LACK OF EFFICIENCY IN ENVISAGING THE USER TRAITS OVER ONLINE SOCIAL MEDIA BASED ON INDIAN METRICS DURING PANDEMIC USING NOVEL DECISION TREE ALGORITHM COMPARING WITH GRADIENT BOOSTED DECISION TREE ALGORITHM

V. Sai Ram Kumar[1], Shri Vindhya A[2*]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

Aim: The primary aim of this research is to increase the intensity percentage of user traits detection to reveal the impact of coronavirus on Twitter users by utilizing machine learning classifier algorithms by comparing Novel Decision Tree algorithm algorithm and Gradient Boosted Decision Tree algorithm.
Materials and Methods: Decision Tree Classifier algorithm with test size=10 and Gradient Boosted Decision Tree algorithm with test size=10 was estimated several times to envision the efficiency percentage with confidence interval of 95% and G-power (value=0.8). Decision Tree classifier constructs regression and classification model in the shape of tree. Gradient Boosted Decision Tree is an advanced version of decision tree that improves the performance by weak learners working sequentially .
Results and Discussion: Decision Tree algorithm has greater efficiency (87%) when compared to Gradient Boosted Decision Tree efficiency (60%). The results achieved with significance value p=0.448 (p>0.05) shows that two groups are statistically insignificant.
Conclusion: Decision Tree algorithm executes remarkably greater than the Gradient Boosted Decision Tree algorithm.

Keywords: Novel Decision Tree, Gradient Boosted Decision Tree, Twitter, Covid, Pandemic, User Traits Detection.

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai,Tamil Nadu,India. Pincode: 602105.
[2*]Project Guide, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu,India. Pincode:602105.

1. Introduction

The aim of this research is to predict User Traits Detection during a pandemic by sentiment analysis on social media(Twitter) using a novel Decision Tree algorithm and to compare the proposed algorithm with Gradient Boosted Decision Tree algorithm. In this case, the investigation aims to enhance the rate of efficiency in detecting user traits. Coronavirus also known as COVID-19 is a contagious illness that took shape as lung syndrome in December 2019. Dry cough, fever, and exhaustion are the most prevalent symptoms of coronavirus (Sarkodie and Owusu 2020). Governments have implemented a variety of precautionary steps to combat the virus's spread, including social isolation, business closures, and educational institution closures etc. As a result, most people communicate via social media platforms. Twitter and other social media sources are real-time communication tools that enable social media users to communicate and interact with many people at the same time (Chan et al. 2020). This platform helps people to express their feelings on pandemic (Rosenberg, Syed, and Rezaie 2020). An answer to detect user nature on pandemic using Decision tree algorithm is aimed in this research. With the huge amount of tweets on COVID-19, different research has been proposed. This research helps for User Traits Detection on pandemic using sentiment analysis. The proposed approach is used to identify user nature on pandemic and to calculate the efficiency of each algorithm used. A related research concludes that the novel Decision Tree algorithm has better detection efficiency and faster detection time (Kumar and Priyanka 2020). The application of this approach helps in detecting user traits and visualizing them with sentiment analysis (Ribeiro et al. 2020; Park, Park, and Chong 2020; Rajput, Grover, and Rathi 2020; Dai and Charnigo 2018; T. Wang et al. 2020; Pastor 2020; Dubey 2020).

Sentiment analysis is used by many researchers for various applications like analyzing customer opinions and brand monitoring etc. On sentiment analysis, 79 papers were published in IEEE Explore and 263 publications in Google Scholar. Yum (Yum 2020) researched in the USA (United States of America) to identify the important aspects of coronavirus by using Twitter data streams. In addition to that, Jain and Sinha (Jain and Sinha 2020) have recognized the influential users on social media using the tweets dataset. Furthermore, during the COVID-19 epidemic, Schild et al. (Tahmasbi et al. 2021) and his colleagues investigated the exposure of emotions on Twitter. Chen and Wang, (Chen, Wang, and Wu 2021) have presented an automated online aggressive conduct interpretation system to interpret tweets about COVID-19. The next part of research focused on using data mining techniques to analyze social network data like tweets. These include social network analysis, e.g., (Ribeiro et al. 2020; Park, Park, and Chong 2020; Rajput, Grover, and Rathi 2020; Dai and Charnigo 2018; T. Wang et al. 2020; Pastor 2020; Dubey 2020)), and disinformation detection, modeling, and forecasting, e.g., (Al-Rakhami and Al-Amri 2020; Kouzy et al. 2020; Pourghomi, Dordevic, and Safieddine 2018; Safieddine, Dordevic, and Pourghomi 2017). Further researches were to identify COVID-19 patients using deep and machine learning algorithms that overcomes detection failures and improved efficiency, e.g., (Ozturk et al. 2020; Oh, Park, and Ye 2020; Nour, Cömert, and Polat 2020; Minaee et al. 2020; Sethy et al. 2020).

Our institution is passionate about high quality evidence based research and has excelled in various domains (Vickram et al. 2022; Bharathiraja et al. 2022; Kale et al. 2022; Sumathy et al. 2022; Thanigaivel et al. 2022; Ram et al. 2022; Jothi et al. 2022; Anupong et al. 2022; Yaashikaa, Keerthana Devi, and Senthil Kumar 2022; Palanisamy et al. 2022). The drawback of the existing User Traits Detection system is less efficient in predicting user nature especially when there is a huge set of data. The main aim of our proposed system is to improve efficiency in predicting user nature by sentiment analysis using a novel Decision Tree Classifier algorithm.

2. Materials And Methods

This research work was carried out at Cyber Forensic Laboratory, Saveetha School of Engineering, SIMATS (Saveetha Institute of Medical and Technical Sciences). The proposed work contains two groups. Group 1 is taken as Decision Tree Classifier and group 2 as Gradient Boosted Decision Tree Classifier. The Decision Tree Classifier algorithm and Gradient Boosted Decision Tree algorithm were executed and evaluated a different number of times with a sample size of 40 (J. Wang 2020; Al-Shargabi and Selmi 2021) with a confidence interval of 95%, and with pretest power of 80% and maximum accepted error is fixed as 0.05.

After data collection, the invalid values, and independent content in the datasets were separated by pre-processing and data cleaning steps. After data cleaning and preprocessing the data, a perfect input for the User traits detection model is created, which are refined into the detection model by python libraries, and efficiency of both novel

Decision Tree Classifier algorithm and Gradient Boosted Decision Tree algorithm is calculated. The studying process of Decision Tree Classifier algorithm and Gradient Boosted Decision Tree algorithms are given below.

Decision Tree Classifier Algorithm

The Decision Tree algorithm is one of the supervised learning algorithms. A Decision Tree is used to create a training model that may be used to identify the target variable's class or data value using basic decision rules derived from previous data, such as training data (Sameer and Sriramya 2021). Figure 1 shows the algorithm for the Decision Tree Classifier algorithm from dataset processing to output and efficiency generation.

Gradient Boosted Decision Tree Classifier Algorithm

Gradient Boosted Decision Tree algorithm is an advanced version of Decision Tree which is similar to AdaBoost. It uses a group of Decision trees significantly to predict the target variable (Priyadarshini, Bazila Banu, and Nagamani 2019). It usually surpasses the Random Forest algorithm. Figure 2 shows the algorithm for the Gradient Boosted Decision Tree algorithm from dataset processing to output and efficiency generation.

Procedure for User Traits detection using Random Forest Algorithm / K-Nearest Neighbor Algorithm:

Step-1: Data gathering

First and foremost, we require the data that will be analyzed later. Scraping tools, APIs, customers' data feeds, and other methods can be used to collect data from social media, specifically Twitter. We can also collect information from consumer reviews on sites such as Google and Yelp. We'll be on the lookout for any mentions of the firm or brand throughout a particular time period. This is a frequent technique in all types of social media listening. I gathered a twitter dataset about covid-19 using different IEEE xplore papers.

Step-2: Analyze the data

It's critical to preprocess data to achieve high-quality results. Data preparation is separated into four steps to make the process easier: data cleaning, data reduction and data transformation.

Outliers are removed, missing values are replaced, noisy data is smoothed, and inconsistent data is corrected using data cleaning techniques. Each of these activities is carried out using a variety of ways, each of which is tailored to the user's preferences or issue set.

The purpose of data reduction is to create a condensed version of the data set that is smaller while maintaining the integrity of the original data set. As a result, efficient yet comparable results are produced.

Transforming the data into a format suitable for data modeling is the final phase in data preparation. Data preprocessing was done by using python libraries like “punkt” and “wordnet”

Step-3: User Traits detection Model

Step- 3.1 : User Traits detection with Decision Tree Algorithm

The purpose of employing a Decision Tree is to build a learning model that can be used to anticipate the target variable's class or value by using fundamental decision rules from past data like training data (Sameer and Sriramya 2021). As a result, in this project, we use a decision tree algorithm to predict user nature on Covid-19 using a dataset of tweets. The decision tree method is implemented using the sklearn tree python module.

Step- 3.2 : User Traits detection with Gradient Boosted Decision Tree Algorithm

A gradient boosted decision tree is a more sophisticated decision tree. Gradient boosting works by repeatedly developing smaller (weak) prediction models, with each model attempting to forecast the error left over from the preceding model (Priyadarshini, Bazila Banu, and Nagamani 2019). As a result, the algorithm has a proclivity towards overfitting. A model that outperforms random guesses by a small margin. It's used to forecast the behavior of Twitter users. The sklearn ensemble python package is used to implement the algorithm.

Step- 4 : Data Visualization

The interpretation of data into a graphical representation of information and data is known as data visualization. By utilizing visual components like charts, graphs, and maps, data visualization tools make it simple to evaluate and grasp trends, outliers, and patterns in data. I used python "numpy" and "panda" libraries to visualize personage traits.

The detection model follows the above procedure. Table 1 gives the source of the dataset and its properties. The dataset is processed using the visualization and NLP (Natural Language Processing) libraries and the data set is processed into a data frame. It is represented in Fig. 3, dependent variables are selected from the data frame and visualized using the matplotlib library. A train and a test set of data are created and implemented using two algorithms to predict user nature. Python programming language was used to implement this work.

Hardware specifications are concerned with the system resource settings allocated for specific devices. The following are minimum hardware

requirements to execute this model processor: intel i3, RAM 4GB, 250 GB HDD storage.
Software specifications are concerned with the resources that must be installed in the target system to get an application to work. The minimal software requirements for this model to work are windows operating system version 7/8/10, python programming language version 3 or above, Jupyter Notebook, or Google Collab.

Statistical Analysis

IBM SPSS v26 is used for statistical analysis (George and Mallery 2019). The independent variable is tweet_id, date, keyword, user_id and the dependent variable is user_nature and text. The independent T-test analysis is performed.

3. Results

Table 1 represents the details and source of the dataset. Table 2 shows the simulated efficiency analysis of novel Decision Tree Classifier and Gradient Boosted Decision Tree algorithms. Table 3 represents group statistical analysis with the mean value of 86.70 and 59.80, the standard deviation of 4.398 and 5.731 for novel Decision Tree Classifier and Gradient Boosted Decision Tree algorithms respectively. Table 4 represents the independent T-test analysis of both the groups with significance value p= 0.448 (p>0.05) states that both groups are statistically insignificant.
Figure 4 shows the bar graph analysis based on the efficiencies of two algorithms. The mean efficiencies of novel Decision Tree Classifier Algorithm and Gradient Boosted Decision Tree algorithm are 87% and 60% respectively. From the results obtained it is inferred that the novel Decision Tree User Traits Detection algorithm is more efficient than the Gradient Boosted Decision Tree algorithm.

4. Discussion

In this research work, Decision Tree Classifier and Gradient Boosted Decision Tree were executed for predicting the efficiency of User Traits Detection of Twitter users. When the two models were validated using the same dataset, it was discovered that the Decision Tree algorithm outperformed the Gradient Boosted Decision Tree algorithm approach. The novel Decision Tree classifier model for predicting the user nature of Twitter users was developed, which makes use of NLB (Natural Language Processing) and visualization libraries to process the user nature by text (user tweet). The proposed model predicts the user's nature using Decision Tree and displays it by using a bar graph. The datasets from different publications assisted in improving the efficiency percentage.
The research affects less development of efficiency in predicting user sentiment on Twitter (Babu et al. 2017). A similar work presents on sentiment classification (Zunic, Corcoran, and Spasic 2020) using the novel Decision Tree algorithm. The results achieved after all iterations on each dataset showed a constant 87% efficiency. The model proposed resulted in reaching more than a 27% rise in efficiency compared to the existing model (Brownlee 2016). Similar research carried out is about negative sentiment detection which is useful for future researchers who are interested in sentiment analysis (Jalil et al. 2021; Cheeti 2021). There are no such opposite findings regarding existing user traits detection for predicting user nature.
Although our proposed system is faster than Gradient Boosted Decision Tree in predicting user nature, it is generally extracted only limited features from the tweets and another limitation of this research work is, it is limited to test only twitter data (Huang et al. 2022). Currently, it is not programmed to embed with other social media. Further, this research work can be improved by deploying a model that analyzes more features from tweets in less time so that wait will be less and it can be embedded with other social media as in this research.

5. Conclusion

In this research work, prediction of efficiency percentage for User Traits Detection using Decision Tree algorithm appears to have enhanced efficiency (87%) when compared to Gradient Boosted Decision Tree algorithm (60%). User Traits Detection has been successfully employed for predicting twitter's user behavior. The results reveal the maximum number of true positives compared to true negatives from all the observations.

DECLARATION

Conflict of Interest

The author declares no conflict of interest.

Authors Contribution

Author VSRK was involved in data collection, data analysis, and manuscript writing. Author SVA was involved in conceptualization, data validation, and critical review of the manuscript.

Acknowledgment

The Authors would like to convey their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (previously known as Saveetha University) for providing the required infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete the research.
1. Cyclotron Technologies, Chennai, Tamilnadu.

2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences (SIMATS).
4. Saveetha School of Engineering.

6. References

Al-Rakhami, Mabrook S., and Atif M. Al-Amri. 2020. "Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter." *IEEE Access : Practical Innovations, Open Solutions* 8 (August): 155961–70.

Al-Shargabi, Amal A., and Afef Selmi. 2021. "Social Network Analysis and Visualization of Arabic Tweets During the COVID-19 Pandemic." *IEEE Access*. https://doi.org/10.1109/access.2021.3091537.

Anupong, Wongchai, Lin Yi-Chia, Mukta Jagdish, Ravi Kumar, P. D. Selvam, R. Saravanakumar, and Dharmesh Dhabliya. 2022. "Hybrid Distributed Energy Sources Providing Climate Security to the Agriculture Environment and Enhancing the Yield." *Sustainable Energy Technologies and Assessments*. https://doi.org/10.1016/j.seta.2022.102142.

Babu, Aarabhi, Mtech Student, Department of Computer Science and Engineering, Sahrdaya College Of Engineering and Technology Kodakara, Kerala, India., Vince Paul, et al. 2017. "Comparitive Study on Sentiment Analysis Techniques and User Behavior Prediction on Twitter Data." *International Journal Of Engineering And Computer Science*. https://doi.org/10.18535/ijecs/v6i2.01.

Bharathiraja, B., J. Jayamuthunagai, R. Sreejith, J. Iyyappan, and R. Praveenkumar. 2022. "Techno Economic Analysis of Malic Acid Production Using Crude Glycerol Derived from Waste Cooking Oil." *Bioresource Technology* 351 (May): 126956.

Brownlee, Jason. 2016. *XGBoost With Python: Gradient Boosted Trees with XGBoost and Scikit-Learn*. Machine Learning Mastery.

Chan, Teresa M., Kristina Dzara, Sara Paradise Dimeo, Anuja Bhalerao, and Lauren A. Maggio. 2020. "Social Media in Knowledge Translation and Education for Physicians and Trainees: A Scoping Review." *Perspectives on Medical Education* 9 (1): 20–30.

Cheeti, Swetha Sree. 2021. *Twitter Based Sentiment Analysis of Impact of COVID-19 on Education Globally*.

Chen, Toly, Yu-Cheng Wang, and Hsin-Chieh Wu. 2021. "Analyzing the Impact of Vaccine Availability on Alternative Supplier Selection Amid the COVID-19 Pandemic: A cFGM-FTOPSIS-FWI Approach." *Healthcare*. https://doi.org/10.3390/healthcare9010071.

Dai, Hongying, and Richard Charnigo. 2018. "A SENTIMENT ANALYSIS OF MERS-CoV OUTBREAK THROUGH TWITTER SOCIAL MEDIA MONITORING." *JP Journal of Biostatistics* 15 (2): 107–25.

Dubey, Akash Dutt. 2020. "Twitter Sentiment Analysis during COVID19 Outbreak." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3572023.

George, Darren, and Paul Mallery. 2019. *IBM SPSS Statistics 26 Step by Step: A Simple Guide and Reference*. Routledge.

Huang, Zhilian, Evonne Tay, Dillon Wee, Huiling Guo, Hannah Yeefen Lim, and Angela Chow. 2022. "Public Perception on the Use of Digital Contact Tracing Tools Post COVID-19 Lockdown: Sentiment Analysis and Opinion Mining." *JMIR Formative Research*, January. https://doi.org/10.2196/33314.

Jain, Somya, and Adwitiya Sinha. 2020. "Identification of Influential Users on Twitter: A Novel Weighted Correlated Influence Measure for Covid-19." *Chaos, Solitons & Fractals*. https://doi.org/10.1016/j.chaos.2020.110037.

Jalil, Zunera, Ahmed Abbasi, Abdul Rehman Javed, Muhammad Badruddin Khan, Mozaherul Hoque Abul Hasanat, Khalid Mahmood Malik, and Abdul Khader Jilani Saudagar. 2021. "COVID-19 Related Sentiment Analysis Using State-of-the-Art Machine Learning and Deep Learning Techniques." *Frontiers in Public Health* 9: 812735.

Jothi, K. Jeeva, K. Jeeva Jothi, S. Balachandran, K. Mohanraj, N. Prakash, A. Subhasri, P. Santhana Gopala Krishnan, and K. Palanivelu. 2022. "Fabrications of Hybrid Polyurethane-Pd Doped ZrO2 Smart Carriers for Self-Healing High Corrosion Protective Coatings." *Environmental Research*. https://doi.org/10.1016/j.envres.2022.113095.

Kale, Vaibhav Namdev, J. Rajesh, T. Maiyalagan, Chang Woo Lee, and R. M. Gnanamuthu. 2022. "Fabrication of Ni–Mg–Ag Alloy Electrodeposited Material on the Aluminium Surface Using Anodizing Technique and Their Enhanced Corrosion Resistance for Engineering Application." *Materials Chemistry and Physics*. https://doi.org/10.1016/j.matchemphys.2022.125900.

Kouzy, Ramez, Joseph Abi Jaoude, Afif Kraitem, Molly B. El Alam, Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Elie W. Akl, and Khalil Baddour. 2020. "Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter." *Cureus* 12 (3): e7255.

Kumar, Dharmender, and N. A. Priyanka. 2020. "Decision Tree Classifier: A Detailed Survey." *International Journal of Information and Decision Sciences*. https://doi.org/10.1504/ijids.2020.10029122.

Minaee, Shervin, Rahele Kafieh, Milan Sonka, Shakib Yazdani, and Ghazaleh Jamalipour Soufi. 2020. "Deep-COVID: Predicting COVID-19 from Chest X-Ray Images Using Deep Transfer Learning." *Medical Image Analysis* 65 (October): 101794.

Nour, Majid, Zafer Cömert, and Kemal Polat. 2020. "A Novel Medical Diagnosis Model for COVID-19 Infection Detection Based on Deep Features and Bayesian Optimization." *Applied Soft Computing* 97 (December): 106580.

Oh, Yujin, Sangjoon Park, and Jong Chul Ye. 2020. "Deep Learning COVID-19 Features on CXR Using Limited Training Data Sets." *IEEE Transactions on Medical Imaging* 39 (8): 2688–2700.

Ozturk, Tulin, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, and U. Rajendra Acharya. 2020. "Automated Detection of COVID-19 Cases Using Deep Neural Networks with X-Ray Images." *Computers in Biology and Medicine* 121 (June): 103792.

Palanisamy, Rajkumar, Diwakar Karuppiah, Subadevi Rengapillai, Mozaffar Abdollahifar, Gnanamuthu Ramasamy, Fu-Ming Wang, Wei-Ren Liu, Kumar Ponnuchamy, Joongpyo Shim, and Sivakumar Marimuthu. 2022. "A Reign of Bio-Mass Derived Carbon with the Synergy of Energy Storage and Biomedical Applications." *Journal of Energy Storage*. https://doi.org/10.1016/j.est.2022.104422.

Park, Han Woo, Sejung Park, and Miyoung Chong. 2020. "Conversations and Medical News Frames on Twitter: Infodemiological Study on COVID-19 in South Korea." *Journal of Medical Internet Research* 22 (5): e18897.

Pastor, Cherish Kay. 2020. "Sentiment Analysis of Filipinos and Effects of Extreme Community Quarantine due to Coronavirus (COVID-19) Pandemic." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3574385.

Pourghomi, Pardis, Milan Dordevic, and Fadi Safieddine. 2018. "The Spreading of Misinformation Online: 3D Simulation." In *2018 5th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*. IEEE. https://doi.org/10.1109/icitacee.2018.8576937.

Priyadarshini, R. K., A. Bazila Banu, and T. Nagamani. 2019. "Gradient Boosted Decision Tree Based Classification for Recognizing Human Behavior." *2019 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. https://doi.org/10.1109/icacce46606.2019.9080014.

Rajput, Nikhil Kumar, Bhavya Ahuja Grover, and Vipin Kumar Rathi. 2020. "Word Frequency and Sentiment Analysis of Twitter Messages during Coronavirus Pandemic." http://arxiv.org/abs/2004.03925.

Ram, G. Dinesh, G. Dinesh Ram, S. Praveen Kumar, T. Yuvaraj, Thanikanti Sudhakar Babu, and Karthik Balasubramanian. 2022. "Simulation and Investigation of MEMS Bilayer Solar Energy Harvester for Smart Wireless Sensor Applications." *Sustainable Energy Technologies and Assessments*. https://doi.org/10.1016/j.seta.2022.102102.

Ribeiro, Haroldo V., Andre S. Sunahara, Jack Sutton, Matjaž Perc, and Quentin S. Hanley. 2020. "City Size and the Spreading of COVID-19 in Brazil." *PloS One* 15 (9): e0239699.

Rosenberg, Hans, Shahbaz Syed, and Salim Rezaie. 2020. "The Twitter Pandemic: The Critical Role of Twitter in the Dissemination of Medical Information and Misinformation during the COVID-19 Pandemic." *CJEM* 22 (4): 418–21.

Safieddine, Fadi, Milan Dordevic, and Pardis Pourghomi. 2017. "Spread of Misinformation Online: Simulation Impact of Social Media Newsgroups." In *2017 Computing Conference*. IEEE. https://doi.org/10.1109/sai.2017.8252201.

Sameer, S. K. L., and P. Sriramya. 2021. "Improving the Efficiency by Novel Feature Extraction Technique Using Decision Tree Algorithm Comparing with SVM Classifier Algorithm for Predicting Heart Disease." *Alinteri Journal of Agriculture Sciences*. https://doi.org/10.47059/alinteri/v36i1/ajas21100.

Sarkodie, Samuel Asumadu, and Phebe Asantewaa Owusu. 2020. "Investigating the Cases of Novel Coronavirus Disease (COVID-19) in China Using Dynamic Statistical Techniques." *Heliyon* 6 (4): e03747.

Sethy, Prabira Kumar, Santi Kumari Behera, Pradyumna Kumar Ratha, and Preesat Biswas. 2020. "Detection of Coronavirus Disease (COVID-19) Based on Deep Features and Support Vector Machine." *International Journal of Mathematical, Engineering and Management Sciences* 5 (4): 643–51.

Sumathy, B., Anand Kumar, D. Sungeetha, Arshad Hashmi, Ankur Saxena, Piyush Kumar Shukla, and Stephen Jeswinde Nuagah. 2022.

"Machine Learning Technique to Detect and Classify Mental Illness on Social Media Using Lexicon-Based Recommender System." *Computational Intelligence and Neuroscience* 2022 (February): 5906797.

Tahmasbi, Fatemeh, Leonard Schild, Chen Ling, Jeremy Blackburn, Gianluca Stringhini, Yang Zhang, and Savvas Zannettou. 2021. "'Go Eat a Bat, Chang!': On the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19." *Proceedings of the Web Conference 2021*. https://doi.org/10.1145/3442381.3450024.

Thanigaivel, Sundaram, Sundaram Vickram, Nibedita Dey, Govindarajan Gulothungan, Ramasamy Subbaiya, Muthusamy Govarthanan, Natchimuthu Karmegam, and Woong Kim. 2022. "The Urge of Algal Biomass-Based Fuels for Environmental Sustainability against a Steady Tide of Biofuel Conflict Analysis: Is Third-Generation Algal Biorefinery a Boon?" *Fuel*. https://doi.org/10.1016/j.fuel.2022.123494.

Vickram, Sundaram, Karunakaran Rohini, Krishnan Anbarasu, Nibedita Dey, Palanivelu Jeyanthi, Sundaram Thanigaivel, Praveen Kumar Issac, and Jesu Arockiaraj. 2022. "Semenogelin, a Coagulum Macromolecule Monitoring Factor Involved in the First Step of Fertilization: A Prospective Review." *International Journal of Biological Macromolecules* 209 (Pt A): 951–62.

Wang, Jie. 2020. *Bayesian Logistic Regression with the Local Bouncy Particle Sampler for COVID-19*.

Wang, Tianyi, Ke Lu, Kam Pui Chow, and Qing Zhu. 2020. "COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model." *IEEE Access: Practical Innovations, Open Solutions* 8: 138162–69.

Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. "Algal Biofuels: Technological Perspective on Cultivation, Fuel Extraction and Engineering Genetic Pathway for Enhancing Productivity." *Fuel*. https://doi.org/10.1016/j.fuel.2022.123814.

Yum, Seungil. 2020. "Social Network Analysis for Coronavirus (COVID-19) in the United States." *Social Science Quarterly*, May. https://doi.org/10.1111/ssqu.12808.

Zunic, Anastazia, Padraig Corcoran, and Irena Spasic. 2020. "Sentiment Analysis in Health and Well-Being: Systematic Review." *JMIR Medical Informatics* 8 (1): e16023.

TABLES AND FIGURES

Table 1. Dataset Name, Extension and Source.

S.NO	DATASET NAME	DATASET EXTENSION	DATASET SOURCE
1	TWEETS DATASET	CSV	IEEE Xplore

Table 2. Efficiency of Decision Tree Algorithm and Gradient Boosted Decision Tree Algorithm.The Decision Tree is 27% more efficient than the Gradient Boosted Decision Tree algorithm.

ITERATION NO.	Decision Tree Algorithm DT(%)	Gradient Boosted Decision Tree Algorithm GBDT(%)
1	93	68
2	92	66
3	90	64

4	89	63
5	88	61
6	86	59
7	84	58
8	83	56
9	82	53
10	80	50

Table 3. Group Statistics of DT and GBDT algorithm with the mean value of 86.70% and 59.80%

GROUP	**N**	**Mean(%)**	**Std.Deviation**	**Std.Error Mean**
Decision Tree	10	86.70	4.398	1.391
GB Decision Tree	10	59.80	5.731	1.812

Table 4. Independent sample T-test is performed for the two groups for significance and standard error determination. The significance value p=0.448 (p>0.05) shows that two groups are statistically insignificant.

	Equal Variance	**Levene's Test for Equality of Variance**		**T-test for Equality of Means**					
		F	**Sig**	**t**	**df**	**Sig (2-tailed)**	**Mean Difference**	**Std. Error Differ-ence**	**95% Confidence Interval of the Difference**

									Lower	**Upper**
Efficiency	**Assumed**	.601	.448	11.775	18	.001	26.900	2.284	22.100	31.700
	Not Assumed			11.775	16.871	.001	26.900	2.284	22.077	31.723

Input:

Tweets dataset

Output:

Accuracy, confusion matrix.

Steps:

1. Remove unwanted variables from the dataset.
2. Load the data(train set, test set).
3. Import Decision Tree Classifier model from sklearn.tree
4. Fit train set and test set to Decision Tree Classifier model.
5. Do K-fold cross validation on our train set
6. Measure the mean accuracy of K-fold cross validation
7. Predict the user nature of test set
8. Measure the accuracy of the model on test set
9. Plot confusion matrix between predicted test set and original test set
10. Compile model

End

Fig. 1. Pseudocode for novel Decision Tree Classifier algorithm.

Input:

Tweets dataset

Output:

Accuracy, confusion matrix.

Steps:

1. Remove unwanted variables from the dataset.
2. Load the data(train set, test set).
3. Import Boosted Decision tree Classifier model from sklearn.ensemble
4. Fit train set and test set to Boosted Decision tree Classifier model.
5. Do K-fold cross validation on our train set
6. Measure the mean accuracy of K-fold cross validation
7. Predict the user nature of test set
8. Measure the accuracy of the model on test set
9. Plot confusion matrix between predicted test set and original test set
10. Compile model

End

Fig. 2. Pseudocode for Gradient Boosted Decision Tree algorithm.

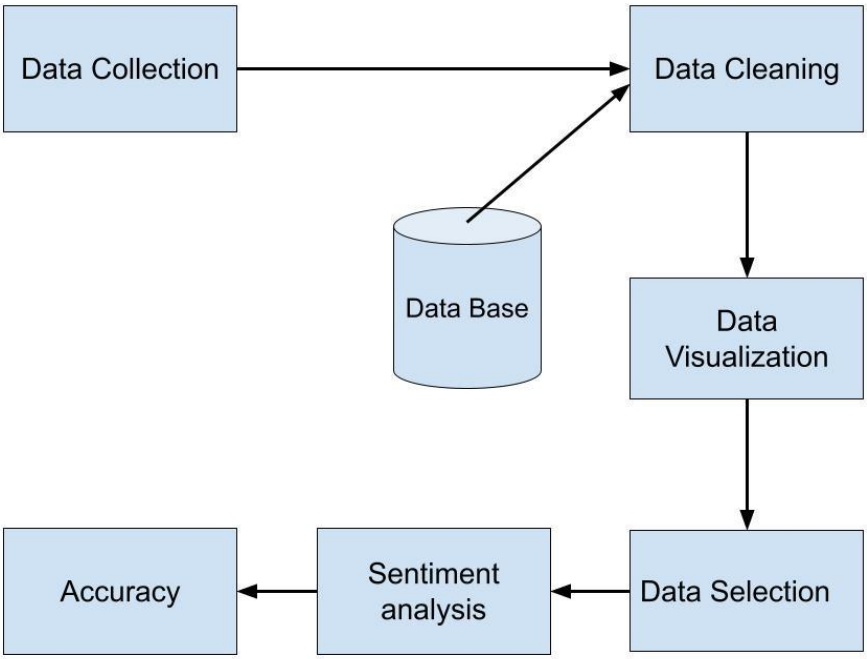

Fig. 3. Architecture for User Traits Detection for predicting user nature using novel Decision Tree algorithm, from dataset collection to accuracy calculation.

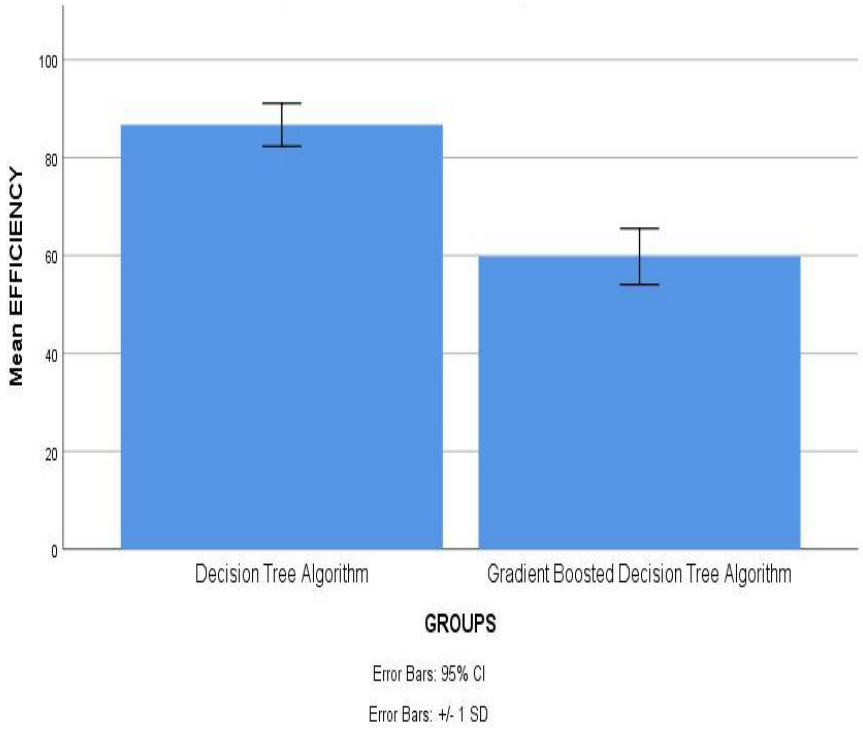

Fig. 4. Bar graph analysis of novel Decision Tree algorithm and Gradient Boosted Decision Tree algorithm. Graphical representation shows the mean efficiency of 87% and 60% for the proposed algorithm (Decision Tree) and gradient Boosted Decision Tree respectively. X-axis : Decision Tree Algorithm vs Gradient Boosted Decision Tree Algorithm, Y-axis : Mean precision ± 1 SD.