Email Spam Detection Using Ensemble Learning

Sanjay S
Department of Computer Science Sathyabama Institute of Science and Technology Chennai, India
sanjaysreeram1@gmail.com

Sasikala S
Department of Computer Science Sathyabama Institute of Science and Technology Chennai, India
103sashi@gmail.com

Anandha Sree. R
Department of Computer Science Sathyabama Institute of Science and Technology Chennai, India
Anandha.sree@gmail.com

***Abstract*—** Email spam is one of the biggest problems of the modern communication technology, which has a significant negative impact on productivity and cybersecurity. Because of the exponential increase in Internet usage, spam emails have increased manyfold, making basic rule-based filtering approaches inefficient owing to their inability to update themselves according to changes in spam behaviour and the high resultant rate of false positives. While there were several approaches to employing machine learning algorithms for spam filtering, studies on spam email filtering using an ensemble learning approach have not been many so far. In this study, spam email filtering using ensemble learning is proposed to make it robust against any change in spamming patterns. Feature vector extraction is performed through the TF-IDF transformation of email texts, while other feature vectors considered in the study are sender details, email length, and the availability of URLs in an email. In this study, a soft voting ensemble classifier combines " Naive Bayes, Logistic Regression And Support Vector Machine " Algorithms for spam email classification. The experiment confirmed that the proposed approach is effective for efficient email spam filtering.

***Keywords*—** *Email Spam Detection, Ensemble Learning, TF-IDF, Machine Learning, Text Classification, Cybersecurity*

1. INTRODUCTION

Email communication has become a vital part of personal and business communication that makes electronic mail services form an important part of our daily lives. Due to the increased volume of email transactions, the number of undesirable messages, or spam has also increased significantly. In addition to the inconvenience it imposes, these spam emails subject their users to significant security risks, which include phishing and financial fraud.

As a result, there has been a high demand to integrate efficient filtering systems into email systems in order to curb the spam traffic [14]. Spam filters are normally automatic, which scans all the messages coming to the system and diverts those that are classified as unwanted to a special spam folder thus improves user experience and system security. A properly developed filtering system can protect the user against phishing, where unwanted emails do not pass their mail inboxes.

Spammers have over time utilized advanced methods like obfuscating text, shortening URLs and other methods of social engineering [7], [12]. These innovations are major difficulties in classifying valid email messages as spam by traditional filtering methods. Besides, the daily increase in the count of emails is further increasing the need to adopt automated and intelligent spam blocking systems in order to maintain safe and reliable communication of information [11].

R. Vasanth Kumar Mehta et al. (eds.), *Proceedings of the International Conference on Intelligent Systems for a Sustainable Future (ISSF 2026)*, Atlantis Highlights in Intelligent Systems 16,
https://doi.org/10.2991/978-94-6239-693-7_108

The majority of the existing spam detection methods are based on rule-based filters or individual machine-learning classifiers, neither of which provides high levels of adaptability to changing spamming patterns and hence lead to high levels of false-positives. This paper will suggest an adaptive ensemble-learning based spam filtering model which combines various classifiers to achieve robustness and improved detection accuracy [1].

2. RELATED WORK

Supervised machine learning methods have been widely studied in the context of spam email classification, based on the features derived using email content, email headers and email hyperlink structure. The first literature was majorly focused on probabilistic or linear classifiers, like Naive Bayes and Logistic Regression, text representations using Bag-of-Words and TF-IDF. Such techniques have been found to perform significantly in the context of classification with carefully designed feature engineering techniques [17], [20]. Since the sophistication of the spam messages has proliferated, scholars have embraced tree based and kernel based classifiers to capture non-linear decision boundaries, in most cases they achieve better detection rates at a cost of compromised generalization with respect to heterogeneous data [9], [14].

In the recent research findings, there has been a gradual development of ensemble learning techniques to reduce the natural constraints of individual classifiers. By aggregating models with complementary strengths, ensemble techniques also contribute to reducing bias–variance trade-offs and stabilizing the results in the case of imbalanced spam datasets [1], [16]. Several studies have reported that the integration of probabilistic, linear, and decision tree-based classifiers using voting mechanisms outperforms single-model approaches in terms of accuracy and robustness of spam detection [4].

Meanwhile, feature extraction techniques have expanded from textual analysis to metadata attributes, including sender information, message length, and link-related characteristics, which play an important role in distinguishing phishing attempts and malicious URLs [12], [15]Most existing systems suffer from either very limited feature representation or isolated classification. This often leads to inconsistent performance in real-world applications of these models. Challenges such as concept drift, false-positive reduction, scalability, and deployment efficiency have been addressed independently rather than within an integrated framework [11], [14].

Based on these observations, this study proposes a soft voting ensemble-based spam filtering approach that jointly considers textual, metadata, and contextual features of the data. It is expected that the integrated framework will improve the reliability and robustness of the classification while delivering Superior Performance Compared To existing standalone or partially integrated solutions [1], [16]

3. METHODOLOGY

The proposed framework conducts email spam detection by integrating text-based feature extraction, metadata analysis, and ensemble-based machine learning classification in a module comprising preprocessing, feature extraction, model training, ensemble classification, and results analysis. Such modular designs enhance the scalability and maintainability of any spam detection system [14]. To enhance robustness and generalization, the framework adopts ensemble learning, thus reducing bias and variance by combining complementary classifiers [1], [16]. More importantly, the combination of probabilistic, linear and tree based model into a single framework has enhanced accuracy in detection, false positive rates, and adaptation to dynamically changing spam patterns significant [4], [13].

3.1 Data Collection and Preprocessing

In this paper, a publicly available email dataset with labels of spam and legitimate (ham) emails, including both subject and body text and some link header features are used in order to identify various spam patterns.

Data pre-processing was done through traditional text-processing methods, including lower casing letters, removing punctuations and other special characters and non-essential stop words in text classification situations [20]. URLs are retained because they contain significant phishing and other malpractice elements in spam emails [15]. " The Dataset was split into " Training and Testing datasets to provide accurate, unbiased results[9].

Type	**Feature/Label**
Input (Text)	Email Subject and Body
Input (Metadata)	Sender Domain, Email Length, URL Count
Feature Extraction	TF-IDF Vectors
Output (Classification)	Spam, Ham
EvaluationMetrics	Accuracy, Precision, Recall, F1-Score

Table I — Key Features And Labels Used In Proposed Syst

3.2. Feature Extraction

Feature Extraction is an important aspect of the proposed email spam detection model, where raw email features are converted into significant numerical forms necessary for machine learning. The hybrid model uses a combination of text features obtained from email messages and features based on metadata attributes, where the model can extract features that describe the context of spam messages. The combined features can enhance the accuracy of classification by acquiring intricate patterns that differentiate spam messages from genuine messages, as stated in [11], [14].

Textural features are identified from the subject line and content of spam or valid emails by utilizing the term frequency-inverse document frequency (TF-IDF) approach, which has been largely used in spam classification owing to its efficiency in identifying key words [17], [20]. Both unigrams and bigrams are employed to determine the significance of words and the dependence of short phrases in spam content [13]. Furthermore, features including the number of embedded URLs, frequency of special characters, email length, number of words, and pattern of sending domains are also considered, which are also strong indicators of spam behavior [12].

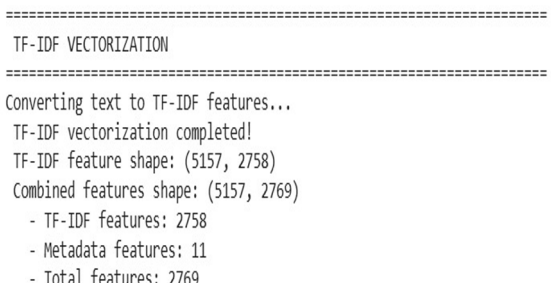

Fig. 1: TF-IDF Vectorization Result And Combined Feature Extraction For the Proposed System

All features were normalized to eliminate dimensionality issues and combined into a single feature vector to enable the significant learning of content and behavioral indicators for proper spam classification [1].

3.3. Base Classifier Models

The email spam classifier uses multiple supervised learning models as base classifiers to capture various aspects of spam emails, as well as normal emails. Each classifier gets independently trained on the same feature vector, which has been extracted using TF-IDF transformation[16], represented by

$$x = [x_1, x_2, \ldots, x_n]$$

where xi is the value of TF-IDF or the metadata feature. Independent training of the model encourages diversity in decision boundaries and is important when attempting to improve ensemble model accuracy and alleviate model bias [16].

The Naive Bayes classifier is effective for high-dimensional spam classification by labeling email through maximization of posterior probability on the assumption of conditional independence among the extracted textual and metadata features attributes.[17]

$$\hat{c} = \arg\max_{(C)} P(c) \prod_{i=1}^{n} P(x_i \mid c)$$

Logistic regression (LR) and support vector machine (SVM) classifiers improved the overall efficacy. LR involves learning linear decision boundaries with a probability output [9], whereas SVM maximizes class separation in high-dimensional spaces, guaranteeing excellent generalization capabilities on sparse text and providing a solid foundation for ensemble methods [1], [17].

3.4. Ensemble Classification Strategy

To improve the accuracy of the classification process and address the limitations of each classification model, the proposed system uses an ensemble learning technique that involves combining several base classifiers to form a powerful predictive model. By incorporating probabilistic, linear, and margin-based classifiers, such as Naïve Bayes, Logistic Regression and Support vector machines, the ensemble technique promotes diversity in the classifiers to address bias and variance issues in the generalization of spam patterns [1].

Soft voting techniques were used in this ensemble approach, where the results of each classifier were combined in the form of an average to calculate the confidence levels of the classes. Based on the confidence level of each class, the output of the classifier corresponds to the highest score class [16]. This technique addresses the diversity in the confidence levels of each model and makes it possible to combine the learning techniques of the classified models in a way that enhances the accuracy of spam detection in emails [9], [13], [15].

3.5. System Architecture

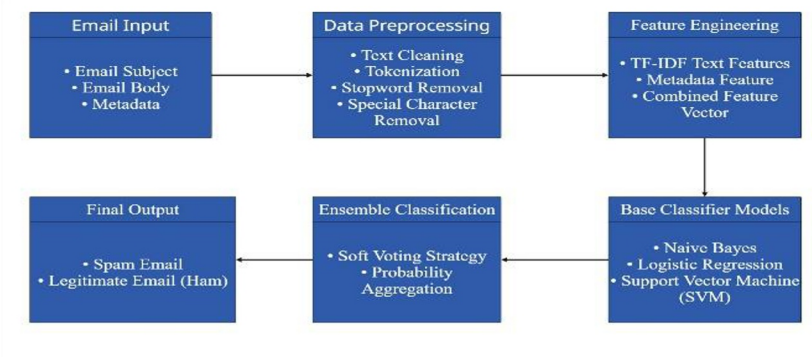

Fig 3: System Flow diagram of the Email Spam Detection Model

3.4. Results and Visualizations

The effectiveness of the suggested email spam-filtering system was measured by using the traditional measures of accuracy such as accuracy, precision, recall, and F1 -score. These metrics are popular because they are able to identify the problem of class-imbalance and reduce the rate of false positives.

It was tested on the test sample in order to determine whether the model predictions are free of bias. The results of all the classification were carefully stored to enable thorough assessment of model performance.

The results of the archival can be utilized to optimize the performance of the model and, possibly, improve the system to a powerful spam-detecting device.

4. RESULTS AND DISCUSSION

The effectiveness of the offered email spam classification system was measured using publicly available data with spam and legitimate email labels in accordance with the methods implemented in the previous research on spam classification [14, 15]. Three major areas were covered in the performance study: the effectiveness of text and metadata feature information, the performance of the classifiers, and the improvement attained using an ensemble approach. This was done to ensure the accurate classification of spam messages with minimal false positives, which is an important area in the real-world applications of email spam classification systems [11].

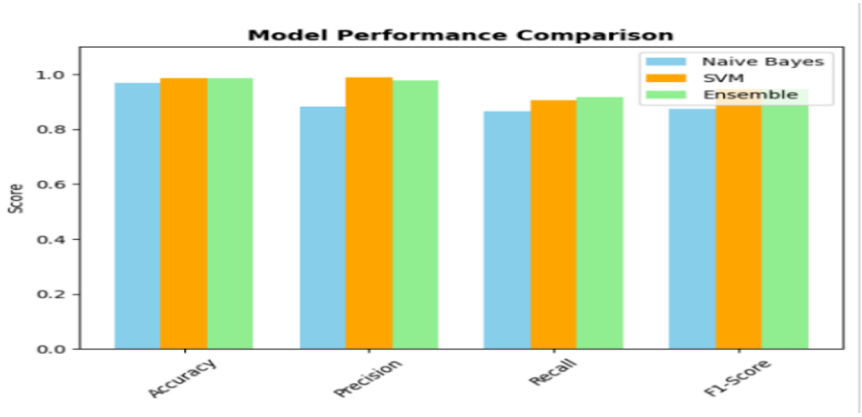

Fig.4: Comparative Performance Of Individual Classifiers And Ensemble Model

The textual features extracted using the TF-IDF method, along with other metadata variables such as the sending domain address, email length, and URL presence, were employed to train and test separate classifiers, such as naïve Bayes, logistic regression, and support vector machine [17], [20]. All classifiers were trained with stable accuracies, ensuring that the learned features were representative. Naïve Bayes performs excellently in recognizing typical spam motifs owing to its probabilistic model [17], Logistic Regression offers reliable hyperplane separation between spam and legitimate messages [9], and SVM has good generalization skills in dealing with

high-dimensional sparse text-feature data [17]

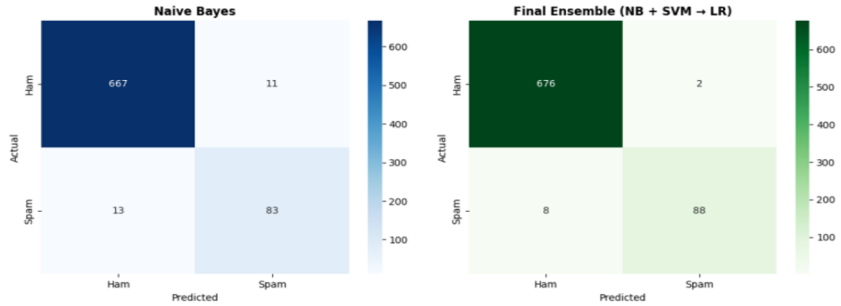

Fig.5: Confusion Matrix Comparison For Individual And Ensemble Learning

An ensemble model was built and a soft voting approach was used to combine the probability outputs of each base classifier, which was found to enhance the robustness and stability of the predictions [1], [16]. The outcome of the experimental analysis shows that the ensemble model performed better than the base classifiers in terms of accuracy, precision, recall, and the value of the F1 measure, and had a significant impact on the decline of incorrect positive instances [4], [13]. The effectiveness and robustness of the proposed system proved to be resilient to different email attributes, such as large URL density and the presence of advertisements in emails. Although the proposed system was validated on a single dataset, the outcomes were sufficient to confirm the effectiveness and applicability of the TF-IDF attributes and ensemble approach for accurate spam email detection in the real world [14], [15].

5. LIMITATIONS

Although the proposed system exhibits efficiency in classifying spam messages in an email using ensemble learning, it has some limitations. In their experimentation, the proposed system was tested on a public dataset, which may not represent all possible cases in real-world email communication traffic [14], [15]. The nature of spam can change from one domain, language, or user context to another, and it may not be easily trained on one dataset in a real-world deployment scenario.

The ensemble approach tends to be based on TF-IDF representations coupled with supportive metadata characteristics. Although TFidf has a high level of performance in terms of lexical salience, it lacks in the ability to resolve more profound semantic connections, and thus, it is unable to identify more complex types of spam, including obfuscated text, image-based spam, and context-dependent phishing attacks [7, 8]. Also, the adversaries can manipulate metadata properties, such as the domain of the sender, message size, and the number of URLs, thus affecting the detection performance [12, 15].

Additionally, the model is not built with the capability of continuous learning, which is inappropriate since spam strategies keep on changing. This kind of evolution may lead to performance deterioration through concept drift [11, 14]. Nevertheless, despite these limitations, the model is a credible basis on which further improvements should be made.

6.CONCLUSION

The current research presents a practical spam identification model based on ensemble learning including features of text extraction, feature enhancement of meta-feature, and discrete supervised classifiers. The new approach uses the TF-IDF representation of the email text, and the meta-features of sender domain, message size, and presence of URL to expand the discriminatory capability between spam and ham. Naive Bayes, a support vector machine and logistic regression classifiers are used as the basic models and their results are aggregated through a soft-voting mechanism.The effectiveness of the system is due to the ability to effectively and accurately filter spam messages by using the complementary skills of the probabilistic, linear, and margin-based models. The ensemble approach is more robust than single-model classifier in the sense that it is able to accommodate heterogeneous spam patterns and has a significantly low false-positive rate. It is empirically shown that the ensemble is superior to single models in email filtering.

7. FUTURE WORK

Future additions to the framework can include scaling to bigger and more diverse data sets, more complex representations of features, e.g. word embeddings or deep-learning-based systems, and includes adaptive learning processes to counteract concept drift and multimodal data, e.g. image-based features.

8. REFERENCES

[1] Fatima, R., Fareed, M. M. S., Ullah, S., Ahmad, G., & Mahmood, S. (2024). An optimized approach for detection and classification of spam email's using ensemble methods. Wireless Personal Communications, 139(1), 347-373.

[2] Gupta, S. M. (2007). Support vector machines based modelling of concrete strength. *Int. J. Intel. Technol*, *3*(1), 12-18.

[3] Jain, V. (2025, April). Intelligent Email Spam Detection: A Machine Learning-Based Approach. In *2025 5th International Conference on Trends in Material Science and Inventive Materials (ICTMIM)* (pp. 1574-1579). IEEE.

[4] Akeel, M., Butt, K. K., Javed, K., Tariq, M., & Yousaf, M. (2025). Email Spam Detection Using Machine Learning with Optimized Feature Engineering and Classification Techniques. *Journal of Computing & Biomedical Informatics*, *10*(01).

[5] Subramaniyaswamy, V., Kshetri, N., Ravi, L., Revathy, G., & Thillaiarasu, N. (2025). *Deep Learning and Blockchain Technology for Smart and Sustainable Cities*. Auerbach Publications.

[6] Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International journal of information management data insights*, *1*(1), 100007.

[7] Liu, X. Deciphering Spam Through AI: From Traditional Methods to Deep Learning Advancements in Email Security.

[8] Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2024). Next-generation spam filtering: Comparative fine-tuning of LLMs, NLPs, and CNN models for email spam classification. *Electronics*, *13*(11), 2034.

[9] Jayapandian, N. (2023, December). Machine learning based spam e-mail detection using logistic regression algorithm. In *2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG)* (pp. 1-6). IEEE.

[10] Keat, L. C., & Ying, T. X. Artificial Intelligence-Based Email Spam Filtering.

[11] Maurya, S. K., Singh, D., & Maurya, A. K. (2023). Deceptive opinion spam detection approaches: a literature survey. *Applied intelligence*, *53*(2), 2189-2234.

[12] Tian, Y., Yu, Y., Sun, J., & Wang, Y. (2025). From Past to Present: A Survey of Malicious URL Detection Techniques, Datasets and Code Repositories. *arXiv preprint arXiv:2504.16449*.

[13] Agarwal, R., Dhoot, A., Kant, S., Bisht, V. S., Malik, H., Ansari, M. F., ... & Hossaini, M. A. (2024). A novel approach for spam detection using natural language processing with AMALS models. *IEEE Access*, *12*, 124298-124313.

[14] Hadi, M. T., & Baawi, S. S. (2024, January). Email spam detection by machine learning approaches: a review. In *International Conference on Forthcoming Networks and Sustainability in the AIoT Era* (pp. 186-204). Cham: Springer Nature Switzerland.

[15] Tusher, E. H., Ismail, M. A., Rahman, M. A., Alenezi, A. H., & Uddin, M. (2024). Email spam: A comprehensive review of optimize detection methods, challenges, and open research problems. *IEEE Access*.

[16] Al-shanableh, N., Alzyoud, M. S., & Nashnush, E. (2024). Enhancing email spam detection through ensemble machine learning: A comprehensive evaluation of model integration and performance. *Communications of the IIMA*, *22*(1), 2.

[17] Kumar, N., & Sonowal, S. (2020, July). Email spam detection using machine learning algorithms. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 108-113). IEEE.

[18] Murthy, G. K., Nayana, C. P., Harshitha, P., & Sandhya, M. (2024, May). Email Spam Detection using Graph based Convolutional Neural Network with MALO. In *2024 Second International Conference on Data Science and Information System (ICDSIS)* (pp. 1-8). IEEE.

[19] Al Saidat, M. R., Yerima, S. Y., & Shaalan, K. (2024). Advancements of SMS spam detection: A comprehensive survey of NLP and ML techniques. *Procedia Computer Science*, *244*, 248-259.

[20] Rusland, N. F., Wahid, N., Kasim, S., & Hafit, H. (2017, August). Analysis of Naïve Bayes algorithm for email spam filtering across multiple datasets. In *IOP conference series: materials science and engineering* (Vol. 226, No. 1, p. 012091). IOP Publishing.