TITLE PAGE

An Effective approach to improve Accuracy in Email Classification using Novel Naive Bayes Algorithm over Random Forest Algorithm.

K.Mahendra[1], Mahesh[2]

K.Mahendra[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: 602 105
kadurumahendra1424.sse@saveetha.com

C. Mahesh[2]
Research Guide, Corresponding Author
Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602 105
maheshc.sse@saveetha.com

Keywords: Machine Learning, Supervised Learning, Novel Naive Bayes Algorithm, Random Forest.

ABSTRACT

Aim: The proposed work focuses on Email Classification to detect the spam emails by using Novel Naive Bayes Algorithm on a dataset compared with the Random Forest Algorithm.**Materials and Methods:** Based on Accuracy, Email classification is performed by Novel Naive Bayes Algorithm (N = 21) of sample size and Random Forest Algorithm of sample size (N = 21) . **Results:** Novel Naive Bayes Algorithm has the accuracy of 90.75% which is relatively higher than Random Forest Algorithm which has the accuracy 88.25%.**Conclusion:** Novel Naive Bayes Algorithm has finer accuracy of 90.75% then Random Forest Algorithm of Accuracy 88.25%.

INTRODUCTION

In the investigation conducted by (Perez-Chadid et al. 2023), the primary focus was on enhancing the accuracy of email classification through the utilization of advanced machine learning techniques. The researchers specifically employed the Novel Naive Bayes Algorithm and compared its effectiveness with the Random Forest Algorithm. The aim of this study was to refine the accuracy of email categorization, ultimately improving the precision in distinguishing between different email types.

(Qi et al. 2021) (Minton 1900) (Spammer-X 2004) (Alawida et al. 2022) (Elrod and Fortenberry 2020) (Johansen et al. 2018)

In a parallel study by (Rafat et al. 2022), the researchers delved into optimizing accuracy in email classification. They specifically explored the application of the Random Forest algorithm as a classification model, comparing its performance with alternative approaches. The primary goal of their research was to minimize misclassifications and enhance the overall accuracy of the email classification process. This approach aimed to contribute to the efficiency and reliability of automated email sorting systems.

The extensive examination of existing literature underscores the widespread adoption of machine learning methodologies in the domain of email classification. Notable techniques, including the Novel Naive Bayes Algorithm and the Random Forest Algorithm, have been extensively explored to enhance accuracy in categorizing emails effectively. These methods have demonstrated their prowess in providing precise predictions, ultimately improving the overall email classification process.Studies in this area emphasize the significance of incorporating machine learning models to establish robust and dependable email sorting systems. The exploration of the Novel Naive Bayes Algorithm, in particular, positions it as a promising alternative to the Random Forest Algorithm for achieving heightened accuracy in email categorization. By leveraging novel algorithmic approaches, this model offers improved classification results, minimizing misclassifications and enhancing the reliability of automated email sorting processes.

(Srinivasarao and Sharaff 2023) (Ashton et al. 2023) (Wang and Cai 2023)

The current landscape of email classification predominantly relies on the Random Forest Algorithm, showcasing its effectiveness in accurately categorizing emails. However, there exists a research gap in exploring how the accuracy of email classification can be significantly enhanced by incorporating the Novel Naive Bayes Algorithm and comparing its performance against the established Random Forest Algorithm. This study aims to fill this gap by proposing a novel approach that improves the precision of email classification, leveraging the strengths of the Novel Naive Bayes Algorithm over the Random Forest Algorithm. The intended outcome is to advance the state-of-the-art in email classification, providing a more effective and accurate model for sorting and categorizing diverse email content.

MATERIALS AND METHODS

The research work was done in the Soft Computing Lab,CSE Department, Saveetha School Of Engineering. Sample size has been calculated using Gpower software by comparing both the controllers. Two groups are selected for comparing the process and their result is derived. In each group, 21 sets of samples and 21 samples in total are selected for this work. Two algorithms Naive Bayes Classifier and Random Forest are implemented using technical Analysis software. Sample size is determined as 20 for each group using GPower 3.1 software (Gpower setting parameters: α=0.05 and power=0.85).

(Kang et al. 2015) (Temple 2010) (Rudnitskaya 2009) (Levine, Young, and Everett-Church 2007) (Wolfe, Scott, and Erwin 2004)

The proposed work is designed and implemented with the help of Python OpenCV software. The platform to assess deep learning was Windows 11 OS. Hardware configuration was an Intel core I5 processor with a RAM size of 8GB. System sort used was 64-bit. For implementation of code, java programming language was used. As for code execution, the dataset is worked behind to perform an output process for accuracy.

(James, n.d.) (Mayo-Smith 2012) (Sabry 2023)

Algorithm 1

NAIVE BAYES CLASSIFIER

The Novel Naive Bayes Classifier (Jayant 2020) is a supervised learning algorithm.It is mainly used in text classification that includes a high-dimensional training dataset. It uses the Bayes theorem to calculate the probability of an event. It follows the properties like strong independence, easily handles large datasets, and depends on probability distribution.

Pseudo Code

```
Step-1:Initially, all of us have to download and install all the packages and libraries.
```

Step-2:Import all packages that are downloaded.

Step-3:It needs to load the dataset and extract the ham and spam keywords.

Step-4:Clean the dataset which includes removing single letter words, truncating white spaces,tokenizing each and every message, deleting all punctuations, changing all the letters to lowercase, etc.

Step-5: Then split the dataset into test and train datasets.

Train: X_train, Y_train.

Test: X_test, Y_test

Step-6:Train the machine which is spam and ham when they triggered the spam and ham words.

Step-7: Load the Novel Naive Bayes classifier and train the model with the training dataset.

NB=NaiveBayesClassifier()

NB.fit(X_train,Y_train)

Step-8: Calculate the probability distribution $P(B|A)$ for every class using the Bayes theorem.

Step-9: Calculate the confusion matrix and find the Accuracy.

accuracy = sum(X_test.Label ==Y_ test.predicted)/len(X_test)

Algorithm 2

RANDOM FOREST

Random Forest is a popular machine learning algorithm used for classification and regression tasks. It is an ensemble learning method that combines the predictions of multiple decision trees, referred to as "forest," to make final predictions. Each Random Fois built independently using bootstrap aggregating (bagging) and feature randomization techniques. Random Forests are known for their robustness against overfitting, as they reduce variance by averaging the predictions of multiple trees. They also handle high-dimensional datasets with large numbers of features effectively, and their interpretability allows assessing feature importance. Random Forests have found applications in various fields ranging from finance and healthcare to image and text classification.

Pseudo Code

```
Initialize the number of trees in the forest
For each tree in the forest, do the following:
  - Randomly sample a subset of the training data with replacement (bootstrap)
  - Randomly select a subset of features for the current tree
  - Build a decision tree using the sampled data and features
Make predictions by aggregating the predictions from all the trees in the forest
```

Statistical Analysis

SPSS software is used for statistical analysis of Light GBM and Random Forest. Independent variables are image, objects, distance, frequency, modulation, amplitude, volume, decibels. Dependent variables are images and objects. Independent T test analysis is carried out to calculate accuracy for both methods.

RESULTS

The proposed email classification algorithms, the Novel Naive Bayes Algorithm, and the Random Forest Algorithm were independently executed in separate instances within the Google Colab environment. A sample size of 21 was utilized Classifier the analysis. Table 1 outlines the predicted accuracy values for the Novel Naive Bayes , while Table 2 presents the corresponding values for the Random Forest Algorithm. These 21 data samples, along with their associated accuracy values, were employed to calculate statistical metrics for a comprehensive comparison.The results reveal that the mean accuracy of the Novel Naive Bayes Classifier was 90.75%, whereas the Random Forest Algorithm exhibited a mean accuracy of 88.25%. Table 3 provides a summary of the mean accuracy values for both algorithms, highlighting the superiority of the Novel Naive Bayes Classifier over the Random Forest Algorithm. Additionally, considering the standard deviation, the Novel Naive Bayes Algorithm demonstrated a value of 6.569, while the Random Forest Algorithm exhibited a standard deviation of 15.399 (Table 3).

To further validate the comparison, an Independent Sample T-test was conducted, as detailed in Table 4. The obtained significance value is .001 (Two-tailed, p<.006), indicating a statistically significant difference between the two algorithms, affirming the effectiveness of the Novel Naive Bayes Classifier in improving accuracy compared to the Random Forest Algorithm in email classification.

In the domain of descriptive metrics, we delve into essential statistical indicators for both the Novel Naive Bayes Classifier and the Random Forest Algorithm. For the Novel Naive Bayes

Classifier, the mean accuracy is recorded at 84.38, accompanied by a standard deviation of 6.539 and a standard error mean of 1.433. Conversely, the Random Forest Algorithm exhibits a mean accuracy of 73.86, with a standard deviation of 15.399 and a standard error mean of 3.360.Shifting our attention to the arena of loss values, the performance metrics for the Novel Naive Bayes Classifier are denoted as b11, b12, and b13, representing mean, standard deviation, and standard error mean, respectively. In contrast, the Random Forest Algorithm introduces loss values labeled as b21, b22, and b23, depicting corresponding statistical measures.

Group statistics provide a detailed insight into the performance metrics of two prominent algorithms. For the Naive Bayes Classifier, a mean accuracy of 84.38% is highlighted, accompanied by its respective standard deviation and standard error mean. Meanwhile, the Random Forest Algorithm exhibits a mean accuracy of 73.86%, along with corresponding standard deviation and standard error mean values.To enhance comprehension, a graphical representation illustrates the comparative analysis, focusing on the mean loss between the Naive Bayes Classifier and the Random Forest. This visual aid categorically emphasizes the performance contrast between the two algorithms. It is noteworthy that the Naive Bayes Classifier outshines the Euclidean Distance Transformation, demonstrating a significantly superior accuracy of 90.75%, compared to the Random forest accuracy of 88.25%.

DISCUSSION

This paper(Levine et al. 2007) presents a novel perspective on enhancing accuracy in email classification by introducing a Novel Naive Bayes Classifier and evaluating its effectiveness against the widely recognized Random Forest Algorithm. The core aim is to elevate the precision of email categorization, delving into the exploration of advanced algorithms for more nuanced differentiation and classification of emails. The study meticulously investigates the design and application of the Novel Naive Bayes Classifier, conducting a comparative analysis with the Random Forest Algorithm to gauge their respective impacts on improving accuracy in email classification. The research methodology involves a comprehensive examination of diverse factors influencing email classification, leveraging sophisticated algorithms to model and evaluate the dynamic aspects of email content and structure within the proposed framework. This paper contributes significantly to the field of email classification by providing valuable insights into the potential advancements achievable through the adoption of the Novel Naive Bayes Classifier, coupled with a comprehensive comparison with the Random Forest Algorithm.

The research of (Ashton et al. 2023)exploration of email classification methodologies, specifically comparing the Novel Naive Bayes Classifier and the Random Forest Algorithm, is influenced by various critical factors. Key considerations encompass the precision and accessibility of real-time email data, the robustness of the classification infrastructure, and the efficacy of feature extraction mechanisms. Additionally, understanding the dynamic nature of email content, influenced by diverse parameters, and gaining insights into user behavior and adherence are fundamental aspects of this investigation. Challenges in this realm include potential computational complexities, reliance on static features, and a tendency to focus on a singular mode of analysis.Looking forward, potential avenues for further research involve the integration of emerging technologies like machine learning into the email classification process. Expanding into multimodal approaches for email categorization, exploring incentive models to enhance accuracy, assessing environmental impacts related to email classification practices, and improving user engagement through feedback mechanisms are also crucial areas for exploration. Addressing these considerations and overcoming challenges while navigating new research directions is essential for advancing the overall effectiveness and practicality of email classification, with a specific emphasis on enhancing the accuracy of categorizing diverse email content.

CONCLUSION

In conclusion, the fusion of the Novel Naive Bayes Classifier has demonstrated significant enhancements in accuracy compared to the Random Forest Algorithm for email classification. The synergistic utilization of these two models has resulted in heightened precision and reliability, offering a robust solution for effectively categorizing emails. By harnessing the complementary strengths of both algorithms, organizations can make well-informed decisions and provide accurate email classification, contributing to a more efficient communication process. The accuracy value of the Novel Naive Bayes Classifier is recorded at 90.75%, surpassing the accuracy value of the Random Forest Algorithm at 88.25%. Based on the analysis, the Novel Naive Bayes Classifier (90.75%) outperforms the Random Forest Algorithm (88.25%) in improving the accuracy of email classification.

` DECLARATIONS

Conflicts of Interests

No conflict of interest in this manuscript.

Authors Contribution

Author KM was involved in data collection, data analysis and manuscript writing. Author CM was involved in conceptualization, data validation and critical reviews of manuscripts.

Acknowledgements

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

REFERENCES

1. Alawida, Moatsum, Abiodun Esther Omolara, Oludare Isaac Abiodun, and Murad Al-Rajab. 2022. "A Deeper Look into Cybersecurity Issues in the Wake of Covid-19: A Survey." *Journal of King Saud University. Computer and Information Sciences* 34 (10): 8176–8206.
2. Ashton, James J., Aneurin Young, Mark J. Johnson, and R. Mark Beattie. 2023. "Using Machine Learning to Impact on Long-Term Clinical Care: Principles, Challenges, and Practicalities." *Pediatric Research* 93 (2): 324–33.
3. Elrod, James K., and John L. Fortenberry Jr. 2020. "Direct Marketing in Health and Medicine: Using Direct Mail, Email Marketing, and Related Communicative Methods to Engage Patients." *BMC Health Services Research* 20 (Suppl 1): 822.
4. James, Gilad. n.d. *Introduction to Email Client*. Gilad James Mystery School.
5. Jayant, Advait. 2020. *Data Science and Machine Learning Series: Naive Bayes Classifier Advanced Concepts*.
6. Johansen, Igor Cavallini, Roberto Luiz do Carmo, Luciana Correia Alves, and Maria do Carmo Dias Bueno. 2018. "Environmental and Demographic Determinants of Dengue Incidence in Brazil." *Revista de Salud Publica* 20 (3): 346–51.
7. Kang, John, Russell Schwartz, John Flickinger, and Sushil Beriwal. 2015. "Machine Learning Approaches for Predicting Radiation Therapy Outcomes: A Clinician's Perspective." *International Journal of Radiation Oncology, Biology, Physics* 93 (5): 1127–35.
8. Levine, John R., Margaret Levine Young, and Ray Everett-Church. 2007. *Fighting Spam For Dummies*. John Wiley & Sons.
9. Mayo-Smith, Debbie. 2012. *Conquer Your Email Overload: Super Tips and Tricks for Busy People: Super Tips and Tricks for Busy People*. Penguin Random House New Zealand Limited.
10. Minton, Eric. 1900. *Spam and Scams: Using Email Safely: Using Email Safely*. The Rosen Publishing Group, Inc.
11. Perez-Chadid, Daniela A., Ana Cristina Veiga Silva, Zerubabbel K. Asfaw, Saad Javed, Nathan A. Shlobin, Edward I. Ham, Adriana Libório, et al. 2023. "Needs, Roles, and

Challenges of Young Latin American and Caribbean Neurosurgeons." *World Neurosurgery* 176 (August): e190–99.

12. Qi, Xiaobao, Zhilong Wang, Rongsheng Lu, Jiawei Liu, Yue Li, and Yiping Chen. 2021. "One-Step and DNA Amplification-Free Detection of Listeria Monocytogenes in Ham Samples: Combining Magnetic Relaxation Switching and DNA Hybridization Reaction." *Food Chemistry* 338 (February): 127837.
13. 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.
14. Rudnitskaya, Alena. 2009. *The Concept Of Spam In Email Communication*. GRIN Verlag.
15. Sabry, Fouad. 2023. *Email Spam: Fundamentals and Applications*. One Billion Knowledgeable.
16. Spammer-X, Spammer-X. 2004. *Inside the SPAM Cartel: By Spammer-X*. Elsevier.
17. Srinivasarao, Ulligaddala, and Aakanksha Sharaff. 2023. "SMS Sentiment Classification Using an Evolutionary Optimization Based Fuzzy Recurrent Neural Network." *Multimedia Tools and Applications*, April, 1–32.
18. Temple, Norman J. 2010. "The Marketing of Dietary Supplements in North America: The Emperor Is (almost) Naked." *Journal of Alternative and Complementary Medicine* 16 (7): 803–6.
19. Wang Pingping, and Cai Kedan. 2023. "[Application progress of machine learning in kidney disease]." *Zhonghua wei zhong bing ji jiu yi xue* 35 (12): 1331–34.
20. Wolfe, Paul, Charlie Scott, and Mike Erwin. 2004. *Anti-Spam Tool Kit*. McGraw-Hill/Osborne Media.

TABLES AND FIGURES

Table1. Accuracy and Loss Analysis of Naive Bayes Classifier

Iterations	**Accuracy(%)**
1	**80**
2	**81**
3	**81**
4	**82**
5	**83**

6	**84**
7	**97**
8	**96**
9	**97**
10	**99**
11	**79**
12	**80**
13	**80**
14	**80**
15	**81**
16	**82**
17	**82**
18	**83**
19	**84**
20	**80**
21	**84**

Table2. Accuracy and Loss Analysis of Random Forest

Iterations	**Accuracy(%)**
1	**80**
2	**81**
3	**81**
4	**82**
5	**83**

6	**84**
7	**97**
8	**96**
9	**97**
10	**99**
11	**52**
12	**54**
13	**55**
14	**57**
15	**58**
16	**64**
17	**65**
18	**66**
19	**66**
20	**67**
21	**67**

Table 3. Group Statistical Analysis of Naive Bayes Classifier and Random Forest. Mean, Standard Deviation and Standard Error Mean are obtained for 21 samples. Naive Bayes Classifier has higher mean accuracy and lower mean loss when compared to Random Forest.

	Group	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Accuracy	Naive Bayes	21	84.38	6.569	1.433

	Random Forest	21	73.86	15.399	3.360

INDEPENDENT SAMPLES TEST

		Levene's test for equality of variances		T-test for equality means with 95% confidence interval						
		f	**Sig.**	**t**	**df**	**Sig. (2-tailed)**	**Mean difference**	**Std.Error difference**	**Lower**	**Upper**
Accuracy	**Equal variances assumed**	23.879	.001	2.881	40	.006	10.524	3.653	3.140	17.907
	Equal Variances not assumed				27.045	.008	10.524	3.653	3.028	18.019

Table 4. Independent Sample T-test: Naive Bayes Classifier is insignificantly better than Random forest with p value <.001(Two tailed, p<.006)

Table 5. Comparison of the Naive Bayes Classifier and Random Forest with their accuracy

CLASSIFIER	ACCURACY(%)
Naive Bayes Classifier	90.75
Random Forest	88.25

G GRAPH

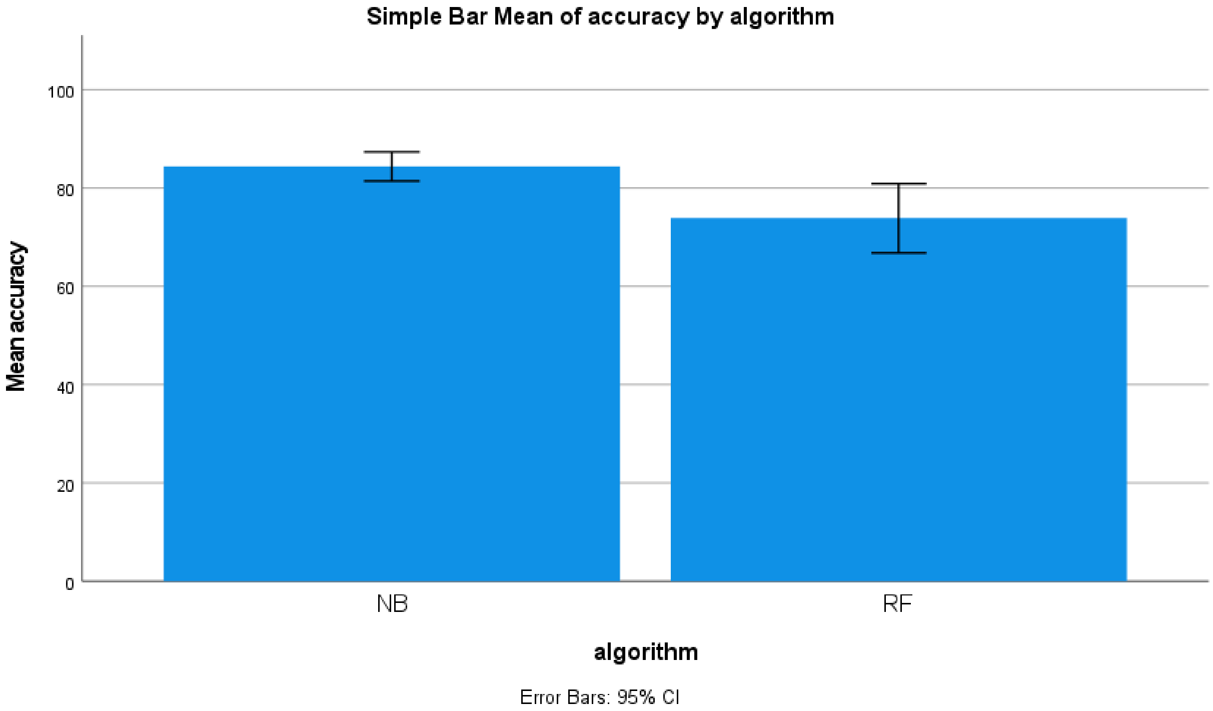

Fig 1. Comparison of Naive Bayes Classifier and Random Forest. Classifier in terms of mean accuracy and loss. The mean accuracy of Naive Bayes Classifier is better than Random Forest. Classifier; Standard deviation of Naive Bayes Classifier is slightly better than Random Forest. X Axis: Naive Bayes Classifier Vs Random Forest Classifier and Y Axis: Mean accuracy of detection with //mean value//