BALTIC JOURNAL OF LAW & POLITICS
A Journal of Vytautas Magnus University
VOLUME 15, NUMBER 4 (2022)
ISSN 2029-0454

Cite: *Baltic Journal of Law & Politics* 15:4 (2022): 140-146
DOI: 10.2478/bjlp-2022-004014

Classification of Spam Emails Using Random Forest Algorithm In Comparison With Naive Bayes Algorithm

Yenimireddy Thirumala Kishon Reddy
Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602105

S. Sobitha Ahila
Project Guide, Corresponding Author, Department of Computer Science and Engineering,
Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602105

Received: August 8, 2022; reviews: 2; accepted: November 29, 2022.

Abstract

Aim: The main aim of the research is to classify the spam emails using Random Forest over Naive Bayes Algorithm. **Materials and Methods:** Random Forest algorithm and Naive Bayes algorithm are implemented in this research work. Sample size of n = 20 calculated using G power software, G power value is between 0.59 and 0.9 and determined as 10 per group with pretest power 80%, threshold 0.05% and CI 95%. **Result:** Random Forest algorithm provides a higher accuracy of 98.33% compared to Naive Bayes algorithm with 88.22% to classify . There is a significant difference between two groups with a significance value of 0.049 ($p<0.05$). **Conclusion:** These results show that the performance of the Random Forest algorithm(98.33%) is better than that of the Naive Bayes algorithm(88.22%) in terms of accuracy.

Keywords

Machine Learning, Novel Tree Specific Method, Random Forest, Naive Bayes, Spam Filtering, Black List, White List.

INTRODUCTION

The aim of the work is about Spam emails classification using Random Forest over Naive Bayes Algorithm. The prediction using machine learning has succeeded in comparing Random Forest over Naive Bayes Algorithm (Douzi et al. 2020). Spam email classification as Novel Tree Specific Method is helpful to classify spam emails which save a lot of time to the users which are using the emails and avoids frauds and theft of personal information from various hackers. The emails can be identified by using the black lists and white lists . Recently unwanted commercial emails are marked as spam emails (Mishra and Thakur 2013). Email spam is one of the major problem that is faced by the every email user.On a dialy basis every email user recieved hundreds of spam emails from various anonymous address. White list and black list plays a major role in classification of non spam from spam emails. Spam email filtering was done by using black lists, white lists, ip addresses and email addresses (Wang, Zeng, and Huang 2020). Applications of spam filtering were used in the classification of spam emails in our phones, laptops and in our offices.

In the last five years, Google Scholar identified almost 1800 articles on Spam email classification using Machine Learning. Spam filtering of emails using Novel Tree Specific Method plays a major role in the classification of normal emails from spam emails. Spam filtering is used to classify the emails. The spam emails can be classified mainly on the basis of black lists and white lists(Santoso 2019). Spam emails are very annoying to the email users who have fallen victim to the spam emails ("Random Forest Algorithm for Spam Filtering Based on Machine Learning" 2015). Spam email classification by Novel Tree Specific Method is one of the most important tasks nowadays. It has been demonstrated that it is possible to use these machine learning algorithms to filter out spam emails from the normal emails (S et al. 2014).Normally White list and Black list contain certain words that are used in the spam and non-spam emails. Novel Tree Specific Method is used in the classification of spam emails. Spam filtering is used to classify emails. The Random forest algorithm belongs to the Novel Tree Specific Method. The purpose of these emails is to release confidential personal information to the hackers and also to steal the passwords,usernames and Bank Verification Number(BVN), by using this information they try to steal money from our bank accounts(Cormack 2008). From all these research papers, the best study paper in my opinion is(Wang, Zeng, and Huang 2020)

Previously our team has a rich experience in working on various research projects across multiple disciplines (Venu and Appavu 2021; Gudipaneni et al. 2020; Sivasamy, Venugopal, and Espinoza-González 2020; Sathish et al. 2020; Reddy et al. 2020; Sathish and Karthick 2020; Benin et al. 2020; Nalini, Selvaraj, and Kumar 2020).The research gap identified from the existing system shows poor accuracy. The study is to improve the accuracy of Classification by incorporating Random Forest and comparing performance with Naive Bayes (Broadhurst and Trivedi 2020). The proposed model improves accuracy (3.55%) to achieve more efficiency.

MATERIALS AND METHODS

This study setting was done in the Soft Computing Laboratory, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. Recall that the testing setup includes both hardware and software configuration choices. The laptop has an Intel Core i7 8th generation CPU with 12GB of RAM, an x86-based processor, a 64-bit operating system, and a hard drive. Currently, the software runs on Windows 10 and is programmed in Python. Once the program is finished, the accuracy value will appear. Procedure: Wi-Fi laptop connected. Chrome to Google Collaboratory search Write the code in Python. Run the code. To save the file, upload it to the disc, and create a folder for it. Log in using the ID from the message. Run the code to output the accuracy and graph. The number of required samples in research are two in which group 1 is Random Forest compared with group 2 of Naive Bayes Algorithm. The samples were taken from the device and iterated 10 times to get desired accuracy with G power 80%, threshold 0.05% and CI 95%. The study uses a dataset spam or not-spam dataset consisting of a collection of spam and non spam emails downloaded from kaggle (mukulkirti 2021).

Random Forest

Random Forest is a classic example of learning and regression techniques suitable for solving data classification problems(Akinyelu and Adewumi 2014) .Random forest algorithm is an branch of the Novel Tree Specific Method. Random forest algorithm is utilized to classify spam emails from normal emails. White list and Black list is commonly used to classify the emails into spam and non-spam emails. This Random forest algorithm classifies the data into the different classes using decision trees. Random forest algorithm has become widely popular over the years and it is being applied to various problems in various fields("To Enhance Phishing Emails Classification Using Machine Learning Algorithm" 2019) .

Pseudocode for Random Forest

```
    Step1: Import packages.
    Step2: Create an input dataset.
    Step3: Analyze the size of the taken input data.
    Step4: Split the datasets for testing and training the dataset.
    Step5: Apply Random Forest
    Step6: Predict the results.
```

Naive Bayes

Naive Bayes algorithms are a collection of classification algorithms based on Bayes Theorem. Naive Bayes is a machine learning algorithm widely used in a variety of classifications (Mishra and Thakur 2013) .

Pseudocode for Naive Bayes

```
    Step1: Import packages.
    Step2: Create an input dataset.
    Step3: Analyze the size of the taken input data.
    Step4: Split the datasets for testing and training the dataset.
    Step5: Apply Naive Bayes
    Step 6: Predict the results.
```

Statistical Analysis

SPSS is a software tool used for statistics analysis. The proposed system utilized 10 iterations for each group with predicted accuracy noted and analyzed. Independent samples t-test was done to obtain significance between two groups. Dependent variable is no.of white list words and independent variable is no.of black list words.

RESULTS

The proposed Novel Tree Specific Method of Random Forest and Naive Bayes were run at a time for performing spam email classification. Table 1 shows the accuracy value of iteration of Random Forest and Naive Bayes. Table 2 represents the Group statistics results which depicts Naive Bayes with mean accuracy of 88.37%, and standard deviation is 3.46. Random Forest has a mean accuracy of 97.90% and standard deviation is 1.28. Proposed Random Forest algorithm provides better performance compared to the Naive Bayes algorithm. Table 3 shows the independent samples T-test value for Naive Bayes and Random Forest with Mean difference as 9.53, std Error Difference as 1.17. Significance value is observed as 0.003 ($p<0.05$). Figure 1 shows the bar graph comparison of mean of accuracy on Naive Bayes and Random Forest algorithm. Mean accuracy of Naive Bayes algorithm is 88.37% and Random Forest algorithm is 97.90%.

DISCUSSION

In this study, classification of spam emails using Random Forest has significantly higher accuracy, approximately 98.33% in comparison to Naive Bayes (88.22%). Random Forest algorithm appears to produce more consistent results with minimal standard deviation.

In this study, classification of spam emails using Random Forest over the Naive Bayes has increased the accuracy (3.55%). The limitation of this research is that it cannot give appropriate results for smaller data like below the size of 5. The proposed work has more time complexity. The similar findings of the paper had an accuracy of 95% with Random Forest which was used to classify the spam emails from normal emails(S et al. 2014). The proposed work has reported that the Random Forest algorithm has 88% accuracy which is used to classify the spam emails. The work proposed by (Mishra and Thakur 2013) shows the Random Forest has a better accuracy of 90%. Random Forest algorithm which is used in both traditional and modern methods as per their research it opposes Random Forest has highest accuracy and aNaive Bayes algorithm will get least accuracy compared to other machine learning techniques which ranges between 60% when compared to other machine learning algorithms will get more accuracy than this (Rafat et al. 2022). By using Random Forest to classify the spam emails it will have key issues to pretend (Guia, Silva, and Bernardino 2019) in this paper shows Random Forest has the least accuracy of 88%. Increasing the dataset's value only tends to get desired accuracy. Random Forest performs better with a combination of other machine learning algorithms.

The limitation of this research is that it cannot give appropriate results for smaller data like below the size of 5. In this model it is not able to consider all given feature variable parameters for training. The future scope of proposed work will be classification of spam emails based on using class labels for lesser time complexity.

CONCLUSION

These results show that the performance of the Random Forest algorithm in classification of spam emails (98.33%) is better than that of the Naive Bayes algorithm (88.22%) in terms of accuracy and efficiency. Random Forest algorithm classifies spam emails better than that of Naive Bayes algorithm.

DECLARATION

Conflict of Interests

No conflict of interests in this manuscript.

Authors Contribution

Author YTKR was involved in data collection, data analysis, and manuscript writing. Author SSA was involved in conceptualization, data validation, and critical review of manuscript.

Acknowledgement

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. VK Technologies, Chennai.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

Akinyelu, Andronicus A., and Aderemi O. Adewumi. 2014. “Classification of Phishing Email Using Random Forest Machine Learning Technique.” *Journal of Applied Mathematics*. https://doi.org/10.1155/2014/425731.

Benin, S. R., S. Kannan, Renjin J. Bright, and A. Jacob Moses. 2020. “A Review on Mechanical Characterization of Polymer Matrix Composites & Its Effects Reinforced with Various Natural Fibres.” *Materials Today: Proceedings* 33 (January): 798–805.

Broadhurst, Roderic, and Harshit Trivedi. 2020. “Malware in Spam Email: Risks and Trends in the Australian Spam Intelligence Database.” https://doi.org/10.52922/ti04657.

Cormack, Gordon V. 2008. *Email Spam Filtering: A Systematic Review*. Now Publishers Inc.

Douzi, Samira, IPSS, Faculty of Science, University Mohammed Rabat, Morocco, Feda A. AlShahwan, Mouad Lemoudden, and Bouabid El Ouahidi. 2020. “Hybrid Email Spam Detection Model Using Artificial Intelligence.” *International Journal of Machine Learning and Computing*. https://doi.org/10.18178/ijmlc.2020.10.2.937.

Gudipaneni, Ravi Kumar, Mohammad Khursheed Alam, Santosh R. Patil, and Mohmed Isaqali Karobari. 2020. "Measurement of the Maximum Occlusal Bite Force and Its Relation to the Caries Spectrum of First Permanent Molars in Early Permanent Dentition." *The Journal of Clinical Pediatric Dentistry* 44 (6): 423–28.

Guia, Márcio, Rodrigo Silva, and Jorge Bernardino. 2019. "Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis." *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. https://doi.org/10.5220/0008364105250531.

Mishra, Rachana, and R. S. Thakur. 2013. "Analysis of Random Forest and Naïve Bayes for Spam Mail Using Feature Selection Catagorization." *International Journal of Computer Applications*. https://doi.org/10.5120/13844-1670.

mukulkirti. 2021. "Naive_bayes_Theorem." Kaggle. September 29, 2021. https://kaggle.com/mukulkirti/naive-bayes-theorem.

Nalini, Devarajan, Jayaraman Selvaraj, and Ganesan Senthil Kumar. 2020. "Herbal Nutraceuticals: Safe and Potent Therapeutics to Battle Tumor Hypoxia." *Journal of Cancer Research and Clinical Oncology* 146 (1): 1–18.

Rafat, Khan Farhan, Qin Xin, Abdul Rehman Javed, Zunera Jalil, and Rana Zeeshan Ahmad. 2022. "Evading Obscure Communication from Spam Emails." *Mathematical Biosciences and Engineering: MBE* 19 (2): 1926–43.

"Random Forest Algorithm for Spam Filtering Based on Machine Learning." 2015. *Electronic Engineering and Information Science*. https://doi.org/10.1201/b18471-53.

Reddy, Poornima, Jogikalmat Krithikadatta, Valarmathi Srinivasan, Sandhya Raghu, and Natanasabapathy Velumurugan. 2020. "Dental Caries Profile and Associated Risk Factors Among Adolescent School Children in an Urban South-Indian City." *Oral Health & Preventive Dentistry* 18 (1): 379–86.

Santoso, Budi. 2019. "An Analysis of Spam Email Detection Performance Assessment Using Machine Learning." *Jurnal Online Informatika*. https://doi.org/10.15575/join.v4i1.298.

Sathish, T., and S. Karthick. 2020. "Gravity Die Casting Based Analysis of Aluminum Alloy with AC4B Nano-Composite." *Materials Today: Proceedings* 33 (January): 2555–58.

Sathish, T., D. Bala Subramanian, R. Saravanan, and V. Dhinakaran. 2020. "Experimental Investigation of Temperature Variation on Flat Plate Collector by Using Silicon Carbide as a Nanofluid." In *PROCEEDINGS OF INTERNATIONAL CONFERENCE ON RECENT TRENDS IN MECHANICAL AND MATERIALS ENGINEERING: ICRTMME 2019*. AIP Publishing. https://doi.org/10.1063/5.0024965.

Sivasamy, Ramesh, Potu Venugopal, and Rodrigo Espinoza-González. 2020. "Structure, Electronic Structure, Optical and Magnetic Studies of Double Perovskite Gd2MnFeO6 Nanoparticles: First Principle and Experimental Studies." *Materials Today Communications* 25 (December): 101603.

S, Sarju, S. Sarju, Riju Thomas, and Shyni C. Emilin. 2014. "Spam Email Detection Using Structural Features." *International Journal of Computer Applications*. https://doi.org/10.5120/15485-4265.

"To Enhance Phishing Emails Classification Using Machine Learning Algorithm." 2019. *International Journal of Recent Technology and Engineering*. https://doi.org/10.35940/ijrte.c6542.118419.

Venu, Harish, and Prabhu Appavu. 2021. "Experimental Studies on the Influence of Zirconium Nanoparticle on Biodiesel–diesel Fuel Blend in CI Engine." *International Journal of Ambient Energy* 42 (14): 1588–94.

Wang, Lu, Guohui Zeng, and Bo Huang. 2020. "Naive Bayesian Algorithm for Spam Classification Based on Random Forest Method." *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/1486/3/032021.

TABLES AND FIGURES

Table 1. Accuracy Values for Random Forest and Naive Bayes

S.NO	Naive Bayes	Random Forest
1	88.22	98.33
2	89.24	98.50
3	90.37	98.00
4	92.89	96.54
5	93.22	95.69
6	93.45	99.25
7	87.45	99.00
8	86.32	97.45
9	84.39	99.57
10	83.23	96.67

Table 2. Group Statistics Results-Naive Bayes has an mean accuracy (88.37%), std.deviation (3.46), whereas for Random Forest has mean accuracy (97.90%), std.deviation (1.28).

Group Statistics					
	Groups	N	Mean	Std deviation	Std. Error Mean
Accuracy	RF	10	97.90	1.28	0.40
	NB	10	88.37	3.46	1.15

Table 3. Independent Samples T-test -Random Forest seems to be significantly better than Naive Bayes 0.049 ($p<0.05$)

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

ISSN 2029-0454
2022

								Lower	Upper
Equal variances assumed	6.75	0.049	8.122	17	0.003	9.53	1.17336	7.05443	12.00557
Equal variances not assumed			7.785	9.963	0.003	9.53	1.22419	6.80096	12.25904

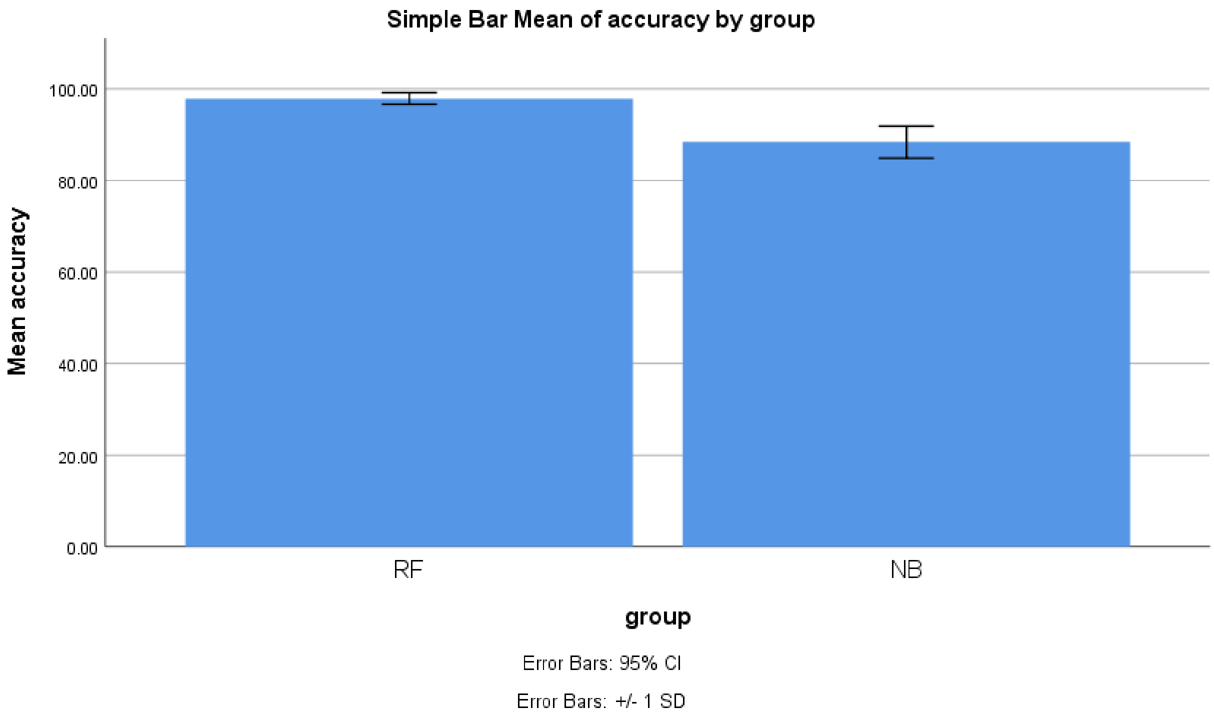

Figure. 1. Bar Graph Comparison on mean accuracy of Naive Bayes (88.22%) and Random Forest(98.33%). X-axis:Naive Bayes ,Random Forest algorithms, Y-axis: Mean Accuracy with ±1 SD.