TO IMPROVE ACCURACY TO DETECT FAKE NEWS IN SOCIAL MEDIA USING RANDOM FOREST 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: To predict accuracy by using Machine Learning Algorithms to find fake news published in Social Media to discover the best accuracy in determining which news is fake and which is true. Random Forest(RF) and Naive Bayes (NB) are two approaches for detecting anomalies.
Materials and Methods: The dataset for the false news identification was collected from www.kaggle.com. The two groups are Random Forest (N=10) and Naive Bayes (N=10). By Adding G power and to fix 80% is the minimum power of the analysis and maximum accepted error is fixed as 0.5 with threshold value as 0.0805% and Confidence Interval is 95%.
Results: A Novel Random Forest (RF) Detection Algorithm has been found to be useful in detecting fake news. The accuracy of the Random Forest(RF) algorithm is (82.60%), 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=0.0414 ($p<0.05$) was met, i.e. alpha is 0.01 with a 95% confidence level.
Conclusion: The Novel Random Forest (RF) Detection Algorithm looks to outperform Naive Bayes when it comes to recognising fake news on social media.

Keywords: Novel Random Forest (RF) 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

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

Fake news may be detected by several academics. There are 465 papers regarding false news detection in IEEE xplore, and 73 articles about fake news detection in ScienceDirect. The accuracy of detecting fake news in social media using Random Forest(RF) was found to be (82.60%) (Ireton and Posetti 2018). Whereas the accuracy of the Naive Bayes method was found to be (72.40%) (Chiluwa and Samoilenko 2019).

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)

As people live , they can estimate which words, sentences, or paragraphs are fake and which are genuine, but by reading this article (Zimdars and Mcleod 2020), others may be able to tell which are fake and which are genuine. There was no doubt in the public's mind about what was fake and what was real (Chakraborty, n.d.). The main backstep is that they used a lot of qualities and attributes in existing algorithm.; as a result, It uses fewer attributes in our news algorithms to provide accuracy (Maglogiannis, Iliadis, and Pimenidis 2020a).

2. Materials and Methods

The work for research was done in the Open Source lab in Saveetha School of Engineering(SSE), SIMATS, Chennai. The requested work is being investigated. With G power set to 0.8, the value 0.8 be set to minimum power, the value 0.5 maximum tolerable error set and threshold set to 0.0714 percent, and confidence interval set to 95 percent, clincalc.com used to find the sample size. Previous research was used to calculate the mean and standard deviation for size calculation (Coles 2018). Novel Decision Trees Detection Based 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 taken. One is a new algorithm which produces more difference than the existing algorithm which is using sample groups which is the Novel Random Forest (RF) Detection Algorithm (N=10) and it produces an accurate value. The second sample group is Naive Bayes(N=10) which is pre pre-existing approach which takes different values of different calculations and suggests the percentage of information and suggests which is suitable. The MNIST dataset is used to discover all of the digits included in the dataset, as well as to train and test the Novel Random Forest (RF) Detection Algorithm. Over 1000 data points in the form of text news were acquired as a sample from www.kaggle.com 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 Random Forest(RF) and Naive Bayes approaches.

Data Preparation

The Novel Random Forest 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 (Gupta et al. 2022).

Statistical analysis

For statistical analysis for our study, here IBM SPSS version 21 statistical software is used. The independent variables are datasets and the dependent variables are shape and size and the accuracy. The T test (independent) analysis was carried out, to calculate the accuracy for both methods (Malla and Alphonse 2022).

Random Forest

Novel Random Forest (RF) Detection Algorithm is a model for supervised learning that is an enhanced version of decision trees (DT). RF is made up of a huge

number of decision trees that work together to predict the expected output of a class, with the final prediction based on the class that obtained the most votes. Due to little correlation across trees, the error rate in random forest is low when compared to other models. Our random forest model was trained with various parameters, such as varying numbers of estimators, in a grid search to find the optimal model that can accurately predict the outcome from equation 1. There are several strategies for deciding a split in a decision tree based on a regression or classification problem.

$Gini=(pi)^2$, where i=1: (1)

Pseudocode

Step 1. Dataset Imported correctly and provides the data path.
Step 2. Preprocess the data that has been imported.
Step 3. Tokenize the input and select the classification.
Step 4. Determine the frequency of terms and create a document term matrix.
Step 5. Using an algorithm for machine learning to evaluate the data.
Step 6. Finally, use the Algorithm to check the effectiveness and accuracy.

Naive Bayes

Naive Bayes is one of the algorithms for machine learning that falls under the category of supervised learning classifiers. Where each word in this document has its own unique format. The posterior classification has been done and plot has been formed using equation 2. The Naive Bayes algorithm is a working principle 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.

$P(X|Ci)=P(xk|Ci)=P(x1|Ci)xP(x2|Ci)x...xP(xn|Ci)$ (2)

Pseudocode

Step 1. The first step is to import the data.
Step 2. Preprocess the imported data.
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

In training the algorithm test size (N=10), The Novel Random Forest (RF) Detection Algorithm 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, Random Forest out performs the Naive Bayes method based on precision and predictability. Table 1 represents the data collected from the dataset's N=10 samples for Random Forest and Naive Bayes. As used in the Classification of Random Forest, 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 Random Forest, the groupID is 1, the groupID 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 Random Forest 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 (82.60 %) was Support Vector Machine. In table 2, 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 0.0714 and a confidence range of 95%. Random Forest has accepted a statistically significant value ($P<0.05$) after performing the SPSS computation. It was depicted by a simple bar Mean of Accuracy Random Forest error range (0.99 - 0.98) and Loss error range (0.11 - 0.22) in Fig. 1.

4. Discussion

Lastly, 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. The algorithm Random Forest produces accuracy

in this way(82.60%) (Shirsat 2018). By the comparison algorithm which may be NaiveBayes(72.40%). As a result, these two algorithms can have distinct specializations to demonstrate their accuracy (Maglogiannis, Iliadis, and Pimenidis 2020b).

As shown in Fig. 1. Our proposed methodology achieved high headway rates for both allocated plots to some extent (Rice 2018): When the successful robotized attack rate is 1%, manual human test plans are considered flawed, according to (Raza and Ding 2022). 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 (Pasumpon Pandian et al. 2019). Despite the truth of detecting fake information on social media, there can be lots of information that characterizes information in more than one ways, such as fake and real, resulting in accuracy in detecting news (Palani, Elango, and Viswanathan K 2021). This is useful for detecting or identifying the difference between authentic and fake news, as well as false propaganda and manipulated language. In this process a detection of fake news provides accurate information which is true and fake by producing the resulting accuracy in detecting news and in future however it may create a great impact on fake news it may be useful in easy prediction (Taskin, Kucuksille, and Topal 2021). These algorithms provide what was genuine and fake in circulating social media. By using these detection processes and can suggest which was real and fake (O'Brien 2018).

5. Conclusion

In this process,the main thing was to find fake news published in the social media by taking the dataset which was already present in the kaggle and by using machine learning methods like Random Forest(RF) which it produces accuracy in detecting news is (82.60%) and Naive Bayes algorithm
Which it produces(72.40%). Among these two algorithms Random Forest(RF) produces more accuracy than existing Naive Bayes 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

Chakraborty, Tanmoy. n.d. Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers. Springer Nature.

Chiluwa, Innocent E., and Sergei A. Samoilenko. 2019. Handbook of Research on Deception, Fake News, and Misinformation Online. IGI Global.

Chun, Russell, and Susan J. Drucker. 2020. Fake News: Real Issues in Modern Communication.

Coles, T. J. 2018. Real Fake News: Techniques of Propaganda and Deception-Based Mind Control, from Ancient Babylon to Internet Algorithms. Red Pill Press.

Greifeneder, Rainer, Mariela Jaffe, Eryn Newman, and Norbert Schwarz. 2020. The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation. Routledge.

Gupta, Ashish, Han Li, Alireza Farnoush, and Wenting Jiang. 2022. "Understanding Patterns of COVID Infodemic: A Systematic and Pragmatic Approach to Curb Fake News." Journal of Business Research 140 (February): 670–83.

Ireton, Cherilyn, and Julie Posetti. 2018. Journalism, Fake News & Disinformation: Handbook for Journalism Education and

Training. UNESCO Publishing.

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.

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.

Maglogiannis, Ilias, Lazaros Iliadis, and Elias Pimenidis. 2020a. Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II. Springer Nature.

2020b. Artificial Intelligence Applications and Innovations. AIAI 2020 IFIP WG 12.5 International Workshops: MHDW 2020 and 5G-PINE 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings. Springer Nature.

Malla, Sreejagadeesh, and P. J. A. Alphonse. 2022. "Fake or Real News about COVID-19? Pretrained Transformer Model to Detect Potential Misleading News." The European Physical Journal. Special Topics, January, 1–10.

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.

O'Brien, Nicole J. 2018. Machine Learning for Detection of Fake News.

Palani, Balasubramanian, Sivasankar Elango, and Vignesh Viswanathan K. 2021. "CB-Fake: A Multimodal Deep Learning Framework for Automatic Fake News Detection Using Capsule Neural Network and BERT." Multimedia Tools and Applications, December, 1–34.

Pasumpon Pandian, A., Tomonobu Senjyu, Syed Mohammed Shamsul Islam, and Haoxiang Wang. 2019. Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI - 2018). Springer.

Raza, Shaina, and Chen Ding. 2022. "Fake News Detection Based on News Content and Social Contexts: A Transformer-Based Approach." International Journal of Data Science and Analytics, January, 1–28.

Rice, Dona Herweck. 2018. Deception: Real or Fake News? Teacher Created Materials.

"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.

Taskin, Suleyman Gokhan, Ecir Ugur Kucuksille, and Kamil Topal. 2021. "Detection of Turkish Fake News in Twitter with Machine Learning Algorithms." Arabian Journal for Science and

Engineering, October, 1–21.
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.
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.
Zimdars, Melissa, and Kembrew Mcleod. 2020. Fake News: Understanding Media and Misinformation in the Digital Age. MIT Press.

Tables and Figures

Table 1. Accuracy Values for RF and NB. And took 10 iterations of each algorithm to produce the best accuracy or algorithm.

S.NO	RANDOM FOREST	NB
1	95.80	94.80
2	94.09	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	75.00	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 0.0414 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	**RANDOM FOREST**	10	82.6000	7.95613	3.55809
	NB	10	72.4000	9.15423	4.09390

Table 3. Independent Samples T-test-LR seems to be significantly better than NB and have the significant value which accepts the p value is less than 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	.144	.0414	1.881	8	.048	10.20000	5.42402	-2.30781	22.70781
Equal variances not assumed	0.125		1.881	7.848	.034	10.20000	5.42402	-2.30781	22.75022

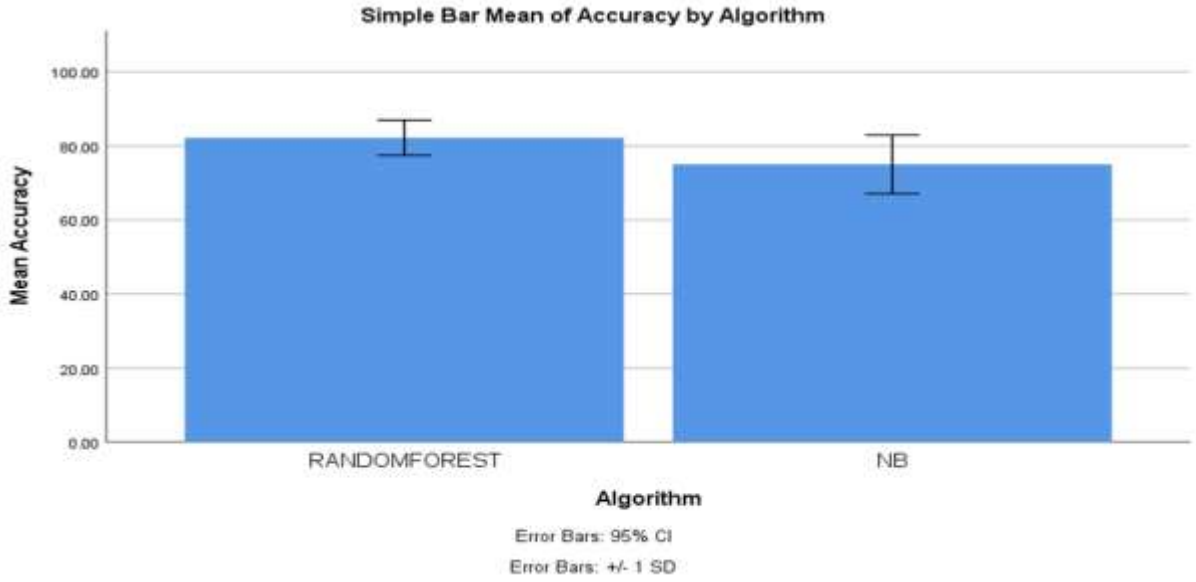

Fig. 1. Comparison of Random Forest and Naive Bayes algorithms in terms of accuracy Random Forest (82.60%) is better than the pre existing algorithm (72.40%) accuracy. X-axis: RF vs NB and Y-axis is mean accuracy ± 1 SD.