

An Approach to Provide Clear Findings for Identifying Unwanted Messages in Electronic Communication Systems

Chenghao Xu

*Shiley-Marcos School of Engineering, University of San Diego, San Diego, USA
chenghaoxu@sandiego.edu*

Abstract. Unwanted messages in electronic communication, such as spam and phishing attempts, continue to pose significant risks, including information theft, malware distribution, and data loss. Traditional rule-based and keyword-based filtering methods have become less effective due to the evolving tactics used by malicious actors. This study presents a comprehensive machine learning framework for detecting unwanted messages, which combines both textual features—such as term frequency-inverse document frequency (TF-IDF)—and structural features including message length, word count, frequency of special characters, and ratios of uppercase letters. The framework is evaluated using three widely adopted classification algorithms: logistic regression, linear support vector machine, and random forest, applied to a large, publicly available Kaggle dataset of electronic messages. Experimental results demonstrate that the random forest model outperforms the other methods, achieving a precision-recall score of 0.986 and an area under the ROC curve of 0.998. These findings highlight the advantage of integrating diverse engineered features with classical machine learning techniques for effective and interpretable spam detection. Further analysis of feature importance and classification errors provides additional insight into model behavior and error patterns. Future research will focus on incorporating advanced deep learning and explainability methods to further improve detection accuracy and transparency in real-world communication systems.

Keywords: Email Spam Detection, Machine Learning, Random Forest, Explainable AI, Text Classification.

1. Introduction

Electronic communication is a necessary factor in the contemporary digital age not only at the workplace, but in personal life. This is one kind of communication that offers a medium where primary emphasis on the communication of information globally is exhibited. Its mass usage also provided grounds that enables electronic communication to be a significant target of system attacks. Unwanted communication in the form of the messages is no longer restricted to advertisements that are not demanded by persons. Such messages exhibit evolution towards a form that offers considerable avenues of attacks in the process of acquiring information using false messages, spreading of malicious programs, and identity theft. The sheer magnitude of the traffic by the unwanted message does not only consume network resources and memory space but also enhances

the threat of breaching the security of the message that is highly prejudiced to the protection of the system. It has been demonstrated through previous research that attacks concerning unwanted messages endure and are constantly in the state of development. Such development sets the groundwork that renders methods that are based on searching certain words ineffective [1]. The conditions arising lead to a high demand on systems that identify undesired messages which reflect intelligence and automation. These systems need to be capable of correctly distinguishing messages that have legitimate properties and that which have malicious or unwanted properties.

The study of spam detection in email has shifted away to rule-based systems, then to the statistical, then to the Machine Learning models, and, more recently, to Deep Learning systems. Machine learning-based methods enable systems to discover patterns which differentiate spam and voluminous data to achieve a better detection. Recent research shows that deep network approaches are highest in terms of spam detection. Patterns in word sequences in messages have been discovered through Long Short-Term Memory and Gated Recurrent Unit and this leads to increased detection accuracy [2]. The in-both-way methods based on transformer structure that have been used to analyze text in an email are able to find context and meaning relationships and perform much better [3,4]. Moreover, techniques, which apply methods of biological systems, were also presented to improve the level of detection of spam by the systems which are based on Deep Learning, and the accuracy of the system cluster of messages [5].

The major limitation of these models that demonstrate high levels of predictive performance is related to their nature. A range of Deep Learning models can also be viewed as black-box models, so the functionality of the decision-making process is not easily analyzed by a person. In the research done on the security of computers, prediction accuracy is not enough. In cases where a valid message is blocked unintentionally, or in the case where malicious message is not blocked by the system, those controlling the system need to be capable of studying the motivations behind such actions. Past research has revealed that the lack of the features to explore lowers the trust people have in the system, complicates the process of finding issues, and poses a problem with adherence to the regulations that demand transparency of the decision making process [6,7].

The recent researches have paid more attention to the methods that can merge several approaches and be examined to obtain the possibility to provide high detection accuracy as well as the model transparency. Procedures of the Artificial Intelligence which can be discussed are integrated with the models of the Machine Learning of the previous work to give more possible explanations of the decisions and have a good performance [8,9]. Multiple model systems detecting spam have also been introduced to achieve a balance between accuracy, the size of computation and spam features which permit examination [10]. In line with this developing trend in the literature on security, the work presented here suggests a system based on the use of the Machine Learning process of efficient spam detection of email. Compared to the strategies that are considered closed systems mainly oriented at accuracy, the presented system involves effective feature work with the models of message group representing that will allow them to be examined to achieve not only reliable spam detection but also understand how these decisions are made. The manner presented here can be considered a reliable useful or practical context of the task of filtering email in real environments by explicitly defining the key characteristics and trends that make messages be defined as spam.

2. Methodology

2.1. Research design and problem formulation

The given study is a group assignment problem of email spam detection. Every email message is an input document and the label shows whether it is legitimate message or spam: 1 means spam and 0 means not spam. The model trains a model of email text and binary label with machine learning techniques.

Research design has a normal procedure. First, Kaggle email spam dataset is an open dataset that is imported and examined. This analysis shows the arrangement and dispersion of labels. Secondly, text processing and feature development convert raw email text into numerical forms. These shapes can be used based on machine learning models. Third, the training and evaluation of several classification approaches are performed. It takes place in the same environment to offer an equal comparison. Lastly, the model with the most effective performance is evaluated on a test set that was held out. They include a variety of measures used in assessment and the behavior that is projected is evaluated. This review analyzes the confusion matrices and summary statistics.

The training of all classifiers uses the same form of features to offer comparability between classifiers. The split of training and test data and the method of the validation is also evaluated on similar groups. The paper deals with four popularized classifiers, including Logistic Regression, Linear Support Vector Machine, Random Forest, and Gaussian Naive Bayes. These models present varied learning strategies. Some of the approaches are linear methods, marginal based classification, multiple tree learning and likelihood based probabilistic. It has this diversity which enables systematic analysis of the impacts of various assumptions on spam detection performance on the identical dataset.

2.2. Data preprocessing and feature engineering

The operation is performed over a marked email spam dataset that is found at Kaggle. This data comprises textual raw email communications and binary saltiness pointers. A number of processing steps and feature development steps are used before model training. Such steps transform the text into numerical features.

2.2.1. Text cleaning and normalization

All the emails are translated to lowercase. This will eliminate the use of capitalization of words as different tokens. Informative characters are eliminated. These are superfluous whitespaces and unnecessary punctuations. Removal reduces noise. Simultaneously, spam-related information is saved in characters. These are symbols like the numerical signs, which are set up in the form of the number sign, dollar, and pattern addressing URLs like the phrase, http. After cleaning, emails which end up being empty or almost empty are thrown away. This avoids the cases of meaningless sample inclusion in the training process.

2.2.2. Tokenization and stop-word removal

After the normalization the process tokenizes individual message words. The method eliminates typical function words that contain the, and, and of, taking a standardlist, and this molarizes dimensionality and enhances signaling in the information. It might use a process that involves

stemming or any other forms that might map variations of words to a common stem thus facilitating generalization throughout the collection of data.

2.2.3. TF-IDF vectorization

Term Frequency Inverse Document Frequency converts the tokenized messages into numeric vectors. In this approach, the weighting of words that occur more often in a specific message and occur less often in the entire set of messages, weighs down, and thus this method would be appropriate in the classification. The textual representation of the study is represented with single words and the vocabulary is limited to the most common terms in the data. The representation of every message is a sparse, high-dimensional vector of weighted frequencies of terms.

2.2.4. Construction of statistical features

Besides textual manifestation, the method also elicits a few of the basic statistical features that reflect structural attributes of messages. These aspects consist of the overall count of characters and words, the percentage of demis in the message, the amount of special characters, such as ! and \$ and the amount of patterns, which look like URLs or domains. Such characteristics express aspects of messages that can usually demonstrate undesired conduct, including excessive dotting or numerous links. After weighted summation of the term vector and statistical features characterizing structural properties, the final representation of every message is acquired.

2.2.5. Feature scaling and 5 TrainTest split

The entire dataset is divided into a training and test set in a 80/20 ratio and stratification done in such that the distribution aligns between the two subsets in individual message category. In classifiers that are sensitive to the feature scales such as Logistic Regression and Linear Support Vector machine, scaling is performed within the data processing channel. This guarantees that preprocessing steps only fit to the training data and are used throughout on the test data to avoid information leakage between the troughs and this may compromise the results.

2.3. Model training and evaluation procedure

After preprocessing and feature construction, the four classifiers are trained and assessed using the same protocol that provides similar assessment.

2.3.1. Cross-validation on the training set

A process is used to assess the training set and it splits the data into five sections. Here, four parts are trained and the remaining part is validated. This is repeated on all the parts and the performance measures are averaged to give a more stable estimate of the performance of the model on data that it has not seen during training.

In this validation process, the analysis would use various measures to determine the performance. These are the rate of accurate classifications, accuracy in identifying spam, accuracy in identifying spam and a composite measure, which is a tradeoff between preciseness and recall. These measures also offer a comparison that takes into consideration the variations in classes and makes it possible to compare the classifications of various strategies.

2.3.2. Model selection and test evaluation

The choice of models is mainly based on the combined measure of the spam class since the measure reflects the focal point of the detection of spam messages against the false detection of legitimate messages. Once the model with the highest performance has been so selected, both approaches are trained on the entire training data with the same fixed parameters, and the results compared on the distinct test data.

The measures are computed on the same to measure final performance on the test set. Furthermore, each approach has detailed tables displaying the results of classification, but the emphasis is put especially on the Random Forest approach that works overall the best. These tables give some understanding of the correct spam identification, correct legitimate identification, false spam classification and false missed classifications, which can further examine classifications errors and their difference in further practical use.

3. Experimental results and analysis

3.1. Dataset overview

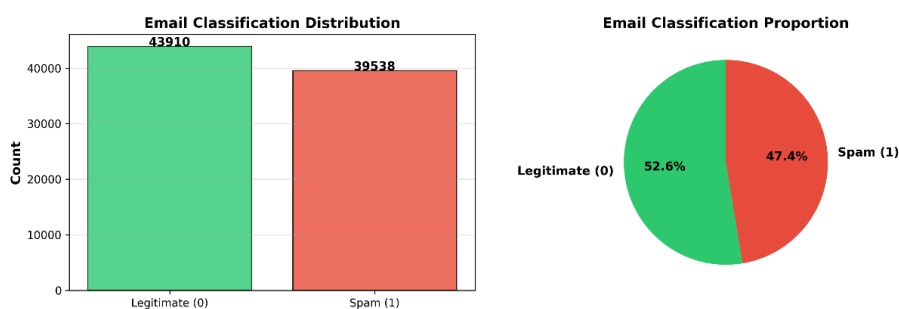


Figure 1. Distribution of email classes in the Kaggle dataset (picture credit: original)

The dataset considered by the study comprises of Kaggle set messages. The distribution of types of messages within this collection is given in figure 1. In the dataset, there is a total of 83, 448 messages of which 43, 910 are legitimate and form 52.6% and 39, 538 are spam which compose 47.4%. The relatively balanced distribution removes the risk of extreme differences in the sizes of classes, and makes normal measures, like the rate of correct classifications and the combined measure of precision-recall, to be useful in comparing approaches. The multitude and size of the data also favor training and testing which generates predictable outcomes with supervised classification strategies.

3.2. Model performance comparison

The performance of the models compared in these three techniques is found in Figure 2. Analysis reveals that all models have good results with values of accuracy showing more than ninety seven percent in the analysis. It shows that the process of identifying spam containing messages is well based on a combination of the characteristics of the term frequency approach with the features that describe other characteristics. The model that employs the method of integrating various decision structures exhibits performance that is demarcated by that of the rest of the two methods in the evaluation.

Combined decision structures model presents the most favorable value of the measure of combining precision and recalls, and the measure is the primary one used to assess in study. The measure enables evaluation that puts into consideration the rate of correct identification and the rate of complete detection of messages with spam. The two models based on linear approaches provide values of precision that seem comparable to the other model, but the values of recall in the models under the comparison seem to be low. This would mean that there is a trend in these strategies as some of these messages that have spam are not correctly categorized. The combined decision structure model indicates a high score on recall and high score on the precision model, and the trend indicates that the model can deliver a superior performance where the lost messages, which have spam, pose issues to the protection of the system.

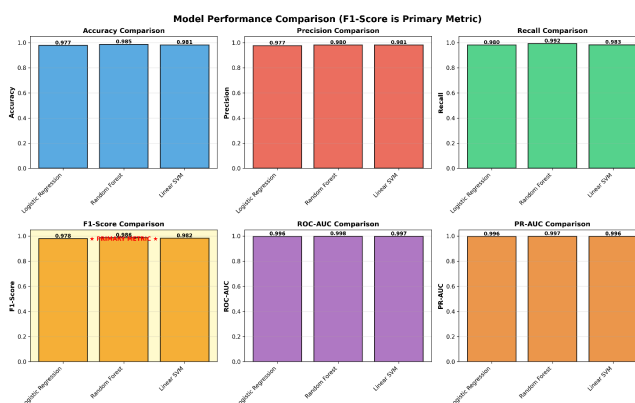


Figure 2. Comparison of model performance across classifiers (picture credit: original)

3.3. ROC, precision–recall, and threshold analysis

The curves of the models in the assessment in relation to the operating characteristic and precision versus the recall have been described in figures 3 and 4 respectively. The curves of all the models used in the analysis depict a trend of going towards the upper left region of the display and this shows that there is great segregation in the classification of messages having spam and messages that do not have spam. The combined decision structure model indicates the most extensive area under the curve of the operating characteristic and this trend is indicative of the consistency in performance at varying rates of errors in legitimacy message classification.

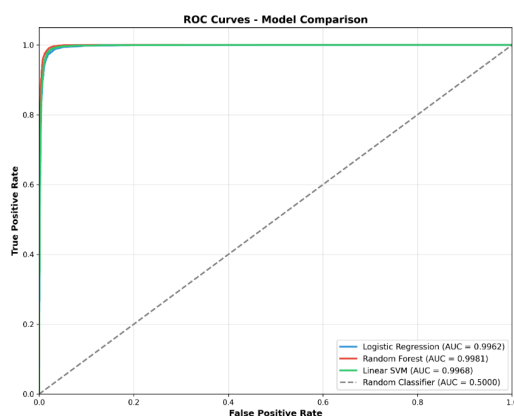


Figure 3. ROC curve of the Random Forest classifier (picture credit: original)

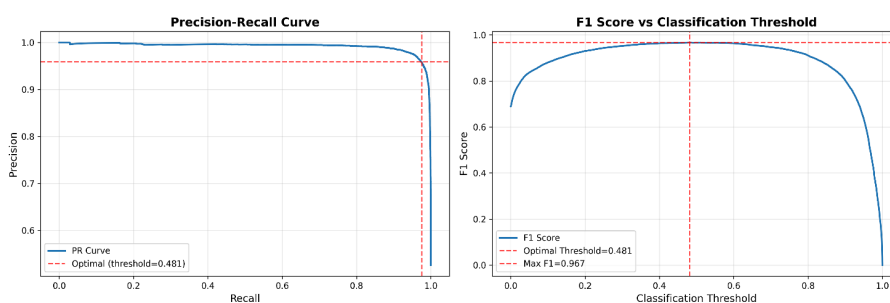


Figure 4. Precision–recall curve of the Random Forest classifier (picture credit: original)

The curve with precision against recall also gives further support to the efficiency of the model with combined decision structures, and this is reflected in the fields where recall has high values. The trend gives the spam filtering systems value since the method will minimize the spam messages that the system will not detect. The analysis which compares various points to consider making classification decisions in Figure 3 and 4 illustrates that the point which offers the best performance upon using combined decision structures takes a location of about zero point four eight, and this is where the measure of a combination of precision and recall would be high in the assessment. This indicates that the proper choice of the point on which classification decisions are made gives some enhancement in the model performance in comparison to the situations where the system operates without fine-tuning.

Figure 5 and 6 provide the matrix of confusion and analysis of the feature importance analysis of the Random Forest method of classification. According to the matrix, the model recognizes the biggest part of both types of email with few exceptions of the positive and negative results. This finding is in favor of high performance as shown by the results in numbers.

The analysis examining feature importance provides understanding of how the Random Forest model makes decisions. Features showing high importance include terms such as "attached," "http," and other elements relating to content, and also include measures using statistics that relate to the structure of messages. These features appear consistent with characteristics that spam commonly shows, such as the use of links and attached files occurring frequently. The model highlights features that show the most influence and this offers a degree that allows interpretation and supports practical use in systems that filter email in actual settings.

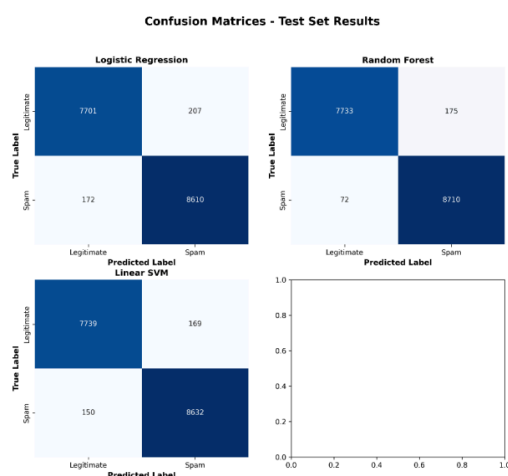


Figure 5. Confusion matrix illustrating classification patterns of the Random Forest model (picture credit: original)

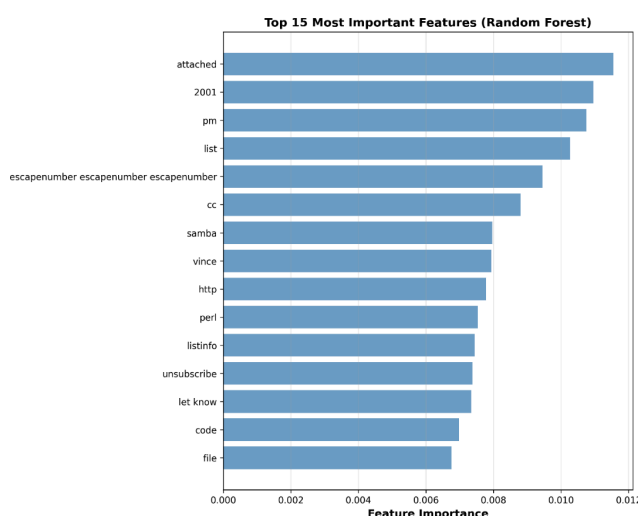


Figure 6. Feature importance derived from the Random Forest model (picture credit: original)

4. Discussion

The Test resulted proves, that classical machine learning, in combination with smart modeled features may be powerful, unfailing agents for the detection of email spam. Among the classifiers tested, Random Forest beats Logistic regression and linear Supported Machines in a large number of statistical evaluations such as F1, PR, ROC. This advantage comes from the ensemble tree-based structure of Random Forests which captures the complex interaction of reverse target-I, product features and its statistically engineered features, whereas linear methods exploit simpler decision boundaries, and thus may fail to completely exploit the complex interaction between text and statistics.

Moreover, the significant finding is the effectiveness of integrating supplementary structural features of emails with textual content. The added features like message length, special characters, and the spelling of the URL work together to improve the randomness between completely legitimate and completely different emails. Their primary intent was to demonstrate that detection of email senders basing solely on the text-based features is a matter open to salient limitations.

There are still many crucial limitations. As an example, first, to the extent that the proposed framework uses supervised learning using a relatively small static set of training samples, it may not be easy to adapt the system to attacks and new forms of spamming; second, because the systems based on the TF-IDF features do not reflect the deeper semantic relationships between words, they may not be able to respond to more subtle types of spam.

In addition, online or iterative learning can be added to the framework to be made dynamically adapt to the changing spam patterns of an actual deployment environment. Besides, new explainable AI methods like local explanations might also enhance openness and credibility.

5. Conclusion

This paper creates a systematic algorithm of the spam email detection process that is the integration of the importance of terms scores with the custom numerical characteristics of the message content. A combination of these attributes converts the unstructured email text into structured information to use in classification. A huge real-world dataset provided by Kaggle was used to evaluate the methodology to make sure that the results are applicable to practice. Three common classification

algorithms were evaluated: a probabilistic model, a linear separation model, and a tree-based predictor, all of which were tested with similar protocols. The findings determine the best combination of features and models to use and still have model interpretability and transparency in the process.

The quantitative findings also demonstrate that the Random Forest classifier has better performances on important metrics, such as F1-score, ROC-AUC, and PR-AUC. In comparison to simpler linear models, Random Forest has the ability to model complex textual and structural feature interactions, which is especially beneficial with regards to spam detection. The study contains the thorough analyses, including the performance metrics reporting and a range of other comprehensive ones: ROC curves, precision-recall plot, threshold optimization, confusion matrix, and feature importance assessment. These studies offer explanations on model decision-making and behavior and patterns of errors, thus helping to out in the practical application and interpretable behavior in email filter systems.

The results show that classical machine learning used in conjunction with effective feature engineering can compete with more complex models especially in situations where interpretability and computational efficiency are of importance. Even though the present structure is based on fixed labeled data, it forms a powerful basis in the future work. The combination of neural network models and interpretable reasoning tools as well as adaptive learning techniques can further improve the privacy of the spam detection and performance under changing conditions.

References

- [1] Nikhil, K., Sanket, S. (2020) Email spam detection using machine learning algorithms, *International Journal of Computer and Information Technology*. Available: <https://www.ijcit.com/index.php/ijcit/article/view/417/11>
- [2] Saleem, S., Islam, Z. U., Hasan, S. S. U., Akbar, H., Khan, M. F., and Ibrar, S. A. (2025) Spam email detection using long short-term memory and gated recurrent unit, *Applied Sciences*, vol. 15, no. 13, Art. no. 7407, 2025, doi: 10.3390/app15137407.
- [3] Nasreen, G., Khan, M., Younus, M., Zafar, B., and Hanif, M. K. (2024) Email spam detection by deep learning models using novel feature selection technique and BERT, *Egyptian Informatics Journal*, vol. 26, Art. no. 100473, 2024. Available: <https://www.sciencedirect.com/science/article/pii/S1110866524000367>
- [4] Guo, Y., Mustafaoglu, Z., and Koundal, D. (2022) Spam detection using bidirectional transformers and machine learning classifier algorithms, *Journal of Computational and Cognitive Engineering*, vol. 2, no. 1, pp. 5–9, doi: 10.47852/bonviewJCCE2202192.
- [5] Nicholas, N. N. and Nirmalrani, V. (2024) An enhanced mechanism for detection of spam emails by deep learning technique with bio-inspired algorithm, *e-Prime – Advances in Electrical Engineering, Electronics and Energy*, vol. 21, Art. no. 100504, doi: 10.1016/j.prime.2024.100504.
- [6] Bouke, M. A., Alramli, O. I., and Abdullah, A. (2024) XAIRF-WFP: A novel XAI-based random forest classifier for advanced email spam detection, *International Journal of Information Security*, vol. 24, no. 1, Art. no. 5, doi: 10.1007/s10207-024-00920-1.
- [7] Filali, A. et al. (2024) The role of explainable AI in mitigating spam threats, *Procedia Computer Science*. Available: <https://www.sciencedirect.com/science/article/pii/S1877050924010627>
- [8] Abutalha, M., Tasnim, F., Nazmul, N., and Tasnim, A. (2024) Explainable AI-based framework for efficient detection of spam from text using an enhanced ensemble technique, *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15596–15601. Available: <https://etasr.com/index.php/ETASR/article/view/7901/3869>
- [9] Uddin, M. A. and Mahiuddin, M. (2024) Explainable Email Spam Detection: A Transformer-based Language Modeling Approach, in *Proc. 27th Int. Conf. Computer and Information Technology (ICCIT)*, doi: 10.1109/ICCIT64611.2024.11022595.
- [10] Kadam, P. (2025) Email spam detection using hybrid model, *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*. Available: <https://www.ijraset.com/research-paper/email-spam-detection-using-hybrid-model>