IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 3, March 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203063 |

# A Comprehensive Study on Email Spam Detection Techniques

**Aaron Horta, Charul Gupta, Meeka Gandham, Shweta Sharma, Akshith K J, Kavitha R**

Dept. of Computer Science and IT, JAIN (Deemed to be University), Bengaluru, India

**ABSTRACT**: The analysis of email spam detection methods is pivotal in the realm of cybersecurity. This study meticulously explores a variety of techniques utilized for discerning spam emails, encompassing traditional methodologies alongside sophisticated approaches such as machine learning algorithms. The examined techniques include Support Vector Machines, Naive Bayes Multinomial Classifier and Random Forest Classifier with a primary objective of identifying the most effective approach among them. By scrutinizing each method's strengths and weaknesses, encompassing factors like detection accuracy and computational efficiency, this research seeks to offer a comprehensive overview of their utility in combatting email spam. Additionally, the study investigates the impact of human elements, such as user education and awareness initiatives, in bolstering overall detection performance. By amalgamating insights from various scholarly endeavors, this inquiry not only enriches the understanding of contemporary spam detection practices but also delineates pathways for future research endeavors and informs the formulation of resilient cybersecurity protocols.

**KEYWORDS**: Email spam detection, Machine learning algorithms, Support Vector Machines, Naive Bayes Classifier, Random Forest Classifier, Cybersecurity.

## I. INTRODUCTION

In the contemporary era of pervasive digital connectivity, electronic mail (email) stands as a cornerstone of modern communication, facilitating rapid and economical dissemination of information across global networks. However, this widespread adoption of email has also given rise to a burgeoning issue: the proliferation of spam emails. Often characterized by their unsolicited nature and bulk dissemination, these spam emails inundate inboxes worldwide, posing a multifaceted challenge to both individual users and organizations alike. Over the years, the volume of spam has surged dramatically, propelled by various factors including the exponential growth of internet usage and the proliferation of botnets, which exploit compromised computers to propagate unsolicited messages. Despite advancements in spam detection algorithms and techniques, achieving flawless accuracy in identifying and filtering out spam remains an elusive goal, beset by the inherent complexities of unstructured data and the vast array of features inherent in email content. Thus, the quest to identify the optimal spam classification algorithm assumes paramount importance, driven by the imperative to enhance both the quality and efficiency of email filtering mechanisms amidst the ever-evolving landscape of cyber threats.

## II. LITERATURE REVIEW

Email spam, commonly known as junk email, constitutes unsolicited messages sent to recipients without explicit consent. These messages, often commercial in nature, pose significant risks as they may contain links to phishing websites or malware-infected attachments. Spammers employ various tactics to harvest email addresses, including automated programs known as spambots, which scour the internet for targets. The proliferation of spam has surged since the early 1990s, presenting a pervasive challenge to email users worldwide. Spam emails manifest in diverse forms, ranging from outright scams to ostensibly legitimate business ventures. They often tout pharmaceutical drugs, weight loss programs, job opportunities, and online gambling services. Moreover, spam serves as a conduit for fraudulent activities, such as advance-fee scams and phishing schemes, where perpetrators deceive users into divulging sensitive information. As such, the detection and filtration of spam emails have become imperative to safeguard users and organizations from potential harm.

Among the plethora of approaches, Kishore Kumar et al. [1] conducted a comparative study on email spam classification using data mining techniques. They employed the TANAGRA data mining tool to analyze a spam dataset from the UCI machine learning repository. Their methodology involved feature construction, selection, and evaluation of 15 different classification algorithms. Through rigorous experimentation, they identified the best classifier for spam detection, achieving promising accuracy rates. From traditional methods like C4.5 and ID3 decision trees to more

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203063 |

sophisticated approaches such as Support Vector Machines (SVM) and Random Forests, a wide spectrum of algorithms has been evaluated for their accuracy and effectiveness in spam classification.

The study by Taylor and Ezekiel (2020) proposes a model for spam email detection utilizing Support Vector Classifier and Random Forest Classifier. Employing a University Collection London (UCL) spam base dataset, the authors preprocess the data using min_max_scaler to ensure proper scaling for effective analysis. With 58 columns in the dataset denoting various email attributes, the model is trained and tested, achieving an accuracy of 89.21% with Support Vector Classifier and 95.36% with Random Forest Classifier. Notably, Random Forest Classifier emerges as the more accurate algorithm and is subsequently employed for spam classification. This research underscores the efficacy of machine learning techniques in combating spam emails, offering insights into algorithm performance and dataset preprocessing.[3]

Dhiman, Jakobsson, and Yen (2017) present a methodology for utilizing Unicode confusable characters to obscure scam messages effectively evading existing email filters while maintaining message readability. Through a systematic process of selecting high-fidelity confusable characters, the authors develop an algorithm capable of generating visually identical but encoded differently scam messages. By testing these obfuscated messages against email filters deployed by major providers like Yahoo, Hotmail, and Gmail, the study reveals the limitations of current spam detection systems. Despite advancements in spam filter technology, obfuscated scam messages successfully bypass detection, highlighting the need for enhanced detection mechanisms. The authors propose strategies to counter homograph attacks, emphasizing the importance of proactive implementation to mitigate risks associated with evolving spam tactics. The research underscores the urgency for improved protective measures to safeguard users against increasingly sophisticated scam techniques.[4],[8]

Rathod and Pattewar developed a spam detection system using a Bayesian Classifier, aiming to address the pervasive issue of spam emails in digital communication. Their focus lies on content-based filtering techniques to distinguish between legitimate and spam emails, considering aspects like email body, subject, and URLs. By evaluating the classifier's performance with metrics such as accuracy, error rate, precision, and recall, they provide insights into its effectiveness. The research contributes to existing literature on spam classification, emphasizing the importance of robust spam detection systems for user privacy and digital security.[2]

Ma, Yamamori, and Thida (2020) conducted research comparing the effectiveness of Naive Bayes Classifier and Support Vector Machine (SVM) for email spam classification. Their study focused on the growing issue of spam emails and the necessity for automatic email spam filters. Through experimentation using data from the Enron corpus, they found that SVM consistently outperformed Naive Bayes Classifier across various training data sizes, demonstrating higher accuracies, precision, recall, and F-measure. This suggests that SVM is more effective in accurately classifying spam emails, highlighting its importance in combating spam and enhancing email security. Further research avenues include exploring richer email content and other supervised learning algorithms to improve performance.[5]

Sharma and Bhardwaj delve into machine learning-based spam email detection, focusing on the Naïve Bayes classifier and J48 decision tree algorithm. They detail the Naïve Bayes classifier's probability-based approach, highlighting its efficiency in supervised learning and multiclass capability. The Multinomial Naïve Bayes classifier, utilized in their work, represents data as word vectors and calculates probabilities based on word counts. Their spam mail detection system integrates these classifiers into a hybrid bagged approach, combining their strengths for email classification. The workflow involves dataset preparation, preprocessing, feature selection, and classification using the hybrid bagged approach. Experimental results showcase the system's accuracy, precision, recall, and other metrics, underscoring the effectiveness of the hybrid bagged approach. The study concludes by suggesting future enhancements such as incorporating boosting techniques for further improvement.[6],[9]

Sharma and Sharma's study delves into email spam filtering, focusing on origin-based and content-based techniques. Origin-based methods like blacklists and whitelists use network data to identify spam, while content-based filters analyze email content using rules or machine learning. Despite their effectiveness, origin-based techniques face challenges in maintaining accuracy. Content-based filters, including rule-based, Bayesian, SVM, and ANN, offer advanced approaches, showing promise in accurately classifying spam. The study also discusses comparative research on spam filtering algorithms, highlighting the need for ongoing investigation. It proposes future research directions, such as improving cruel spam recognition using feature selection and classifiers like SVM and Naïve Bayes. Overall,

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203063 |

the study emphasizes the importance of robust spam filtering mechanisms and suggests avenues for further advancement.[7]

## III. PROPOSED METHODOLOGY

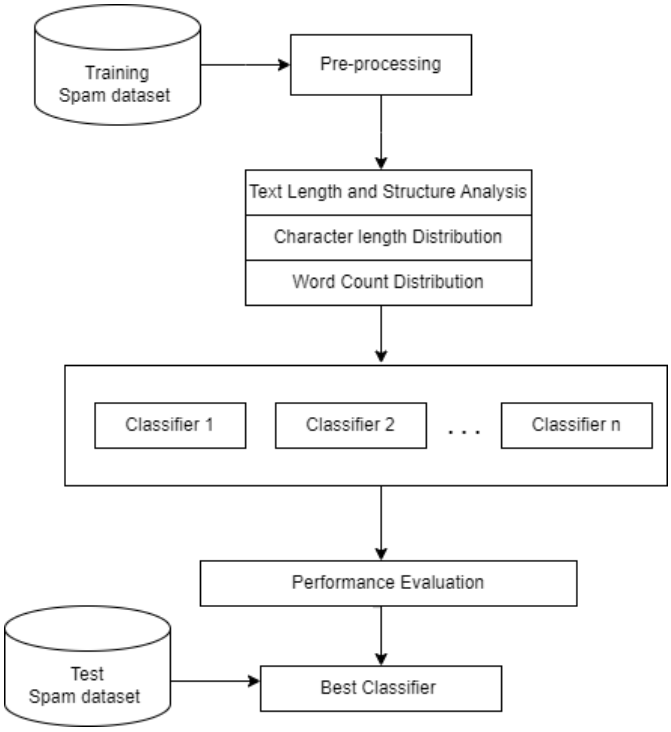


Fig.1 Proposed System Architecture

*A. Spam dataset*

Spam datasets are essential for training and evaluating spam filtering systems. They consist of labeled email samples, enabling researchers to develop and test machine learning models for accurate spam classification. These datasets play a vital role in improving spam detection technology by providing a diverse range of email examples to train algorithms and adapt to evolving spamming techniques.

B. *Pre-Processing*

In modern data analysis practices, the journey often begins with the intricate process of preprocessing, where raw data is refined and shaped to ensure its quality and suitability for further analysis. This critical phase acknowledges the ubiquitous nature of incomplete, noisy, or aggregated datasets prevalent in real-world scenarios. Through meticulous steps including data cleaning, integration, transformation, and reduction, the dataset undergoes a transformational journey aimed at enhancing its overall integrity and utility. Within this framework, specific techniques such as Word Tokenization and word cloud generation are employed, facilitating the breakdown of text data into smaller, analyzable units and providing visual representations of word frequencies, respectively. These preprocessing steps lay the foundation for subsequent analyses, empowering researchers to glean valuable insights and uncover meaningful patterns from the data.

By prioritizing robust preprocessing methodologies, researchers can effectively navigate the complexities inherent in contemporary data analysis endeavors. The adoption of techniques like Word Tokenization and word cloud generation not only ensures data integrity but also facilitates streamlined analyses by breaking down textual information into manageable components and offering visual representations of key insights. As researchers embark on their analytical journey, these preprocessing techniques serve as invaluable tools, empowering them to derive actionable insights and make informed decisions based on a foundation of meticulously prepared data.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203063 |

C. *Classifiers*

There are three main algorithms that we have focused on while accomplishing this research. Those are Naïve bayes Multinomial classifier, Support Vector machine classifier and Random Forest classifier which is a decision tree-based classifier.

Naïve Bayes Multinomial Classifier

The Naive Bayes Multinomial (MNB) classifier assumes a pivotal role in the realm of spam detection, showcasing its indispensability through its adeptness at handling textual data, particularly in the intricate domain of email analysis. Its probabilistic approach imbues it with the prowess to discern patterns within emails, rendering it tailor-made for scrutinizing vast volumes of electronic correspondence. At its core, the MNB classifier operates by undergoing training on meticulously labeled datasets comprising both spam and legitimate emails. Through this process, it meticulously dissects the frequency of occurrence of individual words within each category, effectively deciphering the distinctive linguistic nuances characteristic of spam versus legitimate correspondence.

Upon encountering a novel email, the MNB classifier springs into action, leveraging its acquired knowledge to calculate the likelihood of each word appearing within the email, contingent upon its categorization as spam or ham. This intricate computation involves the multiplication of individual probabilities, culminating in an overarching probability assessment regarding the email's affiliation with each category. Through this nuanced analysis, the MNB classifier empowers spam detection systems with the capability to make informed predictions, thereby facilitating the accurate identification and segregation of unsolicited spam from genuine communication.

Support Vector Machine Classifier

Support Vector Machines (SVMs) offer a formidable alternative to Naive Bayes in the realm of spam detection. Unlike Naive Bayes, SVMs are adept at capturing intricate relationships between words, without relying on the assumption of word independence. By mapping emails into a high-dimensional feature space using techniques like bag-of-words or TF-IDF, SVMs uncover nuanced patterns within the email content. This enables the algorithm to discern optimal hyperplanes that effectively segregate spam and legitimate emails, ensuring accurate classification. In practice, SVMs excel at handling non-linear relationships between words, contributing to their high accuracy in spam detection, particularly when coupled with meticulous feature engineering efforts.

Random Forest Classifier

The Random Forest Classifier, renowned for its efficacy in spam detection, employs an ensemble approach by amalgamating multiple decision trees. This methodology is characterized by the creation of a diverse forest, wherein each decision tree is trained on a random subset of features extracted from the email dataset. This deliberate diversification strategy mitigates the risk of overfitting and ensures robustness in classification.

As new emails arrive for classification, each decision tree independently assesses them based on its unique set of learned decision rules. Subsequently, the classifier aggregates the individual predictions from all the trees and determines the final classification through a majority vote mechanism. This collective decision-making process harnesses the collective wisdom of the entire forest, resulting in a robust and reliable spam detection system capable of effectively discerning between spam and legitimate emails.

## IV. EXPERIMENTATIONS

A. *About Dataset*

The spam dataset utilized in the research was sourced from Kaggle, a widely recognized repository for research data accessible on the internet, providing a valuable resource for scholarly investigation. Comprising approximately 5574 instances of messages in the English language, the dataset was meticulously tagged to differentiate between legitimate (ham) and spam messages, facilitating the classification and analysis of email content. Kaggle's repository serves as a convenient and reliable platform for researchers to access diverse datasets, contributing to the advancement of various fields through data-driven analyses and experimentation. The dataset comprises messages organized line by line, featuring two columns: "v1" indicating whether the message is "ham" (regular) or "spam" (unwanted), and "v2" containing the original text. From the Grumble text website, 425 spam messages were manually gathered, sourced from discussions where users shared their encounters with spam messages without disclosing specific content. Identifying spam messages necessitated meticulous scrutiny of various web pages. Furthermore, a subset of 3,375 randomly selected regular messages from the NUS SMS Corpus (NSC) was incorporated. The NSC, consisting of approximately

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203063 |

10,000 legitimate messages, was amassed for research at the Department of Computer Science, National University of Singapore. These messages, primarily from Singaporean individuals, including many university students, were contributors who were informed of the public sharing of their contributions.

*B. Text length and Structure analysis*

| | num_characters | num_words | num_sentence |
|---|---|---|---|
| count | 5169.000000 | 5169.000000 | 5169.000000 |
| mean | 78.977945 | 18.455794 | 1.965564 |
| std | 58.236293 | 13.324758 | 1.448541 |
| min | 2.000000 | 1.000000 | 1.000000 |
| 25% | 36.000000 | 9.000000 | 1.000000 |
| 50% | 60.000000 | 15.000000 | 1.000000 |
| 75% | 117.000000 | 26.000000 | 2.000000 |
| max | 910.000000 | 220.000000 | 38.000000 |

Fig.2 Text Length and Strcuture Analysis

In the initial stages of spam detection, text length and structure analysis play a pivotal role. Upon scrutinizing the dataset, notable trends emerge:

1. Concise Communication: The data showcases a preference for brevity, with an average sentence length of approximately 79 characters and a word count around 18. While succinct messages may raise suspicion, this criterion alone carries limited weight. Spammers often attempt to compress significant information into concise emails, thus necessitating a more nuanced approach to detection.
2. Varied Sentence Structures: A critical observation pertains to the considerable diversity in both sentence length (ranging from 2 to 910 characters) and word count (spanning from 1 to 220). This wide-ranging disparity often signifies spam emails, notorious for their grammatical inconsistencies and lack of structural coherence. Consequently, substantial deviations in sentence structure serve as a more reliable indicator during preprocessing.

C. *Character Length Distribution*

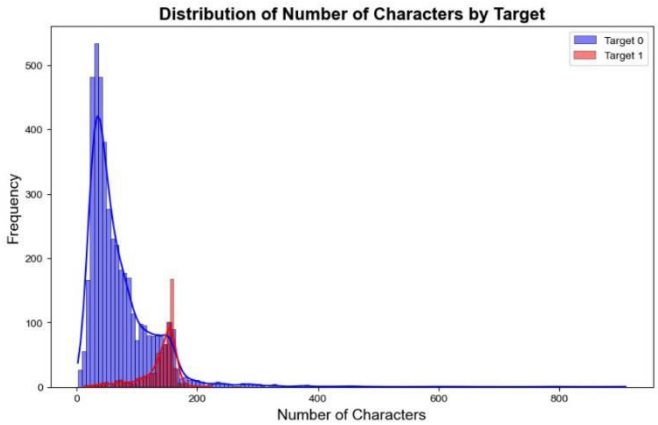


Fig.3 Character Length Distribution

In the domain of initial spam filtering, analyzing character length distribution assumes significance as it provides valuable insights into distinguishing between legitimate emails and spam. The observed bimodal pattern, characterized by peaks around 20 and 60 characters, serves as a useful heuristic for categorizing emails into two distinct groups. Emails falling within Target 0, with a peak around 60 characters, are likely to be legitimate communications due to their typically more extensive content. Conversely, those categorized under Target 1, exhibiting a peak around 20

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203063 |

characters, are indicative of spam emails, which often employ shorter lengths and repetitive phrases to evade detection. However, while character length analysis offers valuable initial indicators, it alone cannot serve as a definitive determinant of spam. A comprehensive spam detection strategy necessitates the incorporation of additional factors, such as identification of spam-specific elements within the email content, examination of sender details and email headers for suspicious patterns, and the utilization of machine learning algorithms trained on diverse datasets to enhance filtering accuracy.

By amalgamating character length analysis with these supplementary techniques, email filtering systems can effectively discern between legitimate correspondence and unwanted spam, thereby fortifying the integrity of communication channels and ensuring that users receive only relevant and authentic content in their inboxes. This multifaceted approach not only enhances the accuracy of spam detection but also reduces the risk of false positives and negatives, thereby optimizing the efficiency and reliability of email filtering mechanisms in combating the ever-evolving challenges posed by spam and ensuring a seamless and secure user experience.

D. *Word Count Distribution*

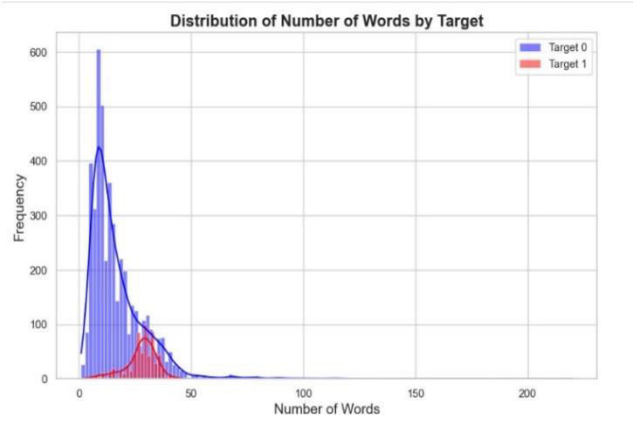


Fig.4 Word Count Distribution

Similar to character length, analyzing word count distribution serves as another initial indicator in spam detection. The observed distribution exhibits a bimodal pattern, with distinct peaks around specific word counts. This suggests the presence of two distinct email groups with differing word usage tendencies:

- Target 0 (Likely Ham): This target presumably corresponds to legitimate emails. The word count distribution peaks around 10 words, aligning with the notion that legitimate emails generally contain more content compared to spam.
- Target 1 (Likely Spam): This target likely represents spam emails. The distribution exhibits a peak around 5 words, which coincides with the tendency of spam emails to be shorter and potentially employ repetitive keywords or phrases to circumvent spam filters.

The x-axis represents the number of words in the email, and the y-axis represents the frequency of emails with that specific word count. It's crucial to acknowledge that word count distribution alone is an insufficient indicator of spam. A comprehensive spam detection strategy necessitates incorporating other factors such as:

- Spam-specific keywords and phrases: Identifying these elements within the email content is vital.
- Sender information and email headers: Scrutinizing these details can reveal suspicious patterns associated with spam.
- Machine learning algorithms: Utilizing algorithms trained on substantial datasets of labelled spam and legitimate emails significantly bolsters the accuracy of spam filtering systems.

## V. RESULT

Based on the statistics of email dataset that is shown in the below figure we can determine that The Support Vector Machine gave an accuracy score of 0.98 and a precision score of 0.97. The same accuracy was scored by the Random Forest Classifier but it was slightly better than Support Vector Machine with regards to precision with a score of 0.98.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203063 |

and the Naïve Bayes multinomial classifier had the same accuracy but precision score of exact 1.0 which makes it overfit.

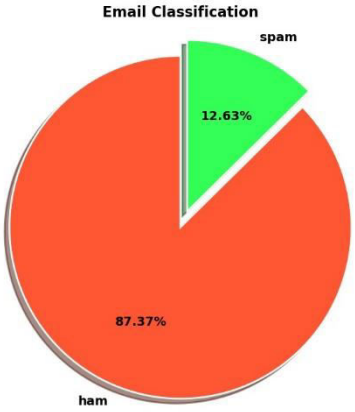


Fig.5 Email classification dataset pie chart

```
For:  SVC
Accuracy:  0.9758220502901354
Precision:  0.9747899159663865

For:  NB
Accuracy:  0.9709864603481625
Precision:  1.0

For:  RF
Accuracy:  0.9758220502901354
Precision:  0.9829059829059829
```

Fig.6 Classification model performance

## VI. CONCLUSION AND FUTURE WORK

The research findings indicate that the Random Forest (RF) classifier emerges as the most effective model for spam classification among the trio of models assessed, namely RF, SVM, and NB. This determination stems from an examination of the confusion matrix, which highlights RF's superior performance in several aspects. Notably, RF showcases elevated values along the diagonal, particularly in correctly classifying both negative and positive instances, indicating a higher number of accurate classifications compared to SVM and NB. Furthermore, RF demonstrates lower off-diagonal values, signifying fewer misclassifications relative to the other models. While all three models exhibit some misclassifications, RF's propensity for fewer errors underscores its efficacy in spam classification. Breaking down the performance based on the confusion matrix, RF notably achieves the highest number of correct classifications, characterized by a reduced occurrence of false positives (legitimate emails misclassified as spam) and false negatives (spam emails misclassified as legitimate). Conversely, the Support Vector Machine (SVM) exhibits commendable performance but demonstrates a slightly higher frequency of misclassifications, particularly in terms of false positives. In contrast, the Naive Bayes (NB) model shows the lowest overall accuracy among the trio, with a notable increase in misclassifications across both false positive and false negative categories.

## REFERENCES

[1] Kumar, R. K., Poonkuzhali, G., & Sudhakar, P. (2012, March). Comparative study on email spam classifier using data mining techniques. In *Proceedings of the international multiconference of engineers and computer scientists* (Vol. 1, pp. 14-16). Newswood Limited, Hong Kong.

[2] Rathod, S. B., & Pattewar, T. M. (2015, April). Content based spam detection in email using Bayesian classifier. In *2015 International Conference on Communications and Signal Processing (ICCSP)* (pp. 1257-1261). IEEE.

[3] Taylor, O. E., & Ezekiel, P. S. (2020). A model to detect spam email using support vector classifier and random forest classifier. *Int. J. Comput. Sci. Math. Theory*, *6*(1), 1-11.

[4] Dhiman, Mayank & Jakobsson, Markus & Yen, Ting-Fang. (2017). Breaking and fixing content-based filtering. 52-56. 10.1109/ECRIME.2017.7945054.

[5] Ma, T. M., Yamamori, K., & Thida, A. (2020, October). A comparative approach to Naïve Bayes classifier and support vector machine for email spam classification. In *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)* (pp. 324-326). IEEE.

[6] P. Sharma, U. Bhardwaj, "Machine Learning Based Spam Email Detection" International Journal of Intelligent Engineering and System 11(3), pp. 1-10, 2018

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203063 |

[7] M. Sharma, S. Sharma, "A Survey of Email Spam Filtering Methods" Control Theory and Informatics 7, 2224-5774, 2018.

[8] S. N. Rekha, "A Review on Different Spam Detection Approaches" International Journal of Engineering Trends and Technology (IJETT). 11 (6), pp. 315-318, 2014.

[9] D. Mallampati. "An Efficient Spam Filtering using Supervised Machine Learning Techniques" International Journal of Scientific Research in Computer Science and Engineering. 6(2), pp.33-37, 2018.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details