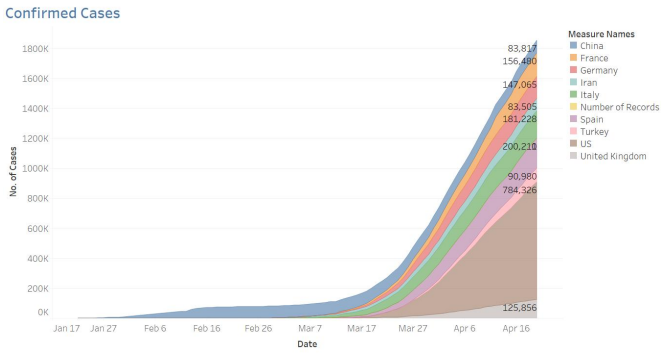

Fig. 3. Confirmed Cases

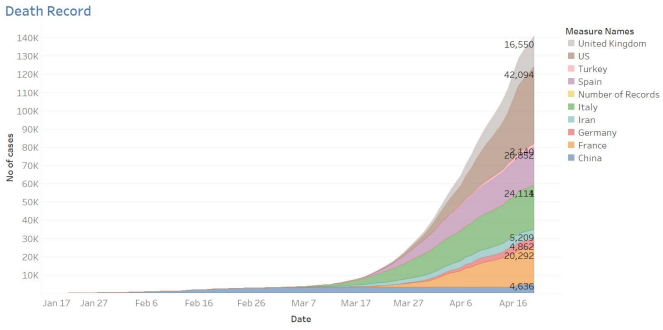

Fig. 4. No. of Deaths

Data is visualized based on the number of cases reported. Separate plots have been made for the total number of confirmed cases, death reported, and the total number of recovered counts. The below graphs can be analyzed based on the area under the curve for each country, higher the area implies higher will be the count. Accordingly, the graphs can be observed on confirmed cases, deaths, and recovered cases.

We have also visualized the death trend with a lag of 2 days i.e., the difference between today's death value and two days before is considered to plot figure 6.

In figure 7, the color gradation is based on the number of cases confirmed in respective countries. Further we analyzed the data for India, to observe the daily trend and growth factor for confirmed, recovered, and death cases in India. Daily trend is calculated at lag of 1 i.e., the difference between todays number of confirmed cases and yesterdays number of

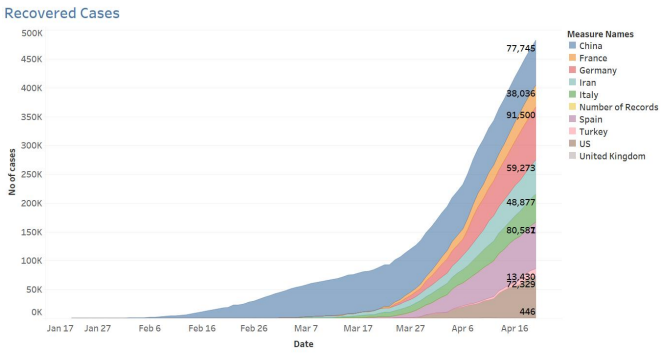

Fig. 5. No. of Recoveries

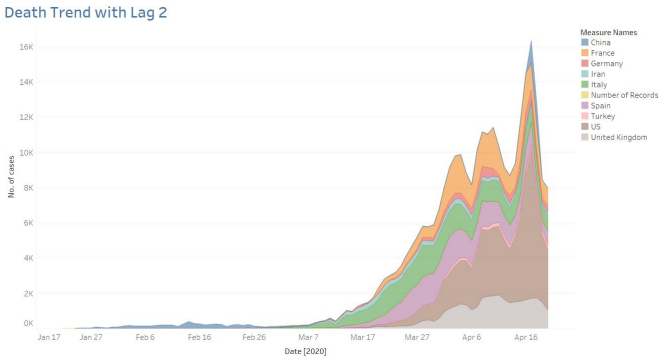

Fig. 6. Death trend with lag 2

confirmed cases is used to calculate the todays increase in count. Similarly, trend is calculated for deaths and recovered cases as seen in figure 8.

A growth factor is calculated dividing todays count by yesterdays count for respective cases. This gives the factor with which each category of cases is reported on a timescale (seen in figure 9).

C. Natural Language Processing

Reading through 44000 scholarly articles or even their abstract is a very tough task which takes months, not even days. We decided to use Natural Language Processing to extract the keywords from the abstract of these scholarly articles. The first thing we did was to see the most used words in these papers which can be seen from the figure 10. For starters we took the sample size as 400.

Words like *This, Abstract, Introduction, The* give us little to no information. So, we decided to remove what is called *stop words* from the abstract of these papers. Stop words are the words that are used most in the language, that is, filler words like articles, pronouns, conjunctions, etc., which contribute almost zero to our analysis. Now that the words were out of our way, we could go deeper with our analysis of the keywords from these articles.

We pooled the keywords into a dictionary, after which we had to choose our algorithm of choice. We went with Latent Dirichlet Allocation which is a type of unsupervised learning

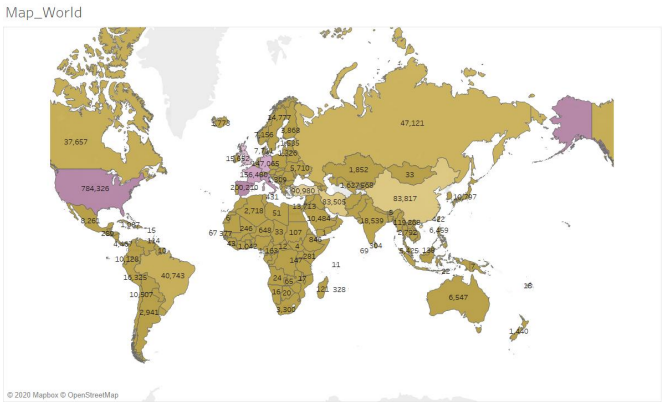

Fig. 7. COVID impact on countries

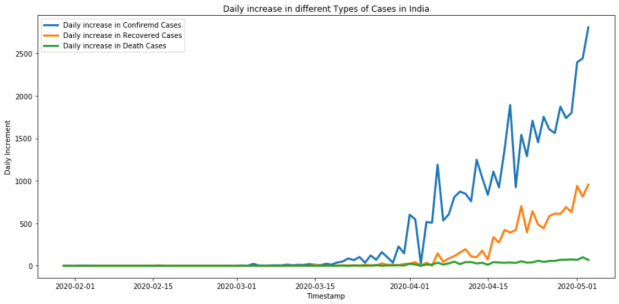

Fig. 8. Daily increase in different types of cases in India

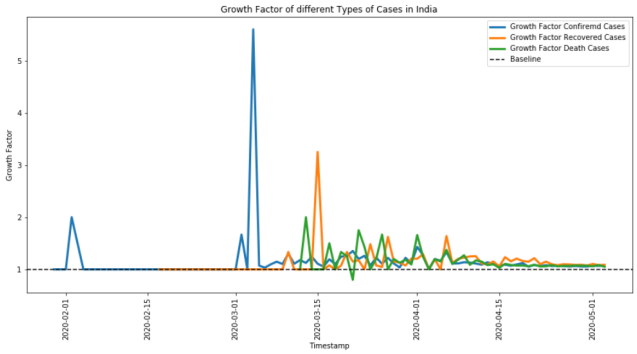

Fig. 9. Growth factor of different types of cases in India

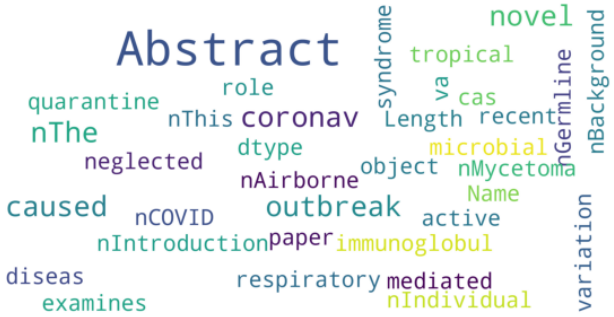

Fig. 10. Keyword Extraction (Sample size – 400 articles)

because it could give us a more in depth understanding of the keywords if they were divided into certain *Topics*. We classified our keywords present in the dictionary and then displayed the weight-age of every keyword in the topics in our figure, the result for which can be seen in figures 11 and 12.

Topic 1 contains words like *structure, epitope, vaccine, agent,* and *disease* which tells us that this group of words is related to the vaccine for the virus.

Topic 2 has the words *virus, viral, human, genome, infection*. This group of words tells us more about what the strain is. It is a viral infection that affects the human genome.

Topic 3 delves more into the complications caused by Corona-virus - infection at the *tissue* and *cellular* level, causing *cytokine* storm which is a severe immune system overreaction[3].

Topic 4, with words such as *protein, drug, cell,* and *target*, gives us information about how it causes these complications. For example, a little bit of research into the relation between Corona-virus and the word protein told us that the strains RNA genome binds itself to the N-protein and the S-protein[4].

Topic 5 is the most important one for our project. The keywords here are *infection, transmission, high, population, model, estimate, epidemic, control, measure, outbreak*. By combining the keywords and a little bit of analysis and research told us that the Corona-virus like SARS and MERS could have been an epidemic having an outbreak in a high population area. The Corona-virus 2019 turned into a pandemic which is basically an epidemic that has spread over a worldwide scale. The control measures for Corona-virus and such related viral infections are cleanliness and sanitation, keeping a safe distance, hydration, and vaccination.

Topic 6 contained words related to clinical studies conducted with patients who had Corona-virus when it first emerged (the 2019 pandemic virus has a different strain). We can assume that these keywords are from the published papers about the clinic trials conducted for the treatment of the virus.

Topic 7 has words like *strain, gene, adenovirus, forager,* etc. We looked up the term adenovirus and this is what the CDC had on their website about adenoviruses –

Adenoviruses are common viruses that cause a range of illness. They can cause cold-like symptoms, fever, sore throat, bronchitis, pneumonia, diarrhea, and pink eye (conjunctivitis). You can get an adenovirus infection at any age. People with weakened immune systems or existing respiratory or cardiac disease are more likely than others to get very sick from an adenovirus infection.

We realized that this is exactly what is happening with Corona-virus right now and it was good for people like us to understand in very simple terms what it meant.

Topic 8, with words such as *sample, detection, use, tool, sequence,* gave us more insight into the detection of the virus.

We looked at the data for a few countries and realized that South Korea followed almost all these steps because they understood the disease and took immediate action. Which is

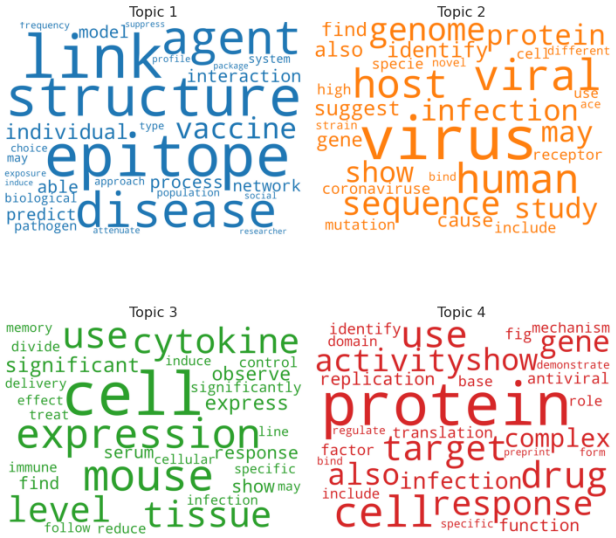

Fig. 11. Keywords classified into topics 1 − 4

Fig. 12. Keywords classified into topics 5 − 8

why South Korea has just 10,600 cases despite having one of the top 25 highest population densities[5]. South Korea peaked early, and their daily cases curve has been decreasing ever since they hit the peak.

On the other hand, the United States of America was hesitant in taking quick action and the number of cases has not started declining yet. Due to loose restrictions, despite being number 174 in the Population Density Rankings, the USA has more than 4 times the number of cases in the next country, Spain.

D. Polynomial Regression

Polynomial Regression is a form of linear regression in which the relationship between the independent predictor variable x and dependent outcome variable y is modeled as an nth degree polynomial.

Sr. No.	Date	PR Prediction
0	2020-05-04	43896.38
1	2020-05-05	46485.65
2	2020-05-06	49287.54
3	2020-05-07	52340.47
4	2020-05-08	55689.07

TABLE I
POLYNOMIAL REGRESSION PREDICTION FOR CONFIRMED CASES INDIA

Linear model,

$$Y = w_0 + w_1 x$$

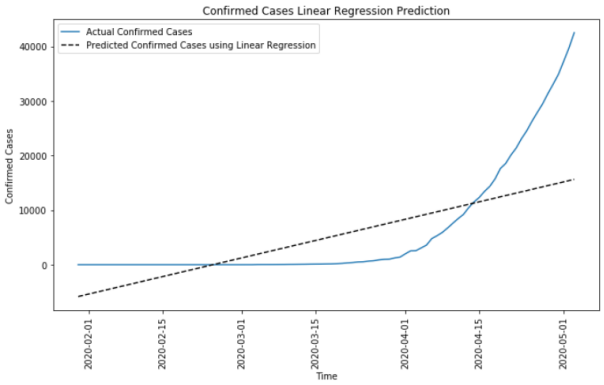

Fig. 13. Linear Regression for Confirmed Cases in India

If the linear model is used to forecast the confirmed cases count in India, from figure 13 we can clearly observe the under fitting of the curve, as the trend of confirmed cases in not linear, our Linear Regression Model is falling apart.

Hence, we consider using the polynomial regression model. polynomial regression is a linear model with a higher order equation, we can add powers of the original features(x) as new features $(x^2, x^3, x^4, , x^n)$ for nth order polynomial. Linear model can be transformed to,

$$Y = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + + w_n x^n$$

We used 8th order polynomial to predict the confirmed cases in India (shown in figures 13 and 14). RMSE (Root Mean Squared Error) is calculated to score the performance of model. RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It is the square root of the average of squared differences between prediction and actual observation. Below is the mathematical representation of the RMSE, where $\hat{y}_i$ is the predicted curve and y is the observed curve.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Root Mean Squared Error observed for Polynomial Regression Model: 517.246

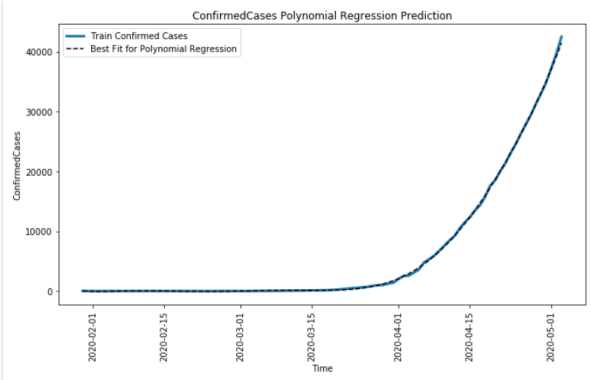

Fig. 14. Polynomial Regression for Confirmed Cases in India

E. Support Vector Machine

We are using support vector machines to regression problems the property of sparseness. In simple linear regression, we minimize a regularized error function given by

$$\frac{1}{2}\sum_{n=1}^{N}\{y_n - t_n\}^2 + \frac{\lambda}{2}||w||^2$$

In laymans term we can say that SVM is Support Vector Machine (SVM) when used for classification and by SVR you mean SVM for regression. The main difference comes in the slack variables used in the 2 techniques. SVM for classification involves assigning one slack variable to each training data point, whereas in SVM for regression, there are two slack variables for each training data point[6].

The optimization function for SVM for classification is given by

$$C\sum_{n=1}^{N}\xi_n + \frac{1}{2}||w||^2$$

and the optimization function for SVM for regression is given by

$$C\sum_{n=1}^{N}(\xi_n + \hat{\xi_n} + \frac{1}{2}||w||^2$$

Figures 15 and 16 show the slack variable demonstrations used in SVM for Classification and Regression respectively.

Figure 17 shows the SVM curve fit on the observed confirmed cases in India and Table II is the prediction using SVM Regression.

Root Mean Square Error observed for Support Vector Machine Regressor: 1717.715

F. Auto-Regressive Model

Consider a stationary time series X_t. Model predicts future behavior based on past behavior. The value of the outcome variable (Y_t) at some point t in time is like regular linear regression directly related to the predictor variable (X_t). Where simple linear regression and AR models differ is that Y is dependent on X and previous values for Y. If a time series X_t is stationary and can be explained by linear combinations

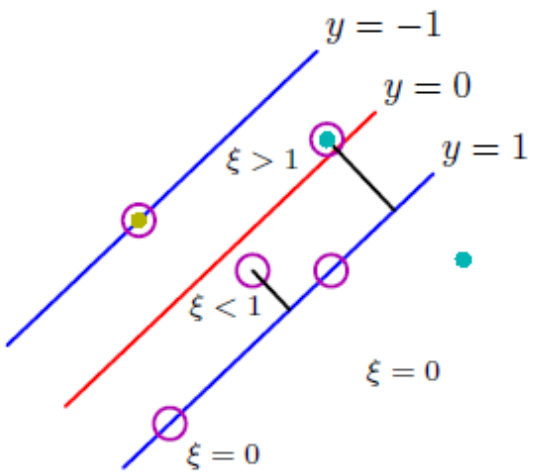

Fig. 15. SVM for Classification

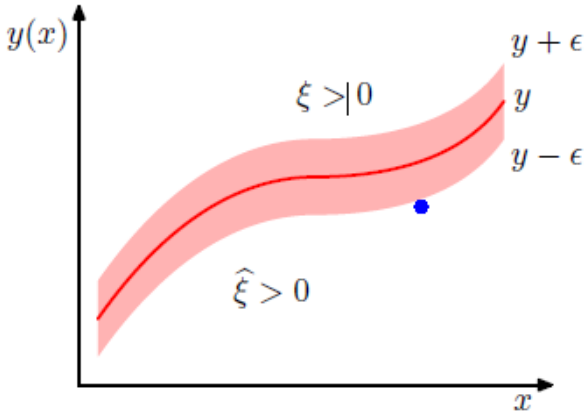

Fig. 16. SVM for Regression

of the past values X_{t-1} and some residual noise term, then the linear model is referred to as an Auto-Regressive (AR) model.

For example,

$$X_t = a_0 + a_1 X_{t-1} + ... + a_p X_{t-p} + \epsilon_t$$

Sr. No.	Date	SVR Prediction
0	2020-05-04	47334.31
1	2020-05-05	50408.70
2	2020-05-06	53647.45
3	2020-05-07	57057.53
4	2020-05-08	60646.09

TABLE II
SUPPORT VECTOR REGRESSION PREDICTION FOR CONFIRMED CASES IN INDIA

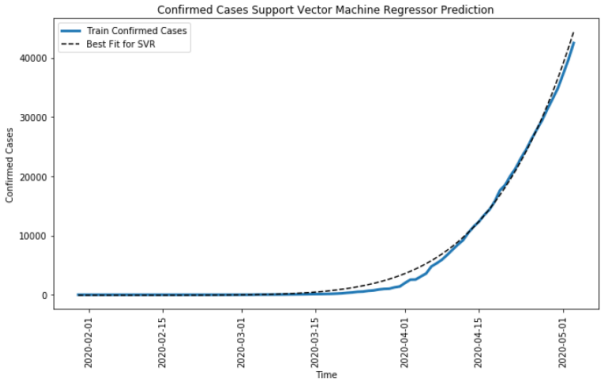

Fig. 17. SVM Prediction curve for India

Where t is a random noise term, a_i are the linear coefficients with the past values and the order p is referred as AR(p) model.

As we observe the trend of confirmed cases in the data visualization section, we can assert that the confirmed cases have an increasing trend. To remove the trend, we take log of the series at lag-1, this makes the series stationary in-order to apply AR model.

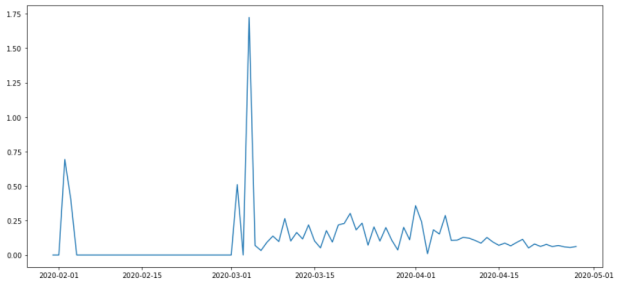

Fig. 18. Log series of Confirmed Cases in India

AR (1) Model:

$$X_t = a_0 + a_1 X_{t-1} + \epsilon_t$$

The correlation factor between X_t and X_{t-1} is called the lag-l auto-correlation function (ACF) of X_t.

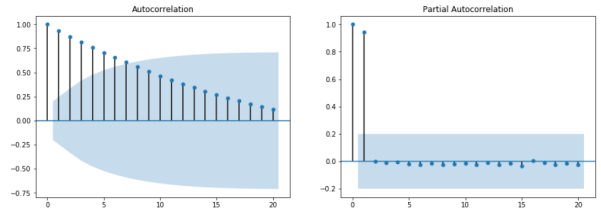

Fig. 19. ACF and PACF Plots for Confirmed Cases in India

In application, the order p of an AR time series is unknown. It must be specified empirically. This is referred to as the order determination (or order specification) of AR models, and it has been extensively studied in the time series literature. Two general approaches are available for determining the value of p. The first approach is to use the partial Auto-correlation function, and the second approach uses some information criteria. The Partial Auto-correlation Function (PACF) of a stationary time series is a function of its ACF and is a useful tool for determining the order p of an AR model.

Let us consider the following AR models in consecutive orders:

$$X_t = a_{0,1} + a_{1,1} X_{t-1} + \epsilon_{1t}$$

$$X_t = a_{0,2} + a_{1,2} X_{t-1} + a_{2,2} X_{t-2} + \epsilon_{2t}$$

$$X_t = a_{0,3} + a_{1,3} X_{t-1} + a_{2,3} X_{t-2} + a_{3,3} X_{t-3} + \epsilon_{3t}$$

...where $a_{0,j}$, $a_{i,j}$, and ϵ_{jt} are, respectively, the constant term, the coefficient of X_{ti}, and the error term of an AR(j) model. These models are in the form of a multiple linear regression and can be estimated by the least-squares method. The estimate $a_{1,1}$ of the first equation is called the lag-1 sample PACF of X_t. The estimate $a_{2,2}$ of the second equation is the lag-2 sample PACF of X_t. The estimate $a_{3,3}$ of the third equation is the lag-3 sample PACF of X_t, and so on. From the definition, the lag-2 PACF $a_{2,2}$ shows the added contribution of X_{t-2} to X_t over the AR (1) model X_t = a_0 + $a_1 X_{t-1}$ + ϵ_{1t} .

The lag-3 PACF shows the added contribution of X_{t-3} to X_t over an AR (2) model, and so on. Therefore, for an AR(p) model, the lag-p sample PACF should not be zero, but $a_{j,j}$ should be close to zero for all j $>$ p. We make use of this property to determine the order p.

Figure 19 shows the ACF and PACF of the stationary log series. We can observe the cliff in the PACF plot at lag 1. Hence, the series is AR (1) model.

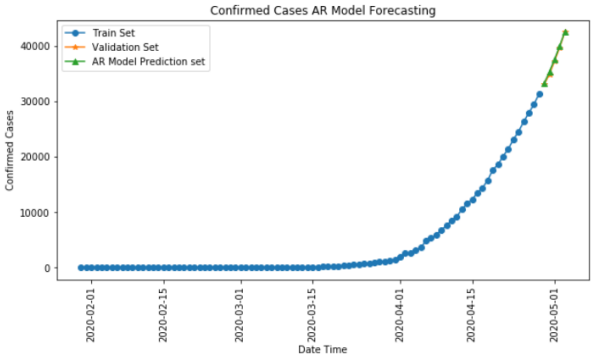

Fig. 20. Confirmed Cases in India AR Model Forecasting

Figure 20, shows the AR Model Forecasting on the observed confirmed cases and Table 3 shows the prediction using AR(1) Model.

Root Mean Square Error observed for AR Model: 271.878

G. Moving Averages Model

A moving average model uses past forecast errors in a regression-like model, rather than using past values of the

Sr. No.	Date	AR Prediction
0	2020-05-04	45307.84
1	2020-05-05	48301.16
2	2020-05-06	51530.87
3	2020-05-07	55011.43
4	2020-05-08	58768.24

TABLE III
AUTO-REGRESSION PREDICTION FOR CONFIRMED CASES IN INDIA

Sr. No.	Date	MA Prediction
0	2020-05-04	54984.42
1	2020-05-05	60396.79
2	2020-05-06	66343.98
3	2020-05-07	72879.03
4	2020-05-08	80060.29

TABLE IV
MOVING AVERAGE PREDICTION FOR CONFIRMED CASES IN INDIA

forecast variable in a regression. A moving average process MA(q) is one where,

$$X_t = c + \epsilon_t + b_1\epsilon_{t-1} + ... + b_q\epsilon_{t-q}$$

We showed above that a stationary AR (1) process can be transformed into an infinite MA process going back to 0. In general, if the AR(p) characteristic equation has roots inside the unit circle, the (therefore) stationary process can be transformed into such an infinite MA process. If we define the characteristic polynomial for an MA(q) process as $z_q + b_1 z_{q1} + ... + b_q = 0$. Then if its roots are inside the unit circle (absolute value less than one), we can invert the MA(q) process into an AR process. This AR process will be stationary. If we observe, the cliff in the ACF and a continuous series in PACF, this is clear indication that the model is MA(q) model and not AR model. Therefore, cliff position is observed as q and same will the order of MA model.

Figure 19 shows the cliff in PACF and not ACF, hence the model is AR and MA. Since, the ACF and PACF plots are made for the confirmed cases of India, they could vary if we try to predict for any other country. We could or could not observe the cliff in ACF and not PACF or a continuous series in ACF and PACF, therefore the model used will be MA or ARIMA respectively for such stationary time series.

If we use MA model for the forecasting, we will observe the considerable amount of difference between the observed and the predicted value which can be seen in figure 21.

Root Mean Square Error observed for Moving Averages Model: 4896.179

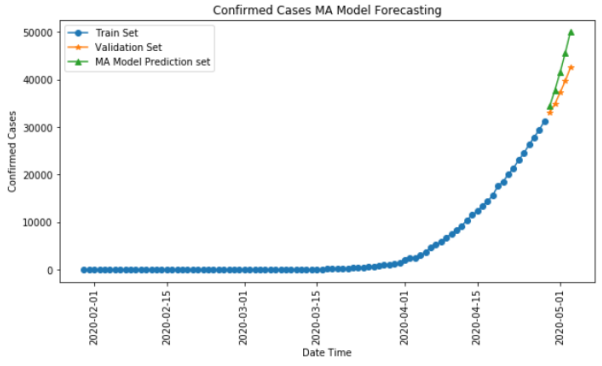

Fig. 21. Confirmed Cases in India MA Model Forecasting

Sr. No.	Date	ARIMA Prediction
0	2020-05-04	52022.61
1	2020-05-05	56601.23
2	2020-05-06	61540.81
3	2020-05-07	66933.26
4	2020-05-08	72763.48

TABLE V
ARIMA PREDICTION FOR CONFIRMED CASES IN INDIA

H. ARIMA Model

Combining AR and MA processes, we can define an ARMA(p,q) process as one where,

$$X_t = c + a_1X_{t-1} + ... + a_pX_{t-p} + \epsilon_t - b_1\epsilon_{t-1} - ... - b_q\epsilon_{t-q}$$

It is convenient to introduce the backward and forward operators B and F. For any time-series Y_t, $BY_t = Y_{t1}$, and $FY_t = Y_{t+1}$. B is sometimes called the lag operator. We can therefore rewrite an ARMA(p,q) process as

$$X_t = c + (\sum_{i-1}^{p} a_iB^i)X_t + \epsilon_t - (\sum_{i-1}^{q} b_iB^i)\epsilon_t$$

We determine the p and q, by observing the ACF and PACF. If there is no cliff observed in ACF or PACF, then we can be assured that the model is ARIMA and we try and tweak the p and q such as we get the lowest value of RMSE.

Figure 22 shows the confirmed cases forecasting using ARIMA model. As the series is not ARIMA we could observe the considerable amount of difference between the observed and the predicted value.

Root Mean Square Error observed for Moving Averages Model: 3544.864

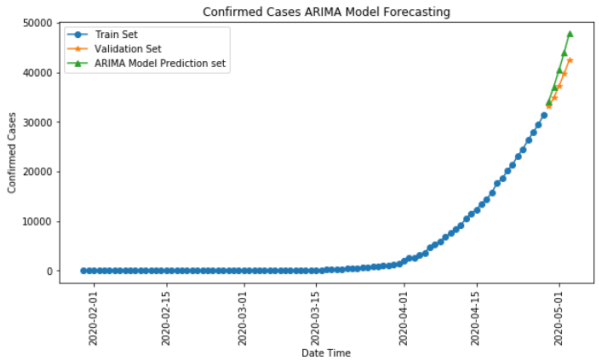

Fig. 22. Confirmed Cases in India ARIMA Model Forecasting

Sr. No.	Model	RMSE Value
1	Polynomial Regression	517.246885
2	Support Vector Regression	1717.715671
3	Auto-Regression	271.878565
4	Moving Average	4896.179313
5	ARIMA	3544.864093

TABLE VI
RMSE VALUES FOR DIFFERENT MODELS

Date	PR	SVM	AR	Official Count
2020-05-04	43896	47334	45308	46434
2020-05-05	46486	50409	48301	49405
2020-05-06	49288	53647	51531	53007
2020-05-07	52340	57057	55011	56351
2020-05-08	55689	60646	58768	59690

TABLE VII
PREDICTIONS COMPARISON WITH ACTUAL COUNT

IV. COMPARISON OF RESULTS

We compared our results of every model we designed by comparing the RMSE values over the same validation set for the data-set for India which can be seen in Table VI.

While the AR model clearly gave us the best result for Confirmed cases in India, we will not discard the other models just yet, primarily because the time series model might change over the coming days. Table number 7 shows the compiled predictions for all the models.

V. FUTURE RESEARCH

In this paper, we only managed to read through the scholarly articles. In future research analysis, we could also consider the local news from a country and monitor the effects of a change in a countrys policy like enforcing lock-down, partial/complete economic closures, travel bans, etc., and their impact in the mitigation of the virus. This will make the prediction and modeling more refined.

Also monitoring the changes in the availability of equipment and PPE could help reduce the death rate in a country from that period on wards. Since our knowledge of medical field was limited, we could not understand or incorporate the effect of the environment in a country and its contribution to the infection spread. This could also be done by the future researchers.

VI. CONCLUSION

Our paper is written during the most uncertain time during a lifetime of a virus. We dont know if the virus has peaked or not, no idea on when a potential vaccine for it might be available, and a major uncertainty over whether there will be secondary waves of the virus or not.

This project is solely based around the analysis of the data that has been readily available to us during the past few weeks. The understanding and implementation of the various algorithms used was our main motive. The accuracy and our results are secondary to us.

REFERENCES

[1] CORD – 19 research data-set from kaggle, https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[2] John Hopkins CSSE Covid – 19 Data, https://github.com/CSSEGISandData/COVID-19

[3] COVID – 19 brings Cytokine Storm, https://www.idse.net/Covid-19/Article/03-20/COVID-19-Brings-Cytokine-Storm/58061

[4] Fighting Corona-virus – Ray Bio Tech , https://www.raybiotech.com/coronavirus-research-products-covid-19/

[5] World Population densities (for Countries), https://worldpopulationreview.com/countries/countries-by-density/

[6] Pattern Recognition and Machine Learning, Christopher M. Bishop