A Big Impact of Social Network Analysis and Machine Learning Algorithms for Predicting Facts of Covid-19 Pandemic

Sonam and Surjeet Kumar

1 Introduction

On the basis of report covered under 'survey of Global Digital Growth 2019–20', there were 304 million people as an active social network users, of which, 2498 people were Facebook users. Therefore, Facebook is a good platform from where people may access to the facts about Covid-19. People share their physical and psychological impact on Facebook portal and subsequently follow the suggestions [1]. People are joining with like, comment, and share on all the posts by WHO at Facebook platform and getting updates related to Covid-19 every day. The algorithm has a crucial role for the advancement of Facebook posts. Day by day, varieties of useful data are handled by using 'big data' analysis [2]. Facebook generates 4 Petabyte data each day. All stored data known as hive that contains 300 Petabyte of data. After analyzing social network we can boost our attractive and quantitative graph metric that provides relationship between nodes and edges. We generally face major challenges choosing appropriate algorithm for social network analysis and prediction. Machine learning algorithms have vital role to make all information available on one platform just like Facebook.

2 Impact of Algorithm on Social Network

Every Facebook user's action starts with commenting on photos, liking to post messages or posting videos [3]. All are shown by algorithm in the final analysis. The goal is to satisfy everyone to ensure the customers only receive information that is valuable to them [4].

Sonam (✉) · S. Kumar
Department of Computer Applications, VBS Purvanchal University, Jaunpur, UP, India

R. Agrawal et al. (eds.), *Cyber Security in Intelligent Computing and Communications*,
Studies in Computational Intelligence 1007,
https://doi.org/10.1007/978-981-16-8012-0_8

Since Facebook algorithms esteem the quality of post and express main factors of audience engagement [5]. Then, work of algorithm on Facebook is [6]:

- Affection
- Variety of theme or text
- Communication
- Contemporary
- Variation
- Productive outcome by Links

3 Major Role of Covid-19 Information Center Posts on Facebook

We have identified Covid-19 information center menu that is available on Facebook [7]. It contains.

Latest updating post—post source is WHO related to India and global Covid-19 cases.

Fact about Covid-19—WHO posts are accurate free from day to day rumors about covid-19.

Recent post—Post from government community and public health organization pages such as recent post on Facebook by WHO, MyGovCoronaHub, MoHFW India, UNICEF, and ABHWCs, etc.

All the request, help, suggestions, and prevention tips are available on Facebook about Covid-19. We collected some unique post from Facebook through WHO pages and plotting data for" like", "comment" and "share". We are showing by Fig. 1.

There are 6 lakhs "likes" on the posts, posted by WHO at Facebook.

According to comment distribution, there are 10,000 comments on posts that are shown in Fig. 2.

There are 50,000 people, sharing WHO posts that is shown in Fig. 3.

4 Using Networkx Technique for Analysis of Facebook

We are using "Networkx" [8] for analyzing social network that draw a specific graph for node and edges on the basis of attributes which are available in dataset [9]. Graph G works as a container for adding the collected form of edges and nodes [10].

It is expressed by properties of graph such as G.node, G.edge and G.degree used to express every value [11]. We generated graph to identify all nodes for multiple public health organization and expressing node.

On the base of WHO Post, the list of all nodes, edges, and degree values is shown in Table 1.

Graph shown in below that expresses all nodes and edges (Fig. 4).

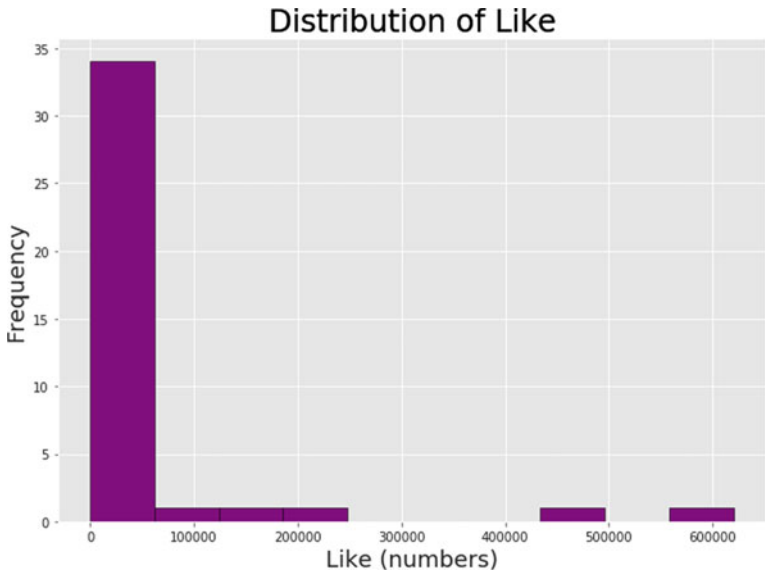

Fig. 1 Distribution of Like

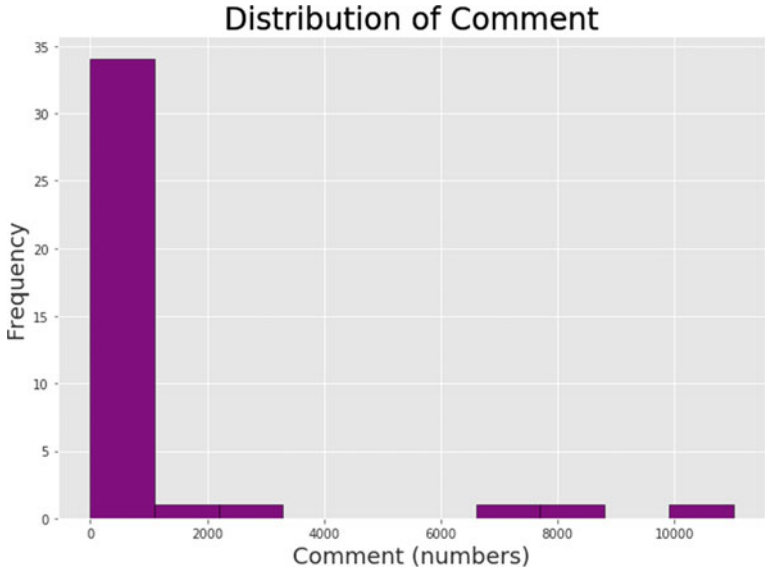

Fig. 2 Distribution of Comment

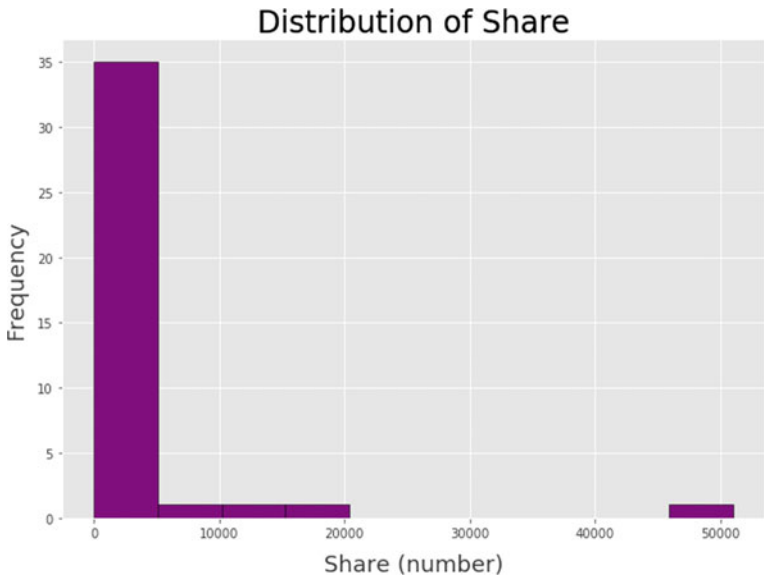

Fig. 3 Distribution of Share

5 Experimental Work and Deliberation of Algorithms

5.1 Random Forest Regressor to Predict Interaction Among People with Public Health Organization

Integrating several decision trees into one model that is the random forest regression technique which is a supervised machine learning method [12]. It mostly uses an ensemble learning algorithm [13]. It also makes a decision tree accompanying of training and generates results as a mean prediction. Hyper parameter [14] refers to attributes which is used for splitting data and identifying accurate percentage.

A big advantage of this ensemble method [15] is to recognize all predicted features. Output would be generated as an average into a specific ensemble property that gets output for every decision tree [8].

We collected 500 posts from all public health organization that is accessible on Facebook social networking sites. We applied algorithm of random forest regressor [16] with splitting train and test with number of estimators is 50 and random state 1 Correlated data are depicted from correlation map among all features. It predicts accurate values on the basis of independent and dependent variable and identifies approximately 6 lakh people are interacting with sWHO posts among 500 posts.

People are interacting with all public health organization on the base of their post shown in Fig. 5.

Table 1 Feature of networkx

Total _Number_of _nodes: 60
Total_Number_of_edges: 10
List of all nodes: [‘1’, ‘2’, ‘3’, ‘4’, ‘5’, ‘6’, ‘7’, ‘8’, ‘9’, ‘10’, ‘11’, ‘12’, ‘13’, ‘14’, ‘15’, ‘16’, ‘17’, ‘18’, ‘19’, ‘20’, ‘21’, ‘22’, ‘23’, ‘24’, ‘25’, ‘26’, ‘27’, ‘28’, ‘29’, ‘30’, ‘31’, ‘32’, ‘33’, ‘34’, ‘35’, ‘36’, ‘37’, ‘38’, ‘39’, ‘40’, ‘41’, ‘42’, ‘43’, ‘44’, ‘45’, ‘46’, ‘47’, ‘48’, ‘49’, ‘50’, ‘60,400’, 1, 2, 6, 3, 4, 5, 8, 9, 7]
List of all edges: [(1, 2, {}), (1, 6, {}), (2, 3, {}), (2, 4, {}), (2, 6, {}), (6, 7, {}), (3, 4, {}), (3, 5, {}), (4, 8, {}), (4, 9, {})]
Degree for all nodes: {‘1’: 0, ‘2’: 0, ‘3’: 0, ‘4’: 0, ‘5’: 0, ‘6’: 0, ‘7’: 0, ‘8’: 0, ‘9’: 0, ‘10’: 0, ‘11’: 0, ‘12’: 0, ‘13’: 0, ‘14’: 0, ‘15’: 0, ‘16’: 0, ‘17’: 0, ‘18’: 0, ‘19’: 0, ‘20’: 0, ‘21’: 0, ‘22’: 0, ‘23’: 0, ‘24’: 0, ‘25’: 0, ‘26’: 0, ‘27’: 0, ‘28’: 0, ‘29’: 0, ‘30’: 0, ‘31’: 0, ‘32’: 0, ‘33’: 0, ‘34’: 0, ‘35’: 0, ‘36’: 0, ‘37’: 0, ‘38’: 0, ‘39’: 0, ‘40’: 0, ‘41’: 0, ‘42’: 0, ‘43’: 0, ‘44’: 0, ‘45’: 0, ‘46’: 0, ‘47’: 0, ‘48’: 0, ‘49’: 0, ‘50’: 0, ‘60,400’: 0, 1: 2, 2:4, 6: 3, 3: 3, 4: 4, 5: 1, 8: 1, 9: 1, 7: 1}
Total_Number_of_self-loops: 0
List of all nodes with self-loops: []
List of all nodes we can go to in a single step from node 1:50 [2, 3, 5, 6]

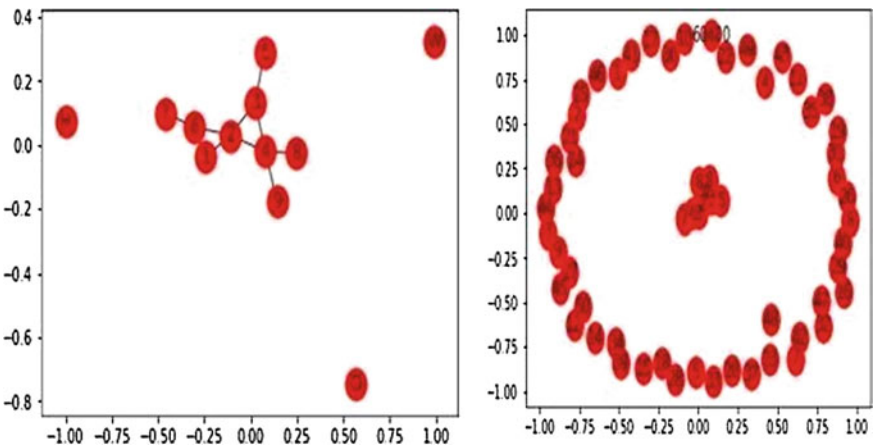

Fig. 4 Networkx: nodes and edges

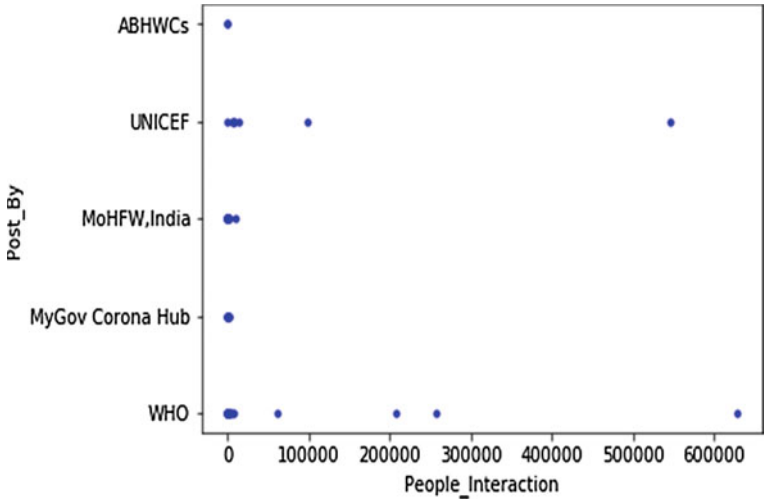

Fig. 5 People interaction with all health community

Interaction between people with WHO post that is depicted on plot shown in Fig. 6.

On the base of our dataset we applied train and test after that we identified predicted value (Fig. 7).

After training dataset we applied algorithm for relation between predicted and test values shown in Fig. 8.

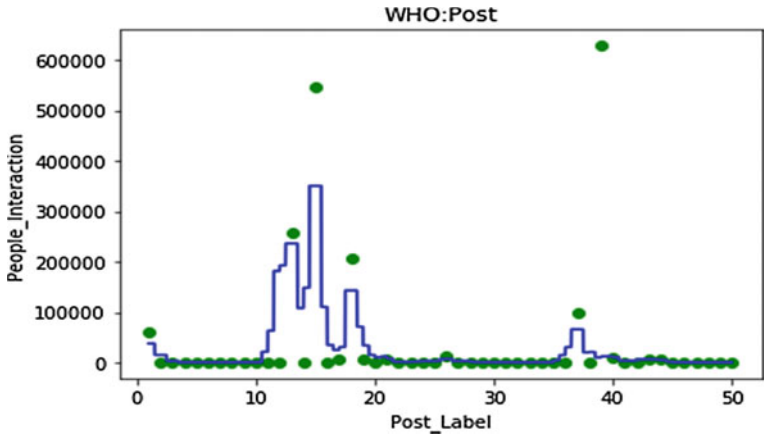

Fig. 6 People interaction with WHO

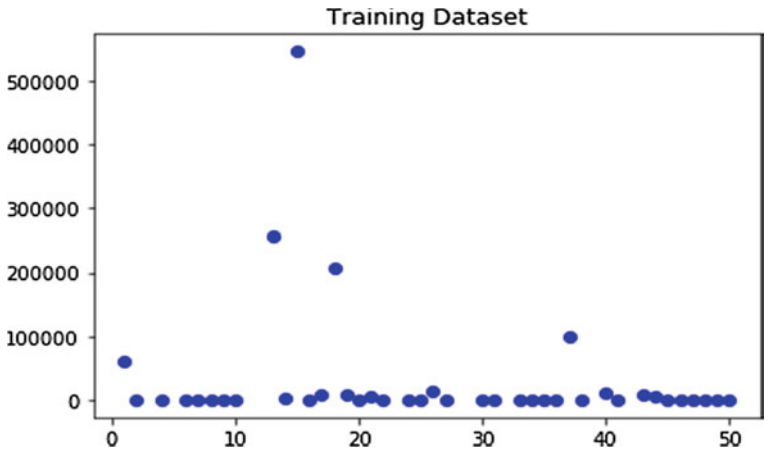

Fig. 7 Analyze for training data

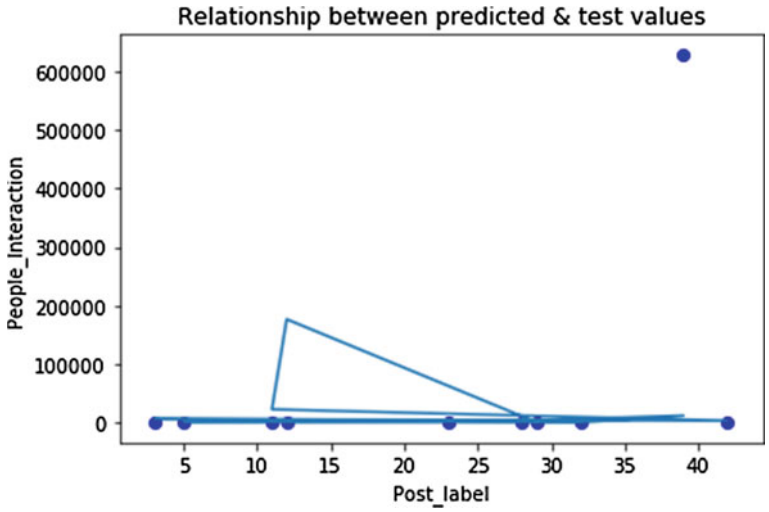

Fig. 8 Analysis for test data and predicted value

5.2 Analysis Feature of Time Series Forecasting: Arima and Sarima Model

A time series is a sequencing step of recording metrics at regular intervals, whereas forecasting is nothing but which step and on where we identify to predict future values that the series will adopt. We have applied model that is Arima [17] and Sarima to predict new case of Covid-19 disease till October.

Arima model is the best forecasting method. It refers as automatic regression integrated moving average which is a forecasting algorithm that is based on some ideas such as data available according to the previous values of time series for predicting future values. Arima model can be applied that exhibits pattern by non-seasonal time series and are not random white noise. Autoregressive [18] as a linearly regression model that uses lags as predictor.

An Arima model combines three methods—Autoregressive, Integrated, and moving average that takes 3 parameters such as p, d, q.

p—Order of automatic regression terms.

d—Difference that requires for making time series stationary.

q—Order of moving average terms. It refers to the number of lagging prediction errors that should be entered into the ARIMA model.

Purely, autoregressive model in which Yt refers as past values that have own lags [19]. That's why Yt is used as a function for Lags of Yt. Arima algorithm is identified by formula such as Eq. 1.

$$Yt = \alpha + \beta 1\ Yt - 1 + \beta 2\ Yt - 2 + ... \beta p\ Yt - p + €1 \tag{1}$$

In this model, the time series differs at least once to stabilize it, and then AR and MA terms are combined. Therefore, Predicted value of Yt = constant + linear combination lags of Y (up to p lags) + linear combination of lagged forecast errors (up to q lags). That's why equation becomes as Eq. 2.

$$Yt = \alpha + \beta 1\ Yt - 1 + \beta 2\ Yt - 2 + \ ... \beta p\ Yt - p\ €t + \Phi 1\ €t - 1 + \Phi 2\ €t - 2 + .. + \Phi q\ €t - q \tag{2}$$

5.3 How Arima Model Works on Dataset?

At first we collected data for predicting new case of Covid-19 disease for India, downloaded from WHO link [20] https://covid19.who.int/region/searo/country/in. We found 184 rows and 5 columns of data from January to July. After that we identified new cases from 30th January to 31st July, 2020. The best solution is applying common method to differentiate it. We are showing dataset as a plot expresses as (Fig. 9).

According to plotting figure we have found everyday cases are different so the series is not stationary, it can be made after differencing. After differencing once, series is called a unified order 1 and denotes as I(1) or I(d)1. If autocorrelation is positive for several delays such as 10 or above, sequences need to be different. In another hand, if lag 1 autocorrelation is too negative, series may be overly differentiated. According to p-value time series [19] is indeed stationary.

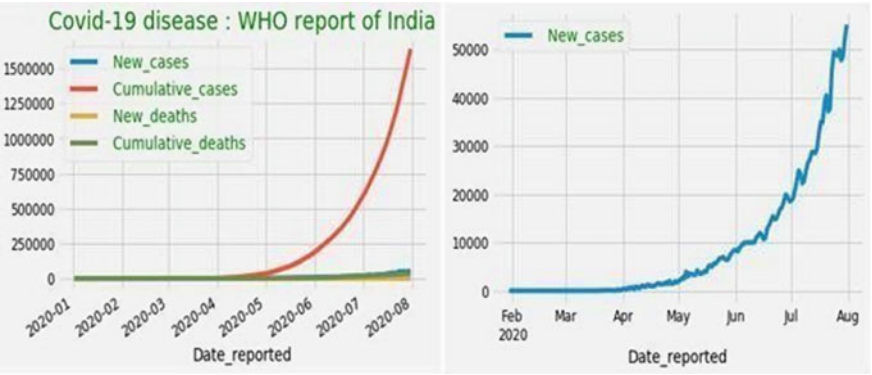

Fig. 9 Plotting data from WHO dataset

So, convert from series to stationary, the method is differentiation that means subtract past value from current value. The value of d is least number of differentiation need to make from series to stationary. If d = 0, then time series is stationary.

In our cases, we calculated the differentiates of data frame values that is comparing with other next value.

That is cases_diff = cases.diff (periods = 1) we used periods = 1 shifted to calculate difference that accepts neutralize value that is showing on autocorrelation figure.

Autocorrelation is the similarity between observations as a function at the time lag between them –1 to +1. The AIC is defined by a simple equation from the sum of squares and aggregate degree of freedom for the two models.

Mean square error indicates distances between regression line and set of points. Distances refer as error that mean using mean squared error we found average according to set of errors. We used Arima order to identify AIC value and extract mean square error [21]. After fitting Arima model. We deducted AIC values and minimum value indicates better model and predicted value on test dataset [22] (Fig. 10).

Fig. 10 Autoregressive: test and prediction

P = d = q = range (0, 2) is better for predicting AIC value that's order is (0, 1, 1) and same as AIC value with fitting ARIMA model [23].

To determine a proper model for a given time series data, it is necessary to carry out the ACF analysis.

5.4 *Seasonal Autoregressive Integrated Moving Average (SARIMA) Model*

This proposed model is known as the Seasonal ARIMA (SARIMA) model. In this model seasonal differencing of appropriate order is used to eliminate non-stationarity from the series. A first-order seasonal difference is the difference between an observation and the corresponding observation from the previous year and is calculated as Z t = Y t – Y t – S during month-wise time series $s = 12$ and for quarterly time series $s = 4$.

This model is generally termed as the SARIMA (p, d, q) × (*P, D, Q*) *s* model. The seasonal period, s, defines the number of observations that make up a seasonal cycle. The value of s is fixed in the series. If we have daily data and the seasonal period is the length of the month, s will be approximately 30, but it will vary from month to month. Histogram features is shown in Fig. 11.

With daily data we can have weekly seasonality, with s = 7, monthly, with s = 30 and yearly, with s = 365. We have assumed that there is only one type of seasonality and at the ending of section we will observe how to extend the methods presented to various seasonal periods [24] (Fig. 12).

Using algorithm we have observed forecast analysis which successfully predict values for upcoming months from August to October (Fig. 13).

Finally, we have predicted New Cases value is up to 104,894 on Date 23/10/2020 that would be approximate value of WHO dataset.

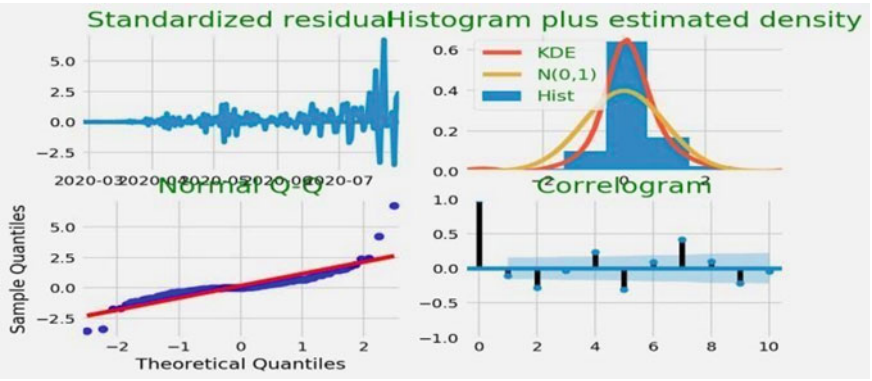

Fig. 11 Histogram features

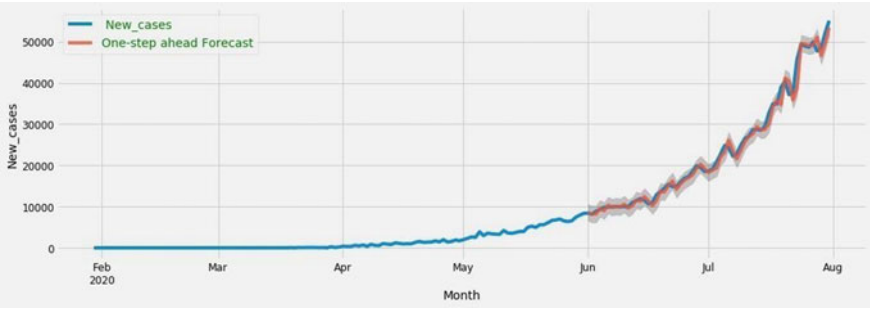

Fig. 12 Analysis for daily cases

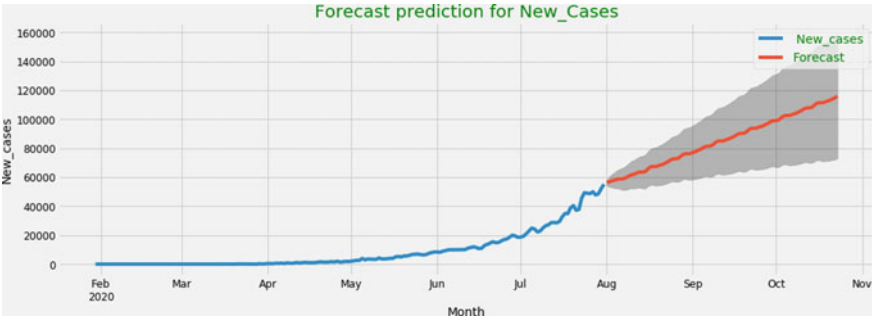

Fig. 13 Forecast future prediction for new cases

6 Conclusion

This work is an attempt to elucidate the importance of employing machine learning tools for getting a meaningful result to receiving ends. The work at first collects dataset of interaction between people and all the respective agencies like AB- HWCs, UNICEF, MoHF, MyGov corona Hub and WHO and proves through applying random forest regressor technique that out of all the agencies, WHO receiving maximum interaction. Also, our work validates the previous notion about the credibility of the two presently available tools ARIMA and SARIMA for future prediction. Here it has been proved by mining dataset up from January to July 2020 and employing the machine learning tool ARIMA and SARIMA to predict new Covid cases up to October 2020. By varying further about the predicted result with WHO database about new cases, it matches approximately validating these two open-source tools. The prediction can generously be used for deploying help, prevention, and propagating public awareness about this pandemic Covid-19.

References

1. G. Bello-Orgaz, J. Jung, D. Camacho, Social big data: recent achievements and new challenges. J. Info. Fusion **28**, 45–59 (2016)
2. H.A. Chalabi, U.C. Apoki, A. Hibah, A. Alsaad, Big data analysis using social networks, in *Conference* (IEEE, Iraq, 2017)
3. H.J. Esfahani, K. Tavasoli, A. Jabbarzadeh, Big data and social media: a scientometrics analysis. J. Dat. Net. Sci. **3**, 145–164 (2019)
4. S. Stieglitz, M. Mirbabaie, B. Ross, C. Neuberger, Social media analytics–challenges in topic discovery, data collection, and data preparation. J. Inf. Manag. **39**, 156–168 (2018)
5. M.M. Mariani, M. Di Felice, M. Mura, Facebook as a destination marketing tool: evidence from Italian regional destination management organizations. J. Tour. Manag. **54**, 321–343 (2016)
6. R. Devakunchari, C. Valliyammai, Big social data analytics: opportunities, challenges and implications on society, in *Conference on Communication, Media, Technology and Design*, Zagreb–Croatia (2016), pp. 27–29
7. Fact of Covid-19 information center, https://www.Facebook.com/coronavirus_info/?page_source=bookmark
8. K.S. Sonam, Analyzing and predicting social networking big data: using network and regression techniques. J. Adv. Sci. Technol. **29**(03), 8087–8096 (2020)
9. R.K. Devi, A machine learning-based online social network analysis for 360-degree user profiling. J. Innov. Tech. **9**(2S2), 2278–3075 (2019)
10. N.B. Lassen, L. Cour, R. Vatrapu, Predictive analytics with social media data, BK-SAGE-SLOAN_QUAN-HAASE-160238-CHP20.indd, in The *Sage Handbook of Social Media Research Methods* (2016)
11. T. Kaushik, S. Singhal, J. Mandan, K. Sharma, Social networking analysis: a case study in tools. J. Eng. Adv. Tech. **8**(2), 2249–8958 (2018)
12. C.C. Hsu, Y.C. Lee, P.E. Lu, S.S. Lu, Social media prediction based on residual learning and random forest, in *Conference. SERSC* (2017)
13. X. Gao, J. Wen, C. Zhang, *An Improved Random Forest Algorithm for Predicting Employee Turnover* (2019)
14. M. Fadhil, P. Andras, A systematic analysis of random forest based social media spam classification, in *11th International Conference, NSS 2017, Proceedings*, Helsinki, Finland (2017), pp. 427–438
15. K. Sridevi, B.V.S. Samrat, S. Srihari, Traffic analysis by using random forest algorithm considering social media platforms. J. Res. Tech. **7**(6S), 2277–3878 (2019)
16. V.M. Herrera, M. Taghi, F.B. Khoshgoftaar, Random forest implementation and optimization for big data analytics on Lexis Nexis's high performance computing cluster platform. J. Big Data **68** (2019)
17. V. Chaurasia, S. Pal, Application of machine learning time series analysis for prediction of Covid-19 pandemic. Res. Biomed. Eng. (2020)
18. L.V.D. Alquisola, J.A.B. Coronel, M.F. Reolope, J.N.A. Roque, Pre-diction and visualization of the disaster risks in the Philippines using discrete wavelet transform (DWT), autoregressive integrated moving average (ARIMA), and artificial neural network (ANN), in *3rd International Conference on Computer and Communication Systems (ICCCS)* (2018), pp. 146–149
19. R. Adhikari, R.K. Agrawal, An introductory study on time series modeling and forecasting. J. LAP (Lambert Academic Publishing, Germany) (2013)
20. Collection of dataset for covid19 cases, https://covid19.who.int/dataset
21. Q. Yang, X. Wang, Research on covid-19 based ARIMA MODEL—Taking Hubei, China as an example to see the epidemic in Italy. J. Inf. Pub. Health (2020)
22. D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, M. Ciccozzi, Application of the Arima model on Covid2019 epidemic dataset. J. Data In Brief **29** (2020)

23. A. Hernandez-Matamoros, H. Fujita, P. Meana, Forecasting of covid19 per regions using Arima models and polynomial functions. J. Elsev. Pub. Health Emerg. Collect. **96** (2020)
24. A. Tarsitano, I.L. Amerise, Short-term load forecasting using a two-stage sarimax model. J. Econpaper Energy **133**, 108–114 (2017)