TO IMPROVE ACCURACY TO DETECT FAKE NEWS IN SOCIAL MEDIA USING DECISION TREES COMPARED OVER NAIVE BAYES ALGORITHM

V. Lakshmi Narayana[1], A. Gayathri[2*]

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

Abstract

Aim: The Machine Learning Algorithms to Detect Fake News in Social Media to discover the best accuracy in determining which news is fake and which is true. Decision Trees and Naive Bayes (NB) are two techniques for detecting anomalies.
Materials and Methods: The dataset for the false news identification was collected from kaggle. .The two groups are Novel Based Decision Trees (N=10) and Naive Bayes (N=10). As known, keeping of G-power and minimum power of the analysis is fixed as 80% and maximum accepted error is fixed as 0.5 with threshold value as 0.0805% and Confidence Interval is 95%.
Results: The Novel Decision Trees Detection Algorithm has been found to be useful in detecting fake news. The accuracy of the Decision Trees algorithm is (84.00%), whereas the accuracy of the Naive Bayes technique is (72.40%). These two algorithms are used to improve the detection of fake news. Furthermore, the independent significant value p=0.0496 ($p<0.05$) was met, i.e. alpha is 0.01 with a 95% confidence level.
Conclusion: The Novel Decision Trees Detection Algorithm looks to outperform Naive Bayes when it comes to recognising fake news on social media.

Keywords: Novel Decision Trees Detection Algorithm, Fake News, True News, Naive Bayes, Machine Learning, Social Media.

[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 False news in social media is growing and it may create a lot of difficulties these days, like these stories that are distributed arbitrarily and sarcastically (Juszczuk et al. 2021), hinting that what they are disseminating on social media isn't true (Antipova 2020). It is most common in Indian politics, when actual news is manipulated to create fake news (Choi et al. 2021). The news they are sharing, on the other hand, may have a different connotation and may spread false propaganda to the broader population (Jankowski 2021).

The False news may be detected by conducting several surveys and researchers and studies (Giachanou, Rosso, and Crestani 2021). There are 580 papers regarding false news detection in IEEE xplore, and 608 articles about fake news detection in ScienceDirect. The accuracy of detecting fake news in social media using Decision Trees was found to be (84.00%). Whereas the accuracy of the Naive Bayes method was found to be (72.40%) (D. K. Jain, Kumar, and Shrivastava 2022).

Our team has extensive knowledge and research experience that has translated into high quality publications (K. Mohan et al. 2022; Vivek et al. 2022; Sathish et al. 2022; Kotteeswaran et al. 2022; Yaashikaa, Keerthana Devi, and Senthil Kumar 2022; Yaashikaa, Senthil Kumar, and Karishma 2022; Saravanan et al. 2022; Jayabal et al. 2022; Krishnan et al. 2022; Jayakodi et al. 2022; H. Mohan et al. 2022). They can estimate an article by using these algorithms and the people also know which was fake and genuine, but by reading this article or research others may also understand about it (Xarhoulacos et al. 2021). There was no doubt in the public's mind about what was fake and what was real (Spitale et al. 2021). The main backstep is that they used a lot of qualities and attributes in existing algorithm.; as a result, and use afewer attributes in our news algorithms to provide accuracy.

2. Materials and Methods

The research was executed inside the Open Source laboratory at Saveetha School of Engineering(SSE), Saveetha Institute of Medical and Technical Sciences(SIMATS), Chennai. The requested work is being investigated. With G power set to 0.8, minimum power set to 0.8, maximum tolerable error set to 0.5, threshold set to 0.08 percent, and confidence interval set to 95 percent, the sample size was calculated using clincalc.com. Previous research was used to calculate the mean and standard deviation for size calculation. Novel Decision Trees Detection Algorithm (N=10), which is an existing model, and Naive Bayes (N=10), which is a proposed model, are the two groups employed ("Sample Size Calculator" n.d.). In this approach two sample groups are used. One is the Novel Detection Based Decision Tree new algorithm which is used to give more accuracy compared to existing algorithms of the sample groups of Novel Decision Trees Detection Algorithm (N=10). It gives more improvement. The second approach of the sample group is the existing algorithm of Naive Bayes (N=10). It gives accuracy in comparing the algorithm.At last, comparing a new method gives improvement. The MNIST dataset is used to discover all of the digits included in the dataset, as well as to train and test the Decision Trees. Over 1000 data points in the form of text news were acquired as a sample from kaggle with their respectives in the dataset. This data was collected and saved in a csv file that could be accessed. It can attain accuracy using the Decision Trees and Naive Bayes approaches.

Data Preparation

The Novel Decision Trees is to find all the digits that are stored in the dataset, to train and test through the dataset it comes from . The dataset includes 10000 data in the form of text which are taken as a sample from www.kaggle.com. There are 1000 trained texts and 9000 tested messages or data (Probierz et al. 2021).

Statistical analysis

The IBM SPSS version 21 statistical software is used for statistical analysis for our study. The independent variables are datasets and the dependent variables are shape and size and the accuracy. The independent T test analysis was carried out to calculate the accuracy for both methods (Li et al. 2021).

Decision Trees

Novel Decision Trees Detection Algorithms are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

Pseudocode
Step 1. Import the dataset correctly and provide the data path.
Step 2. Preprocess the data that has been imported.
Step 3. Tokenize the input and select the classification.
Step 4. Form the tree depending upon information.
Step 5. Using a machine learning algorithm to evaluate the data.
Step 6. Finally, use the Algorithm to check the effectiveness and accuracy.

Naive Bayes
Naive Bayes is a machine learning algorithm that falls under the category of supervised learning classifiers. Where each word in this document has its own unique format. The P(X|Ci) gives the posterior classification to produce plot from equation 1. The Naive Bayes algorithm is based on Bayes' theorem, which states that features in a dataset are independent when combined. The chance of occurrence of one feature has no bearing on the probability of occurrence of the other feature. Naive Bayes can outperform the most powerful alternatives for small sample sets. where the Naive Bayes algorithm, which was already in use, gives 82.40 percent.
$P(X|Ci)=P(xk|Ci)=P(x1|Ci)xP(x2|Ci)x...xP(xn|Ci)$ (1)

Pseudocode
Step 1. The first step is to import the data.
Step 2. Preprocess the data that has been imported.
Step 3. Tokenize the input and select the classification.
Step 4. Compute the frequency of terms and analyze the data.
Step 5. Using an assessment algorithm, evaluate the data.
Step 6. Finally, use the Algorithm to check the effectiveness and accuracy.

3. Results

The algorithm which is using sample size (N=10), Decision Tree delivers the observation by analyzing how it creates the, at whatever point it runs at various times. The layers are molded by the cycles, and the precision value varies with the length of running time, delivering the exactness and misfortune for the period shown in Table 1. Because of its enacting capacities and measures, Decision Tree out performs the Naive Bayes method in terms of precision and predictability. Table 1 shows the data collected from the dataset's N=10 samples for Decision Tree and Naive Bayes. It has been used in the Classification of Decision Trees, the datasets are created in SPSS with a sample size of N=10. The grouping variable is given as GroupID, and the testing variable is given as accuracy. For Decision Trees, the groupID is 1, the group ID is 2 for Naive Bayes.Table 2 shows the results of using Group Statistics on the Statistical Package for the Social Sciences (SPSS) dataset. Using Decision Trees and Naive Bayes to do statistical analysis, group statistics indicate a comparison of the accuracy in detecting fake news. The algorithm with the highest accuracy (84.00 %) was Decision Trees. In table2, Naive Bayes has the lowest accuracy with (72.40 %). In Table 3 it shows the Independent Sample T-Test that was used to collect the samples, with the level of significance set at 0.005 and a confidence range of 95%. Decision Trees have accepted a statistically significant value ($P<0.05$) after performing the SPSS computation. It was depicted by a simple bar Mean of Accuracy Decision Trees error range (0.99 - 0.98) and Loss error range (0.99 - 0.98) in Fig. 1.

4. Discussion

Our general results produce accuracy by comparing the machine learning algorithms that were used to examine the true and fake information.These algorithms produce accuracy by comparing it (Tay et al. 2021). The algorithm Decision Trees produces accuracy in this way (84.00%)(Shu and Liu 2019). By the comparison algorithm which may be NaiveBayes(72.40%). As a result, these two algorithms can have distinct specializations to demonstrate their accuracy(Dice 2017).

As shown in Fig. 1. Our proposed methodology achieved high headway rates for both allocated plots to some extent: When the successful robotized attack rate is 1%, manual human test plans are considered flawed, according to the algorithm(V. Jain et al. 2021). Using these two methods, information can be broken down into pieces, tokenized, and it can be determined which information is fraudulent and which is true by providing accuracy(Nagy and Kapusta 2021).
The fake news which was detected in social media, there may be a lot of news that characterizes news in multiple ways, such as fake and real, resulting in accuracy in detecting news (Szczepański et al. 2021). This is

useful in future for detecting or identifying the difference between authentic and fake news and they may be analyzed easily, as well as false propaganda and manipulated language of many kinds (Shirsat 2018).

5. Conclusion

In this analysis,the main thing was to detect fake news in social media by taking the dataset which was already present in the kaggle and by using machine learning methods like Novel Decision Trees Detection Algorithm which produces accuracy in detecting news is (84.00%) and Naive Bayes algorithm. Which it produces(72.40%). Among these two algorithms Decision Trees produces more accuracy than Naive Bayes which it is used as an existing algorithm.

Declarations

Conflict of interests

No conflicts of interest in this manuscript.

Author Contributions

Author VLN was involved in conceptualization, data collection, data analysis, manuscript writing. Author AG was involved in conceptualization, guidance, and critical review of the manuscript.

Acknowledgments

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

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. CK Technologies Pvt Ltd,Chennai,Tamil Nadu
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

6. References

Antipova, Tatiana. 2020. Integrated Science in Digital Age 2020. Springer Nature.

Choi, Jiho, Taewook Ko, Younhyuk Choi, Hyungho Byun, and Chong-Kwon Kim. 2021. “Dynamic Graph Convolutional Networks with Attention Mechanism for Rumor Detection on Social Media.” PloS One 16 (8): e0256039.

Dice, Mark. 2017. The True Story of Fake News: How Mainstream Media Manipulates Millions. Mark Dice.

Giachanou, Anastasia, Paolo Rosso, and Fabio Crestani. 2021. “The Impact of Emotional Signals on Credibility Assessment.” Journal of the Association for Information Science and Technology 72 (9): 1117–32.

Jain, Deepak Kumar, Akshi Kumar, and Akshat Shrivastava. 2022. “: A Hybrid Deep Neural Model with Mixed Fusion for Rumour Detection in Social Data Streams.” Neural Computing & Applications, January, 1–12.

Jain, Vidit, Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, and Yashvardhan Sharma. 2021. “AENeT: An Attention-Enabled Neural Architecture for Fake News Detection Using Contextual Features.” Neural Computing & Applications, August, 1–12.

Jankowski, Jarosław. 2021. “Habituation Effect in Social Networks as a Potential Factor Silently Crushing Influence Maximisation Efforts.” Scientific Reports 11 (1): 19055.

Jayabal, Ravikumar, Sekar Subramani, Damodharan Dillikannan, Yuvarajan Devarajan, Lakshmanan Thangavelu, Mukilarasan Nedunchezhiyan, Gopal Kaliyaperumal, and Melvin Victor De Poures. 2022. “Multi-Objective Optimization of Performance and Emission Characteristics of a CRDI Diesel Engine Fueled with Sapota Methyl Ester/diesel Blends.” Energy. https://doi.org/10.1016/j.energy.2022.123709.

Jayakodi, Santhoshkumar, Rajeshkumar Shanmugam, Bader O. Almutairi, Mikhlid H. Almutairi, Shahid Mahboob, M. R. Kavipriya, Ramesh Gandusekar, Marcello Nicoletti, and Marimuthu Govindarajan. 2022. “Azadirachta Indica-Wrapped Copper Oxide Nanoparticles as a Novel Functional Material in Cardiomyocyte Cells: An Ecotoxicity Assessment on the Embryonic Development of Danio Rerio.” Environmental Research 212 (Pt A): 113153.

Juszczuk, Przemysław, Jan Kozak, Grzegorz Dziczkowski, Szymon Głowania, Tomasz Jach, and Barbara Probierz. 2021. “Real-World Data Difficulty Estimation with the Use of Entropy.” Entropy 23 (12). https://doi.org/10.3390/e23121621.

Kotteeswaran, C., Indrajit Patra, Regonda Nagaraju, D. Sungeetha, Bapayya Naidu Kommula, Yousef Methkal Abd Algani, S. Murugavalli, and B. Kiran Bala. 2022. "Autonomous Detection of Malevolent Nodes Using Secure Heterogeneous Cluster Protocol." Computers and Electrical Engineering. https://doi.org/10.1016/j.compeleceng.2022.107902.

Krishnan, Anbarasu, Duraisami Dhamodharan, Thanigaivel Sundaram, Vickram Sundaram, and Hun-Soo Byun. 2022. "Computational Discovery of Novel Human LMTK3 Inhibitors by High Throughput Virtual Screening Using NCI Database." Korean Journal of Chemical Engineering. https://doi.org/10.1007/s11814-022-1120-5.

Li, Chen, Hao Peng, Jianxin Li, Lichao Sun, Lingjuan Lyu, Lihong Wang, Philip S. Yu, and Lifang He. 2021. "Joint Stance and Rumor Detection in Hierarchical Heterogeneous Graph." IEEE Transactions on Neural Networks and Learning Systems PP (October). https://doi.org/10.1109/TNNLS.2021.3114027.

Mohan, Harshavardhan, Sethumathavan Vadivel, Se-Won Lee, Jeong-Muk Lim, Nanh Lovanh, Yool-Jin Park, Taeho Shin, Kamala-Kannan Seralathan, and Byung-Taek Oh. 2022. "Improved Visible-Light-Driven Photocatalytic Removal of Bisphenol A Using V2O5/WO3 Decorated over Zeolite: Degradation Mechanism and Toxicity." Environmental Research. https://doi.org/10.1016/j.envres.2022.113136.

Mohan, Kannan, Abirami Ramu Ganesan, P. N. Ezhilarasi, Kiran Kumar Kondamareddy, Durairaj Karthick Rajan, Palanivel Sathishkumar, Jayakumar Rajarajeswaran, and Lorenza Conterno. 2022. "Green and Eco-Friendly Approaches for the Extraction of Chitin and Chitosan: A Review." Carbohydrate Polymers 287 (July): 119349.

Nagy, Kitti, and Jozef Kapusta. 2021. "Improving Fake News Classification Using Dependency Grammar." PloS One 16 (9): e0256940.

Probierz, Barbara, Jan Kozak, Piotr Stefański, and Przemysław Juszczuk. 2021. "Adaptive Goal Function of Ant Colony Optimization in Fake News Detection." Computational Collective Intelligence. https://doi.org/10.1007/978-3-030-88081-1_29.

"Sample Size Calculator." n.d. Accessed March 23, 2021. https://www.calculator.net/sample-size-calculator.html.

Saravanan, A., P. Senthil Kumar, B. Ramesh, and S. Srinivasan. 2022. "Removal of Toxic Heavy Metals Using Genetically Engineered Microbes: Molecular Tools, Risk Assessment and Management Strategies." Chemosphere 298 (July): 134341.

Sathish, T., R. Saravanan, V. Vijayan, and S. Dinesh Kumar. 2022. "Investigations on Influences of MWCNT Composite Membranes in Oil Refineries Waste Water Treatment with Taguchi Route." Chemosphere 298 (July): 134265.

Shirsat, Abhijeet. 2018. UNDERSTANDING THE ALLURE AND DANGER OF FAKE NEWS IN SOCIAL MEDIA ENVIRONMENTS.

Shu, Kai, and Huan Liu. 2019. Detecting Fake News on Social Media. Morgan & Claypool Publishers.

Spitale, Giovanni, Sonja Merten, Kristen Jafflin, Bettina Schwind, Andrea Kaiser-Grolimund, and Nikola Biller-Andorno. 2021. "A Novel Risk and Crisis Communication Platform to Bridge the Gap Between Policy Makers and the Public in the Context of the COVID-19 Crisis (PubliCo): Protocol for a Mixed Methods Study." JMIR Research Protocols 10 (11): e33653.

Szczepański, Mateusz, Marek Pawlicki, Rafał Kozik, and Michał Choraś. 2021. "New Explainability Method for BERT-Based Model in Fake News Detection." Scientific Reports 11 (1): 23705.

Tay, Li Qian, Mark J. Hurlstone, Tim Kurz, and Ullrich K. H. Ecker. 2021. "A Comparison of Prebunking and Debunking Interventions for Implied versus Explicit Misinformation." British Journal of Psychology , December. https://doi.org/10.1111/bjop.12551.

Vivek, J., T. Maridurai, K. Anton Savio Lewise, R. Pandiyarajan, and K. Chandrasekaran. 2022. "Recast Layer Thickness and Residual Stress Analysis for EDD AA8011/h-BN/B4C Composites Using Cryogenically Treated SiC and CFRP Powder-Added Kerosene." Arabian Journal for Science and Engineering. https://doi.org/10.1007/s13369-022-06636-5.

Xarhoulacos, Constantinos-Giovanni, Argiro Anagnostopoulou, George Stergiopoulos, and Dimitris Gritzalis. 2021. "Misinformation vs. Situational Awareness: The Art of Deception and the Need for Cross-Domain Detection." Sensors 21 (16). https://doi.org/10.3390/s21165496.

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.

Yaashikaa, P. R., P. Senthil Kumar, and S. Karishma. 2022. "Review on Biopolymers and Composites – Evolving Material as Adsorbents in Removal of Environmental Pollutants." Environmental Research. https://doi.org/10.1016/j.envres.2022.113114.

Tables and Figures

Table 1. Accuracy Values for Decision Trees and NB are recorded and noted the values.

S.NO	DECISION TREES	NB
1	93.80	94.80
2	92.09	92.00
3	93.99	91.00
4	90.00	88.00
5	87.00	87.00
6	95.00	86.50
7	89.00	87.00
8	88.00	79.00
9	85.00	76.00
10	77.00	75.00

Table 2. Independent Sample T-Test is applied for the sample collections by fixing the level of significance as 0.0496 with confidence interval as 95 %. After applying the SPSS calculation, Random Forest(RF) has accepted a statistically significant value(P<0.05).

Group Statistics

	Algorithms	N	Mean	Std Deviation	Std Error Mean
Accuracy	DECISION TREES	10	84.0000	8.18535	3.66060
	NB	10	72.4000	9.15423	4.09390

Table 3. Independent Samples T-test-LR seems to be significantly better than NB. With producing the sig(2-tailed value) is 0.46 which is less than p-value (0.05).

Accuracy	Independent Samples Test								
	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig	t	df	Sig(2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
Equal variances assumed	.072	.0496	2.112	8	.046	11.60000	5.49181	-1.06404	24.26414
Equal variances not assumed	.066		2.112	7.902	.044	11.60000	5.49181	-1.06404	24.29155

Graph:

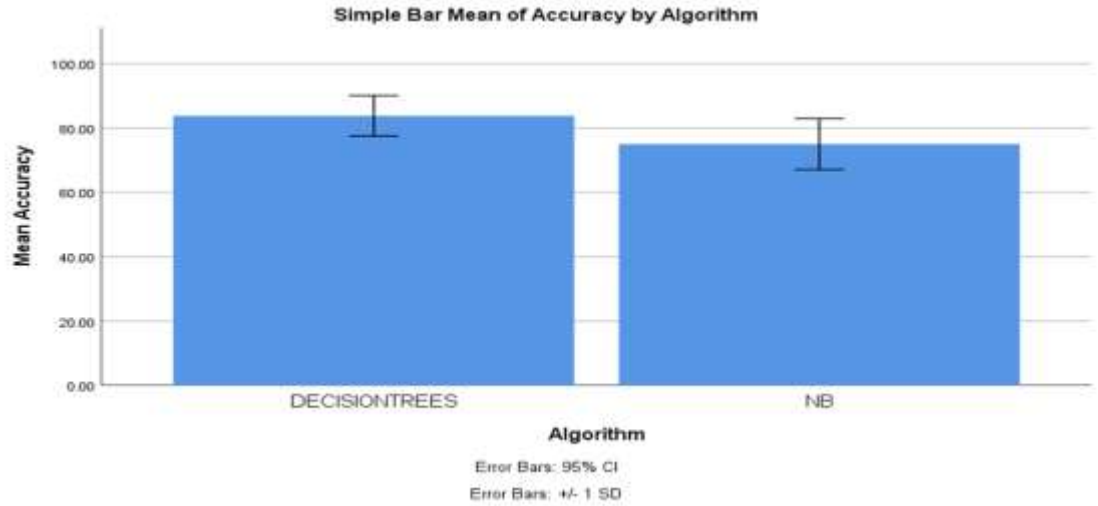

Fig. 1. Comparison of Novel Decision Trees Detection Algorithm and Naive Bayes algorithms in terms of accuracy Decision Trees (84.00%) is better than the pre existing algorithm (72.40%) accuracy. X-axis: DT vs NB and Y-axis is mean accuracy ± 1 SD.