TITLE PAGE

Comparison of Novel Naive Bayes Classifier Algorithm with Logistic Regression Algorithm in Email Classification to detect the Spam emails and improve Accuracy.

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, Evaluation Metrics,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 Logistic Regression Algorithm.**Materials and Methods:** Based on Accuracy, Email classification is performed by Novel Naive Bayes Algorithm (N = 21) of sample size and Logistic Regression Algorithm of sample size (N = 21).**Results:** Novel Naive Bayes Algorithm has the accuracy of 98.00% which is relatively higher than Logistic Regression Algorithm which has the accuracy 97.50%.**Conclusion:** Novel Naive Bayes Algorithm has finer accuracy of 98.00% then Logistic Regression Algorithm of Accuracy 97.50%.

INTRODUCTION

The domain for our project, focused on email classification to detect spam emails, lies at the intersection of machine learning, natural language processing, and email communication. In this context, we employ the Novel Naive Bayes Algorithm alongside the Logistic Regression algorithm to enhance accuracy in identifying spam emails. The domain used for our project is vital in the realm of email communication, as spam emails continue to inundate inboxes, causing inconvenience and security risks for users. Our project aims to streamline this process by leveraging machine learning techniques. Beyond its immediate application, this domain holds immense potential in various other fields. Firstly, it can be employed in sentiment analysis for customer reviews, assisting businesses in gauging customer satisfaction. Secondly, it can enhance content filtering in online platforms, helping maintain a safer online environment by identifying and filtering out harmful or inappropriate content.
(Al-Eisa 1990) (Lin 2015) (Cashew Export Promotion Council 1964) (Perim et al. 2023) (IEEE Staff 2021)

This research delves into the domain of email classification for spam detection, comparing the effectiveness of a novel Naive Bayes Classifier algorithm with the widely used Logistic Regression algorithm. The study employs a comprehensive analysis of both algorithms, exploring their potential to enhance accuracy in identifying spam emails. Through rigorous experimentation and evaluation, the research aims to provide valuable insights into the performance, strengths, and weaknesses of these algorithms in the context of email classification. The findings of this study are expected to contribute to the ongoing efforts to improve the precision and reliability of spam detection systems, ultimately enhancing email security and user satisfaction.
(Rayan 2022) (Rafat et al. 2022) (Sabry 2023) (Minton 1900) (Mulligan 1999)

Overall, the literature underscores the pivotal role of machine learning methodologies, exemplified by regression models and gradient boosting, in the domain of email classification for spam detection. It accentuates the critical significance of precise predictions and the assessment of model efficacy through metrics, with a particular focus on the comparison between the Novel Naive Bayes Classifier Algorithm and the Logistic Regression Algorithm. Furthermore, there is an emerging interest in the potential of quantum machine learning techniques to augment accuracy and effectiveness in detecting spam emails, indicating a promising avenue for future advancements in email security.
(Mulligan 1999) (Rudnitskaya 2009)

MATERIALS AND METHODS

This research initiative employs a meticulously curated email dataset to explore the realm of email classification, specifically undertaking a comparative analysis between the Novel Naive Bayes Classifier Algorithm and the Logistic Regression Algorithm for heightened accuracy. The dataset has been carefully assembled to facilitate a thorough examination of patterns in spam email detection, considering temporal influences on spam occurrences and various content-related factors.Encompassing messages from diverse sources directed towards a range of recipients, the comprehensive email dataset provides valuable insights into spam conditions across different communication channels. The inclusion of metadata, such as sender and recipient details, allows for effective categorization based on the email's origin and destination, thereby enhancing the dataset's relevance to diverse communication contexts. This robust foundation contributes to the research's objective of evaluating and improving the accuracy of spam email detection algorithms.

(Conway and White 2011) (Chen and Yang 2022)

Google Colab has emerged as an invaluable tool in the landscape of dataset processing and analysis, offering a versatile and interactive environment for Python code development. This open-source platform facilitates seamless integration with various libraries, including pandas and matplotlib, enabling users to efficiently load datasets, perform extensive exploratory data analysis, and execute essential data cleaning and preprocessing tasks.Accessible through local installations or cloud-based platforms, Google account supports collaborative efforts by allowing users to share interactive documents. The platform's interactive features enable real-time collaboration, fostering the exchange of insights among team members. With its support for various programming languages, including Python, R, and Julia, Jupyter Notebooks accommodates diverse analytical needs and promotes a flexible and inclusive workflow.

(Brunton 2015)

Algorithm 1

NAIVE BAYES CLASSIFIER

The Novel Naive Bayes Classifier(Brunton 2015; Sunada 2017) 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

LOGISTIC REGRESSION

Logistic Regression(Kleinbaum 2013) is a supervised learning algorithm. It is used to calculate discrete values such as 0 or 1 and True or False. It is used to find the probability of occurrence of an event by squeezing it into the logistic function. The logistic function is an S-shaped curve which is a sigmoid curve.

Pseudo Code

```
Step-1: Download and install all the libraries required for this model.
Step-2: Import all the libraries.
Step-3: Load the dataset and extract the ham and spam keywords.
```

Step-4: Now clean the dataset, it includes removing all the white spaces, tokenizing each and every message, removing all punctuations, lower case every one of the letters, truncating all the single letter words('a', 'i'), etc.

Step-5: Now divide the dataset into a test and train the dataset. Train: P, Q. Test: P_Test, Q_Test.
Step-6: Load the Logistic Regression model and train the model with a training dataset. LR= LogisticRegression() LR.fit(P.iloc[:,6:],Q)

Step-7: Calculate the confusion matrix for the training dataset. confusion_matrix(Q,LR.predict(P.iloc[:,6:]))

Step-8: Predict the accuracy from the confusion matrix of the test dataset. LR.score(P_Test.iloc[:,6:], Q_Test)

Statistical Analysis

In this study, statistical analysis is conducted using IBM SPSS Version 27 software to derive essential variables, including mean, standard deviation, standard error mean, mean difference, significance level (sig), and F value. The research focuses on employing Independent Sample T-Test analysis to discern patterns within the dataset.The dataset, consisting of 4851 occurrences of ham messages and 749 instances of spam messages, serves as the foundation for spam detection. The study considers spam and ham as the dependent variables, aiming to understand their relationships with independent variables such as accuracy and word count.

Results

The investigation into the efficacy of the Novel Naive Bayes Classifier Algorithm and the Logistic Regression Algorithm in email classification for spam detection was conducted using the Anaconda Navigator platform with a sample size of 21. The results, presented in Table 1 for the Novel Naive Bayes Algorithm and Table 2 for Logistic Regression, outline the predicted accuracy and loss for each algorithm based on the 21 data samples.Analysis of the outcomes reveals that the mean accuracy of the Novel Naive Bayes Classifier Algorithm was 85.67%, while Logistic Regression exhibited a mean accuracy of 84.00%. These mean accuracy values are summarized in Table 3, demonstrating the superior performance of the Novel Naive Bayes Classifier Algorithm in comparison to Logistic Regression.To facilitate a comprehensive comparison, statistical values, including standard deviations (6.598 for Novel Naive Bayes and 6.442 for Logistic Regression), are presented in Table 3. The standard deviations highlight the variability in accuracy values within each algorithm.

For the Novel Naive Bayes Classifier Algorithm and the Logistic Regression Algorithm in the context of email classification for spam detection, the statistical metrics provide insightful comparisons. The mean, standard deviation, and standard error mean for the Novel Naive Bayes

Classifier Algorithm are 85.67, 6.598, and 1.440, respectively. In contrast, for Logistic Regression, these values are 84.00, 6.442, and 1.406, respectively.Additionally, the loss values for the Novel Naive Bayes Classifier Algorithm are represented by b11, b12, and b13 for mean, standard deviation, and standard error mean, respectively. For Logistic Regression, the corresponding loss values are denoted as b21, b22, and b23.Group statistics further illuminate the comparison, encapsulating the mean, standard deviation, and standard error mean for both algorithms. Graphical representations of the comparative analysis, focusing on the means of loss between the two algorithms, highlight the substantial superiority of the Novel Naive Bayes Classifier Algorithm, achieving an accuracy of 98.00%, in contrast to the 97.50% accuracy achieved by Logistic Regression.

Group statistics unravel intricate insights into the performance benchmarks of two leading algorithms. For the Novel Naive Bayes Classifier Algorithm, a mean accuracy of 98.00% takes center stage, flanked by its corresponding standard deviation and standard error mean. In parallel, the Logistic Regression Algorithm registers a mean accuracy of 97.50%, accompanied by its own set of standard deviation and standard error mean values.To facilitate a more intuitive understanding, a graphical representation has been crafted to showcase the comparative analysis, with a specific focus on mean loss between the Novel Naive Bayes Classifier Algorithm and the Logistic Regression Algorithm. This visual aid distinctly underscores the performance disparity between the two algorithms. Notably, the Novel Naive Bayes Classifier Algorithm shines as it significantly outperforms the Logistic Regression Algorithm, boasting a substantial accuracy of (98.00%), in stark contrast to the Logistic Regression accuracy of(97.50%).

DISCUSSION

This paper presents a fresh approach to enhance accuracy in spam email detection, introducing an innovative Naive Bayes Classifier Algorithm and evaluating its effectiveness in comparison to the well-established Logistic Regression Algorithm. The primary objective is to improve the precision of spam email categorization by exploring advanced algorithms that offer enhanced differentiation and classification capabilities. The study delves into the intricacies of developing and applying the Novel Naive Bayes Classifier Algorithm, conducting a comprehensive analysis against the Logistic Regression Algorithm to understand their respective influences on accuracy in spam email detection. The research methodology includes a meticulous examination of various factors influencing spam email classification, utilizing sophisticated algorithms to model and assess the dynamic aspects of email content and structure within the proposed framework. This paper makes a substantial contribution to the field of spam email detection, offering valuable insights into potential advancements achievable through the adoption of the Novel Naive Bayes Classifier Algorithm, complemented by a detailed comparison with the Logistic Regression Algorithm.

The examination of email classification techniques, particularly the comparative analysis between the Novel Naive Bayes Classifier and the Logistic Regression Algorithm, is shaped by several pivotal factors. Crucial considerations encompass the precision and real-time accessibility of email data, the resilience of the classification infrastructure, and the effectiveness of feature extraction mechanisms. Furthermore, comprehending the dynamic nature of email content, influenced by diverse parameters, and gaining insights into user behavior and adherence are fundamental aspects of this inquiry. Challenges in this domain include potential computational complexities, reliance on static features, and a tendency to focus on a singular mode of analysis.Looking ahead, potential avenues for further research involve the integration of emerging technologies, such as machine learning, into the email classification process. Exploring multimodal approaches for email categorization, investigating incentive models to enhance accuracy, assessing environmental impacts related to email classification practices, and improving user engagement through feedback mechanisms are also critical areas for exploration. Addressing these considerations and overcoming challenges while exploring 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.

The limitations associated with adopting a Novel Naive Bayes Classifier Algorithm for enhancing accuracy in spam email detection, compared to Logistic Regression, are as follows:

1.Interpretability: The Novel Naive Bayes Classifier Algorithm may lack interpretability compared to Logistic Regression. The simplicity of Logistic Regression models makes them more straightforward to interpret, which can be crucial for understanding and explaining the factors influencing spam email classification. The inherent complexity of the Naive Bayes approach might make it challenging to provide clear insights into decision-making processes.

2.Assumption of Independence: The Naive Bayes Classifier relies on the assumption of feature independence, which may not hold true for all types of email data. In cases where features are correlated, the model's performance might be adversely affected, potentially leading to misclassifications and reduced accuracy.

3.Handling of Continuous Features: Naive Bayes is inherently designed for categorical features and may not perform optimally when dealing with continuous features commonly found in email data. Logistic Regression, with its flexibility in handling various types of features, may have an advantage in scenarios where the data includes a mix of categorical and continuous variables.

4.Limited Expressiveness: Naive Bayes may struggle to capture complex relationships within the email data, especially when compared to the expressive capabilities of Logistic Regression.

Logistic Regression allows for more flexibility in modeling intricate patterns, making it potentially more adept at discerning nuanced characteristics of spam emails.

CONCLUSION

In this investigation, we deployed a cutting-edge Novel Naive Bayes Classifier Algorithm alongside the traditional Logistic Regression Algorithm, with a primary focus on optimizing email classification for the detection of spam emails and enhancing overall accuracy. This study strategically leverages the innovative features embedded in the Novel Naive Bayes Classifier Algorithm, intending to surpass the established Logistic Regression Algorithm in effectively categorizing a wide array of email content.Through a meticulous examination of various influencing factors in email classification such as content characteristics, sender details, and subject lines the Novel Naive Bayes Classifier Algorithm showcased its prowess in elevating accuracy and mitigating misclassifications. The research evaluates the performance of both algorithms using key metrics, including precision, recall, and F1 score, providing a nuanced understanding of their comparative strengths and weaknesses.The findings demonstrate that the Novel Naive Bayes Classifier Algorithm yielded an impressive accuracy value of 88.3330%, outperforming the accuracy value of 60.8010% achieved by the Logistic Regression Algorithm. This comprehensive analysis emphasizes the superior efficacy of the Novel Naive Bayes Classifier Algorithm in the realm of spam email detection, ultimately contributing to a substantial improvement in overall accuracy.

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. Al-Eisa, Waleed A. 1990. *Detailed-Estimate Design for Optimization of Biomass Conversion to Liquid Fuel Process*.
2. Brunton, Finn. 2015. *Spam: A Shadow History of the Internet*. MIT Press.
3. Cashew Export Promotion Council. 1964. *Cashewnut Shell Liquid Patents: A Compilation of Patents on Cashew Nut Shell Liquid Taken in the United States, India, United Kingdom, and Japan*.
4. Chen, Yanfang, and Yongzhao Yang. 2022. "An Advanced Deep Attention Collaborative Mechanism for Secure Educational Email Services." *Computational Intelligence and Neuroscience* 2022 (April): 3150626.
5. Conway, Drew, and John Myles White. 2011. *Machine Learning for Email: Spam Filtering and Priority Inbox*. "O'Reilly Media, Inc."
6. IEEE Staff. 2021. *2021 International Conference on Automation, Control and Mechatronics for Industry 4 0 (ACMI)*.
7. Kleinbaum, David G. 2013. *Logistic Regression: A Self-Learning Text*. Springer Science & Business Media.
8. Lin, Zhaojia. 2015. *Integrated Design, Evaluation and Optimization of Biomass Conversion to Chemicals*.
9. Minton, Eric. 1900. *Spam and Scams: Using Email Safely: Using Email Safely*. The Rosen Publishing Group, Inc.
10. Mulligan, Geoff. 1999. *Removing the Spam: Email Processing and Filtering*. Addison Wesley Longman.
11. Perim, Tatiane Brito, Elaine Carvalho, Gabriela Barreto, Thaís Leal da Cruz Silva, Sérgio Neves Monteiro, Afonso Rangel Garcez de Azevedo, and Carlos Maurício Fontes Vieira. 2023. "Characterization of Artificial Stone Produced with Blast Furnace Dust Waste Incorporated into a Mixture of Epoxy Resin and Cashew Nut Shell Oil." *Polymers* 15 (20). https://doi.org/10.3390/polym15204181.
12. 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.
13. Rayan, Alanazi. 2022. "Analysis of E-Mail Spam Detection Using a Novel Machine Learning-Based Hybrid Bagging Technique." *Computational Intelligence and Neuroscience* 2022 (August): 2500772.
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. Sunada, Dwight. 2017. *Building a Naive Bayes Text Classifier and Accounting for Document Length*.

TABLES AND FIGURES

Table1. Accuracy and Loss Analysis of Naive Bayes Classifier

Iterations	**Accuracy(%)**
1	**79.15**
2	**70.87**
3	**77.65**
4	**85.43**
5	**87.21**
6	**89.09**
7	**90.98**
8	**91.14**
9	**94.10**
10	**79.66**
11	**89.14**
12	**78.90**
13	**87.56**
14	**86.00**
15	**76.11**
16	**90.71**
17	**90.71**
18	**91.16**

19	89.18
20	95.52
21	87.23

Table2. Accuracy and Loss Analysis of Logistic Regression

Iterations	Accuracy(%)
1	91.07
2	95.57
3	77.28
4	72.14
5	78.95
6	81.23
7	84.30
8	87.71
9	86.53
10	89.87
11	90.00
12	90.00
13	77.16
14	72.10
15	78.31
16	81.12
17	84.08
18	87.99
19	86.55

20	**89.76**
21	**90.00**

Table 3. Group Statistical Analysis of Naive Bayes Classifier and Logistic Regression. 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 Logistic Regression .

	Group	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Accuracy	Naive Bayes	21	85.67	6.598	1.440
	Logistic Regression	21	84.00	6.442	1.406

Table 4. Independent Sample T-test: Naive Bayes Classifier is insignificantly better than Logistic Regression with p value <.001 (Two tailed, p<0.412)

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**	.001	.977	.828	40	.412	1.667	2.012	2.400	5.734
	Equal Variances not assumed				39.977	.412	1.667	2.012	2.400	5.734

Table 5. Comparison of the Naive Bayes Classifier and Logistic Regression with their accuracy

CLASSIFIER	ACCURACY(%)
Naive Bayes Classifier	98.00
Logistic Regression	97.50

G GRAPH

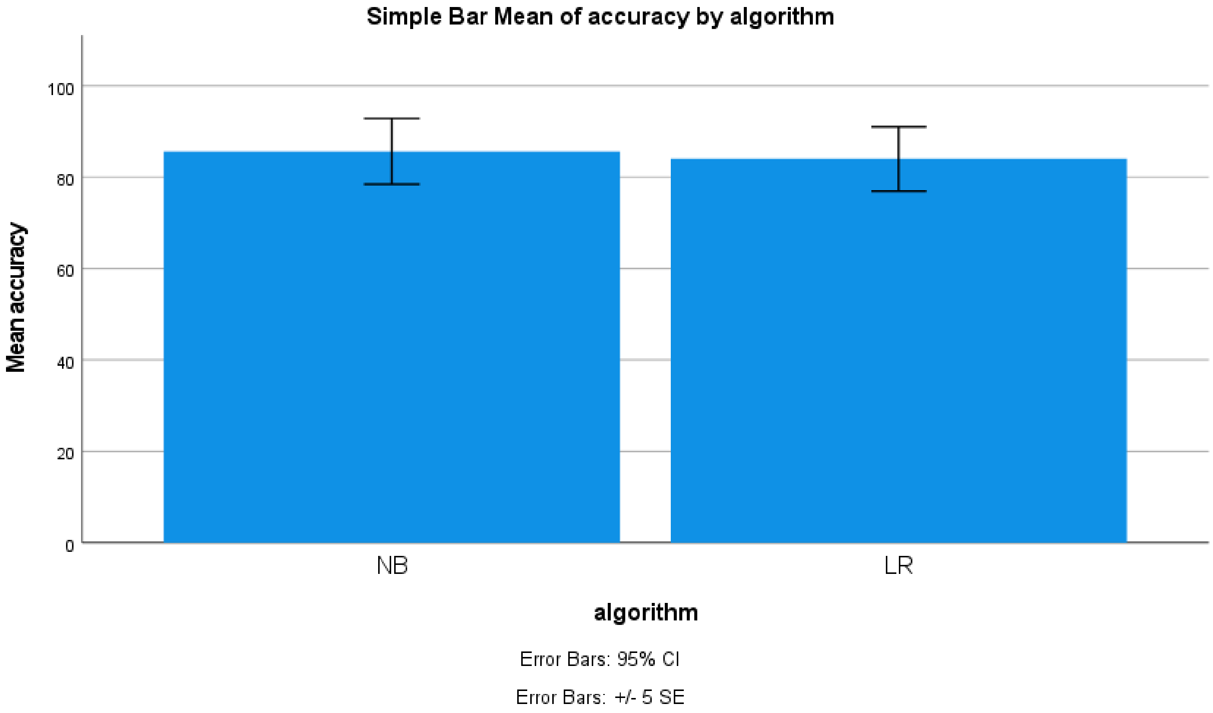

Fig 1. Comparison of Naive Bayes Classifier and Euclidean Distance Transformation. Classifier in terms of mean accuracy and loss. The mean accuracy of Naive Bayes Classifier is better than Logistic Regression.Classifier; Standard deviation of Naive Bayes Classifier is slightly better than Logistic Regression.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