Original Research Article

J. Computing & Int. Systems (2022) 320–324

SACRED HEART RESEARCH PUBLICATIONS

Journal of Computing and Intelligent Systems

Journal homepage: www.shcpub.edu.in

ISSN: 2456-9496

COMPARATIVE ANALYSIS OF BIG DATA RANDOM FOREST AND DECISION TREE ALGORITHMS FOR COVID-19 DISEASE PREDICTION

J Roserin Thiravia[#1], A Priyanga[#2]

Received on 29 NOV 2022, Accepted on 11 DEC 2022

Abstract — Big data is extremely admired at the present time because information can be easily removed from evaluating huge amounts of data. The essential problem is resources. It cannot handle static resources. Big data is one of the main conflicts in statistical science and has many consequences from both algorithmic and theoretical view points. More than five million cases have been documented in more than 200 countries. The Asian countries reported 116,368,471 cases of COVID-19 and 1,277,538 deaths between 2019 and 2022. In Europe, there were 180,627,256 COVID-19 cases and 2,203,389 deaths. As a result, More than five million cases have been documented in more than 200 countries. Comparative analysis of machine learning Random Forest And Decision tree algorithms for covid-19 disease prediction. Using Asian and European countries data. This research looks at how decision tree and random forest algorithms, as well as rapid miner tools, may be used to anticipate COVID-19 instances in Asia and Europe. The decision tree algorithm, according to the findings, outperforms the random forest method.

Keywords - COVID-19 cases, Big data, Asia and Europe, Decision tree, Random forest.

I. INTRODUCTION

Big data analytics is a trendy topic in academia and industry right now. This is a cutting-edge technology for extracting a wide range of information from large amounts of data. Use of analytics in a variety of industries it's also a brand-new research and development area. Researchers can use big data to help them come up with a solution.. This article will look as well as the potential that these methods and tools provide. Big data analytics is being used in a variety of decision domains, which is providing benefits. The corona virus (COVID-19) was first detected. The virus was first discovered in a mild episode of pneumonia. Number of corona virus cases increased dramatically after that. Chinese health experts determined that it was a member of the corona virus family. The epidemic soon spread around the globe. The Latin term "corona" refers to a group of RNA viruses, which is how this virus got its name. Furthermore, the virus increases mortality and has severe economic implications. By May 17, 2020, the WHO forecasts that there will be over four million verified illnesses and 300,000 confirmed deaths worldwide. Using vaccines and lockout measures, the WHO was able to limit the corona virus cases. A lot of cases went unreported as well. A more thorough decision-making alternative is required for pandemic prediction.

The reliability and stability of the suggested approach are investigated and validated using a measurement. The proposed method uses decision tree and random forest classification algorithms. In this, the decision-level approach's accuracy should then be compared against the current relevant work's accuracy.

II. RELATED WORKS

Innumerable countries have experienced a financial crisis as a result of COVID-19 [1]. Face of large-sized data and variety of methodologies have yielded great results on image classification challenges. A CNN-based deep learning technique was applied [4-6]. To examine machine learning approaches, including Corona virus reports [7]. Forecasting model to define the severity of the COVID-19 disease on Canadians [8]. For forecasting the extensive distribution of COVID-19 cases, researchers used linear regression [9]. The ensemble model shows exceptional robustness [11]. Combining machine learning techniques to extract COVID-19 symptoms from a textual clinical report [12]. For predicting COVID-19-related deaths , Generate a large amount of data that may used to analyze processes[13].

The current situation caused by the corona virus epidemic, as well as the topic of data security [15]. It is critical to note that COVID-19 compliance necessitates the use of personal data and adherence to the GDPR [16]. Other academics have concentrated on studying and modeling disease transmission among people around world in order to understand forecast infection and mortality rates [17].

* Corresponding author: E-mail: [1]gthiraviajohn1999@gmail.com, [2]rithupriya16@gmail.com

[1&2] Research Scholar, Department of Computer Science, Dr. Umayal Ramanathan College for Women, Karaikudi , Tamilnadu, India.

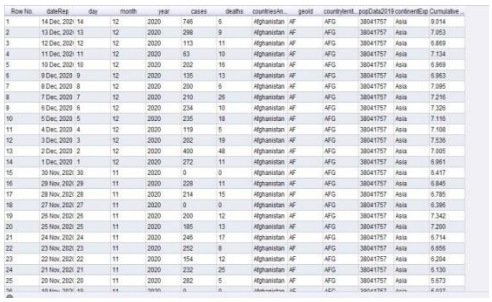
The study presents Prediction model for Pulmonary Arterial Hypertension (PAH) [18]. It also included a survivorship model for people with breast cancer. Using multi model ensemble method lungs, stomach and breast cancer are predicted.[20].Breast cancer data was collected for prediction [21]. Aid in the planning of how these vaccinations should be used [22].

Indonesia's government agreement with a number of vaccine manufacturers [23]. It is commonly used to categorize popular opinion [24]. Systematic Literature Review (SLR) is a strategy to finding, specific research questions [25-27]

III. MATERIALS AND METHODS

A. DATA COLLECTION

The COVID-19 data set attributes are Date, Day, Month, Year, Cases, Deaths, Countries, Geold, Country type, Pop statistics, Continent, Cumulative number are among the elements listed. This can be used to make predictions Asian and European countries' COVID-19 survival . The COVID-19 data set were divided into four categories: continent, country, cases, and deaths. The data set includes variables and numeric values in various formats.



Row No.	dateRep	day	month	year	cases	deaths	countriesAn...	geoId	countryterrit	popData2019	continentExp	Cumulative ...
1	14 Dec, 202(	14	12	2020	746	6	Afghanistan	AF	AFG	38041757	Asia	9.914
2	13 Dec, 202(	13	12	2020	298	9	Afghanistan	AF	AFG	38041757	Asia	7.953
3	12 Dec, 202(	12	12	2020	113	11	Afghanistan	AF	AFG	38041757	Asia	6.869
4	11 Dec, 202(	11	12	2020	63	10	Afghanistan	AF	AFG	38041757	Asia	7.134
5	10 Dec, 202(	10	12	2020	202	16	Afghanistan	AF	AFG	38041757	Asia	6.969
6	9 Dec, 2020	9	12	2020	135	13	Afghanistan	AF	AFG	38041757	Asia	6.963
7	8 Dec, 2020	8	12	2020	200	5	Afghanistan	AF	AFG	38041757	Asia	7.095
8	7 Dec, 2020	7	12	2020	210	26	Afghanistan	AF	AFG	38041757	Asia	7.216
9	6 Dec, 2020	6	12	2020	234	10	Afghanistan	AF	AFG	38041757	Asia	7.326
10	5 Dec, 2020	5	12	2020	235	18	Afghanistan	AF	AFG	38041757	Asia	7.116
11	4 Dec, 2020	4	12	2020	119	5	Afghanistan	AF	AFG	38041757	Asia	7.108
12	3 Dec, 2020	3	12	2020	202	19	Afghanistan	AF	AFG	38041757	Asia	7.536
13	2 Dec, 2020	2	12	2020	400	48	Afghanistan	AF	AFG	38041757	Asia	7.605
14	1 Dec, 2020	1	12	2020	272	11	Afghanistan	AF	AFG	38041757	Asia	6.961
15	30 Nov, 202(	30	11	2020	0	0	Afghanistan	AF	AFG	38041757	Asia	6.417
16	29 Nov, 202(	29	11	2020	228	11	Afghanistan	AF	AFG	38041757	Asia	6.845
17	28 Nov, 202(	28	11	2020	214	15	Afghanistan	AF	AFG	38041757	Asia	6.785
18	27 Nov, 202(	27	11	2020	0	0	Afghanistan	AF	AFG	38041757	Asia	6.386
19	26 Nov, 202(	26	11	2020	200	12	Afghanistan	AF	AFG	38041757	Asia	7.342
20	25 Nov, 202(	25	11	2020	185	13	Afghanistan	AF	AFG	38041757	Asia	7.269
21	24 Nov, 202(	24	11	2020	246	17	Afghanistan	AF	AFG	38041757	Asia	6.714
22	23 Nov, 202(	23	11	2020	252	8	Afghanistan	AF	AFG	38041757	Asia	6.656
23	22 Nov, 202(	22	11	2020	154	12	Afghanistan	AF	AFG	38041757	Asia	6.204
24	21 Nov, 202(	21	11	2020	232	25	Afghanistan	AF	AFG	38041757	Asia	6.130
25	20 Nov, 202(	20	11	2020	282	5	Afghanistan	AF	AFG	38041757	Asia	5.673

Fig 1 Covid -19 data set Sample

B. RESEARCH APPROACH

This methodology entails a combination of numerous methods. It begins with a pre-processing stage to train different Big data approaches.

C. RANDOM FOREST APPROACH

The Random Forest Algorithm is another prominent Big Data Analytics approach. Random forest is a big data model that is commonly used for classification and regression. variety of samples, using the average for regression and the majority vote for classification. "Ensemble" simply refers to the combination of several nodes. Cross validation ensures a higher level of accuracy. It increases the accuracy of a decision tree by reducing over fitting concerns and decreasing variance. The Constructed Classification Structure can also be modified. The dataset was separated into various subsets in this algorithm, which were then expanded with nodes.

The final leaf node value is then taken into account for prediction. The average of all trees, based on the total key features in an RF.

$$RF\ fi_i = \frac{\sum_{j\text{eall tree}} norm\ fi_{ij}}{T}$$

The RFi_i represents the feature's importance, the norm fi sub (ij) represents the normalized importance T is the total number of trees, and I in tree j.

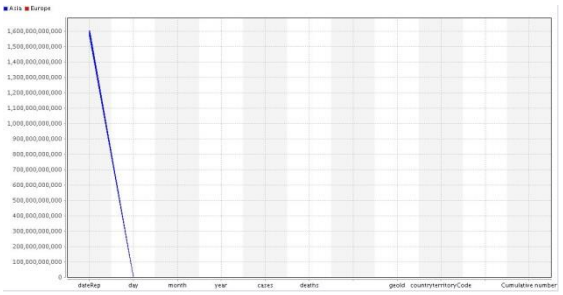


Fig 2 Asia and Europe country view using Random forest algorithm.

D. DECISION TREE METHOD

Among the most well-known Data science algorithms is the decision tree classifier. The Regression and Classification issues in variables are solved using this approach. From the root node, the decision tree compares the values with the new record in the decision branch. They partition the data set using the values in each feature to a point where all data points with the same class are grouped together.

The entropy is a metric that evaluates the randomness of knowledge across the continent, and it is described by:

$$E(S) = \sum_{i=1}^{c} - P_i \log_2 P_i$$

The current situation of the countries is represented by S, while the Probability of survival is represented by P_i.

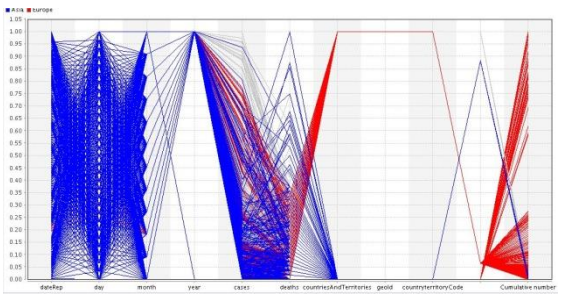


Fig 3 Asia and Europe country views using decision tree.

E. RAPID MINER TOOL

YAE is a rapid miner that was previously known as YAE (Yet Another Learning Environment). Rapid miner is a data science comprehension tool. This means that for data mining jobs, the user does not need to code. The blank procedure in rapid miner has a graphical user interface called the rapid miner. The Rapid miner tool has a local repository that can store our data sets as well as some example data sets. It also has a database connection and certain operators that contain process algorithms. As a result, the rapid miner is referred to as a user-friendly tool.

IV. RESULTS AND DISCUSSION

The Using the COVID-19 data set, this study constructed a Big data model. The collected data set was used in the manual data crawling method. This paper uses this method to investigate the current corona virus outbreak. while illustrating the utility of training data sets for comparing the efficacy On the data set, the accuracies of the two models. The correctness of the random forest was reported (95.92 %). Decision tree model accuracy has the highest accuracy (96.02 %).

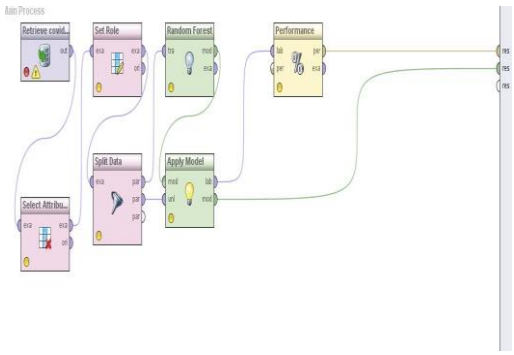


Fig 4. Data Crawling in Random Forest

Fig 4 Using random forest algorithm . Some operators, such as the first operator for data crawling, use a covid-19 data set. The second operator is select attribute, which is used to retrieve the required attribute, such as country, continent, and so on. The set role operator is then used to set attribute values. Split data is the fourth operator, which is used to split data into different formats. The Random Forest operator was then used to classify the data. The algorithm model is represented by the apply model. The performance vector is calculated by performance (classification).

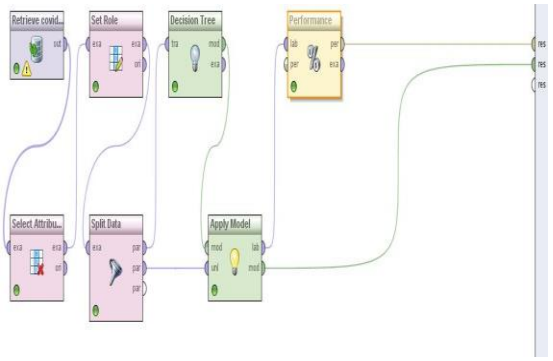


Fig 5. Data Crawling Process in Decision Tree

Fig 5 Using Decision Tree algorithm. Some operators, such as the first operator for data crawling, use a covid-19 data set. The select attribute operator is used to retrieve the desired attribute, such as country, continent, and so on. The attribute values are then set using the set role operator. The fourth operator, split data, is used to divide data into distinct representations. For decision making, the Decision Tree operator was employed. The apply model represents the algorithm model. Performance is used to calculate the performance vector (classification).

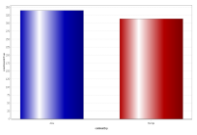


Fig 6. Classification Level of Asian and European Countries

Fig 6 Using the Random Forest and Decision Tree Algorithm, the classification level of Asian and European countries in the COVID-19 data set was determined. Asian countries were hit worse than European countries in this.

```
PerformanceVector:
accuracy: 95.92%
ConfusionMatrix:
True:   Asia    Europe
Asia:   102     8
Europe: 0       86
classification_error: 4.08%
ConfusionMatrix:
True:   Asia    Europe
Asia:   102     8
Europe: 0       86
weighted_mean_recall: 95.74%, weights: 1, 1
ConfusionMatrix:
True:   Asia    Europe
Asia:   102     8
Europe: 0       86
weighted_mean_precision: 96.36%, weights: 1, 1
ConfusionMatrix:
True:   Asia    Europe
Asia:   102     8
Europe: 0       86
absolute_error: 0.000 +/- 0.000
relative_error: 0.00% +/- 0.00%
```

Fig 7 . Performance Vector of Random Forest.

```
PerformanceVector:
accuracy: 96.02%
ConfusionMatrix:
True:   Asia    Europe
Asia:   109     8
Europe: 0       84
classification_error: 3.98%
ConfusionMatrix:
True:   Asia    Europe
Asia:   109     8
Europe: 0       84
weighted_mean_recall: 95.65%, weights: 1, 1
ConfusionMatrix:
True:   Asia    Europe
Asia:   109     8
Europe: 0       84
weighted_mean_precision: 96.58%, weights: 1, 1
ConfusionMatrix:
True:   Asia    Europe
Asia:   109     8
Europe: 0       84
absolute_error: 0.055 +/- 0.048
relative_error: 5.52% +/- 4.84%
```

Fig 8 . Performance Vector of Decision Tree.

Fig 7 & 8 shows Accuracy, Classification Error, Confusion Matrix, Recall, Precision, Relative Error, Absolute Error value for Decision Tree and Random Forest Algorithm. The ability of an instrument to measure the accurate value is known as accuracy. Accuracy is obtained by taking small readings. The closeness of two or more measurements to each other is known as the precision . Recall is the fraction of relevant instances that were retrieved. A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score.

The relative error is ,ratio of the absolute error of the measurement to the actual measurement. Absolute error is, difference between measured or inferred value and the actual value of a quantity.

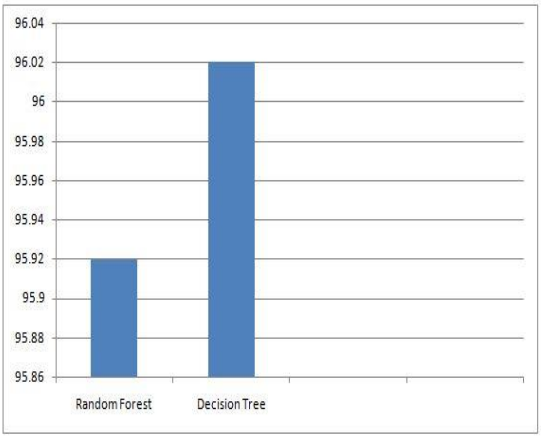


Fig 9 . Comparison of Algorithm Performance

Fig 9 The X-axsis shows Algorithm type and Y-axsis shows Accuracy of the Algorithms.

The Decision Tree and Random Forest can both be observed. In that, the decision tree method was expected to have a higher accuracy than the random forest algorithm in the end of the comparison.

VI.CONCLUSION

Random Forest and Decision Tree Classification Algorithms were utilized on the COVID-19 data set for Asian and European countries during the first and second weeks of January 2022 in this study. During that time period, the study discovered that Asian countries were more hit by Corona virus sickness than European countries. When compared to Random Forest, The Decision Tree has the highest level of accuracy in this paper, with a 96 percent accuracy. As a result, the Decision Tree gave precise results in the Prediction Examination.

ACKNOWLEDGEMENT

Authors are grateful to Principal, Faculty members and Staffs in Our College.

REFERENCES

[1] Abdu Gumaei, Walaa N. Ismail, Md.Rafiul Hassan, Mohammad Mehedi Hassan, Ebtsam Mohamed, Abdullah Alelaiwi, Giancarlo Fortino, A Decision Level Fusion Method for COVID-19 Patient Health Prediction , (oct 2021).

[2] M.Giacalone, D.C.Sinito, M.V.Calciano, V.Santarcangelo, Using Big Data to Record and Represent Compliance in the COVID-19 Era , (oct 2021).

[3] Ijegwa David Acheme, Olufunke Rebecca Vincent, Machine learning models for predicting survivability in COVID-19 Patients, (may 2021).

[4] T.K.Tsang, P. Wu, Y.Lin, E.H, Lau, G.M. Leung, B.J. Cowling, Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland china: a modeling study, Lancet Public Health(2020).

[5] Rahayu GR, Utomo PS, Riskiyana R, Hidayah RN Opportunity Amid Crisis in Medical Education: Teaching During the Pandemic of COVID-19(2022).

[6] Y.Song, S.Zheng, L.Li, X.Zhang,Z.Huang, J.Chen, H.Zhao,Y.Jie, R.Wang, Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images, medRxiv,2020.

[7] S.Lalmuanawma, J.Hussain, L.Chhakchhuak, COVID-19 (SARS-Cov-2) Review of the Pandemic, chaotic solitons fractals(2020) 110059, machine learning and artificial intelligence applications.

[8] V.K.R. Chimmula, L. Zhang, Time series forecasting of COVID-19 transmmision in Canada using LSTM networks, chaos solitons Fractals (2020) 109864.

[9] A.M.U.D. Khanday, S.T.Rabani, Q.R. Khan, N.Rouf, M.M.U.Din, COVID-19 detection using clinical text data using machine learning techniques, Int. J. Inf. Technol. (2020).

[10] H.Y.Cheng, Y.C.Wu, M.H.Lm, Y.L.Liu, Y.Y.Tsai, J.H. Wu, K.H.Pan, C.J.ke, C.MN. Chen,D.P.Liu Applying machine learning models with a ensemble approach for real-time influenza forecasting in Taiwan: development and validation study, J.Med. Internet Res. 22 (2020) e15394.

[11] H.M.Ngie, L.Ndreu, D.G. Mwigereri, Tree based regressor ensemble for viral infectious diseases spread prediction, in: 3rd African Conference on software engineering, Vol. 2689, ACSE,2020, p. 2020.

[12] N.E.Dean, A.P.Y Piontti, Z.J.Madewell, D.A.Cummings, M.D.Hitchings, K.Joshi, R.Khan, A.Vespignani, M.E.Halloran, I.M.Longini Jr, Ensemble forecast modeling for the design of COVID-19 vaccine efficacy trials, vaccine (2020).

[13] E.L.Ray, N.Wattanachit, J.Niemi, A.H.Kanji, K.House, E,Y.Cramer, J.Bracher,A.Zheng, T.K.Yamana, X.Xiong, Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the US, medRxiv, 2020.

[14] Habbash F, Ben Salah A, Almarabheh A, Jahrami H Insomnia and Related Factors During the Delta Wave of the COVID-19 Pandemic in the Kingdom of Bahrain: A Cross-Sectional Study (2022).

[15] Shaheen N, Mohamed A, Attalla A, Diab RA, Swed S, Nashwan AJ, Rababah AA, Hefnawy MT, Soliman Y, Abdelwahab OA, Desouki MT, Khaity A, Shaheen A, Ramadan A, Meshref M

[16] Could the New BA.2.75 Sub-Variant Cause the Emergence of a Global Epidemic of COVID-19? A Scoping Review (2022).

[17] Richter SM, Barnard TG Knowledge, Attitudes, and Perceptions Towards Hand Hygiene of Optometry Students Pre- and Peri-COVID-19 at a Tertiary Institution in Johannesburg, South Africa (2022).

[18] M.M.Sajadi , P.Habibzadeh ,A.Vintzileos ,S.Shokouhi, F.Miralles-Wilhelm , A.Amoroso.Mrach%,2020. Temperature and Latitude Analysis to predicts potential Spread and Seasonality for COVID-19 Available at SSRN 3550308.

[19] R.L. Benza , D.P. Miller , M.Gomberg -Maitlan , R.P.Frantz , A.J.Foreman , C.S.Coffey. Predicting survival in pulmonary arterial hypertension insights from the registry to evaluate early and long-term doi:10.1161/ circulationaha.109.898122. http://circ.ahajournals.org.

[20] Nguyen HT, Nguyen CC, Le Hoang T Falls Among Older Adults During the COVID-19 Pandemic: A Multicenter Cross-Sectional Study in Vietnam (2022).

[21] AlMatham K, AlWadie A, Kasule O, AlFadil S, Al-Shaya O Assessment of Postgraduate Online Medical Education During the COVID-19 Pandemic in Saudi Arabia: A Cross-Sectional Study(2022).

[22] Zehra SS, Fatima W Race Against the Clock: On the Transmission Dynamics of COVID-19 in Africa (2022).

[23] M. Lipsitch and N.E.Dean, Understanding the efficacy of the COVID-19 vaccination , Science (80-), vol.370, no.. 6518, pp. 763-765, Nov.2020,doi: 10.1126\science.abe 5938.

[24] B.Hermawan, Jokowi Bentuk Timnas , Percepatan Pengembangan Vaksin COVID-19 , republika.co.id, 2020. https://republika.co.id/berita/qgcqby354/jokowi-bentuk-timnas-percepatan-pengembangan-vaskin-covid-19 (accessed Jan.22,2021).

[25] T.Mustaqim, K.Umam, and M.A.Muslim, With Vader lexicon polarity detection and the K-nearest neighbor method, Twitter text mining for sentiment analysis on governments' responses to forest fires was possible., J.Phys. Conf. Ser., Vol. 1567, P. 032024, Jun. 2020, doi: 10.1088/1742-6596/1567/3/032024.

[26] R.Watrianthos, M.BobbiKurniawan, Kusmanto, S.Budiman, and B.Ulya, Mapping of Traffic Accidents in Labuhanbatu Regency using GIS Supports, J. Phys. Conf. ser., vol.1566, p. 012104, Jun. 2020, doi: 10.1088/1742-6596/1566/1/012104.

[27] S.Panjaitan et al., Implementation of Apriori Algorithm for Algorithm for Analysis of Consumer Purchase Patterns, J. Phys. Conf. Ser., vol 1255, p. 012057, Aug. 2019, doi: 10.1088/1742-6596/1255/1/012057.

[28] Pristiyono, MulkanRitonga, Muhammad Ali Al Ihsan, AgusAnjar, FauziahHanumRambe, Sentiment analysis of COVID-19 Vaccine in Indonesia using Naïve Bayes Algorithm, (2020).