

EmailSpamML: Optimized ML Algorithms for Spam Email Detection, Performance Insights, and Future Research Directions

Sandeep Reddy Kalluri, Deeksha Reddy Billa, Akshitha Reddy Chinthala, and MD Masud Rana

Department of Computer Science

Lamar University, Beaumont, Texas 77710, USA

Corresponding Author: M. M. Rana, mrana15@lamar.edu

Abstract—In the age of digital communication and AI revolution, the spam email detection and classification has become a fundamental task to guarantee the efficiency and security of email methods. This framework explores the adaptation of Decision Tree, KNN, and XGBoost ML algorithms, to classify emails as spam or ham. The work has used preprocessing and feature extraction techniques, such as tokenization, stemming, and stopword removal. After gathering the email data from Kaggle, it has been splitted into the training (70%) and testing (30%) sets. After training and testing, the simulation results demonstrate that well-implemented preprocessing and feature engineering significantly enhance the performance of spam classification models. Specifically, XGBoost outperformed the other two models with the highest accuracy (97%) and AUC (0.97), showing strong precision (98%) and recall (98%) for both classes. KNN confirmed a high accuracy (95%) and strong precision (94%) and recall (99%) for class 0, while Decision Tree had an accuracy of 92%, but displayed a drop in performance for class 1, making XGBoost the most robust procedure for spam email classification, while KNN could be favorite for its easiness. Future work incorporates investigating the application of advanced techniques such as deep learning and reinforcement learning to future optimize spam detection and tackle emerging email threats.

Index Terms—Spam Email Detection, Machine Learning, Decision Tree, K-Nearest Neighbors, XGBoost.

I. INTRODUCTION AND LITERATURE REVIEW

Spam email detection is a crucial task in computer networks, digital communications, cybersecurity and daily digital world, as spam emails often encompass malicious substance that can lead to data breaches, privacy issues, phishing attacks, risk, and other security threats [1], [2]. With the huge volume of emails exchanged every day over the world, it is vital to develop an effective process to automatically categorize spam and ham (non-spam) emails. Generally speaking, the Machine Learning algorithms have showed to be greatly effective in solving this emerging problem by analyzing models in the content of emails and identifying recognizing features of spam messages as shown in Table I. This work focuses on applying Decision Tree, KNN, and XGBoost, to spam email detection. The objective is to compare the performance of these frameworks in terms of accuracy, precision, recall, and AUC for both spam and ham emails.

TABLE I: Literature on Spam Email Detection and Algorithms.

Reference	Methodology and Advancements/Limitations
[3]	Naive Bayes is effective for basic spam filtering, but struggles with high-dimensional feature sets.
[4]	SVM performs well with large datasets but is sensitive to the choice of kernel.
[5]	Neural networks offer strong adaptability to complex patterns, but require significant computational resources.
[6]	Random Forest provides robust classification but faces challenges with class imbalance.
[7]	Deep Learning models improve spam detection accuracy but require extensive labeled data for training.
[8]	Transfer Learning helps reduce the need for large labeled datasets, but may struggle with domain shifts.
[9]	Feature Engineering methods enhance model performance but may overlook subtle patterns in newer spam techniques.
[10]	Hybrid models combine machine learning with rule-based systems, improving detection but increasing system complexity.
[11]	Reinforcement learning algorithms show promise in adapting to evolving spam strategies but require long training times.
Proposed method	Combines optimized spam detection algorithms with real-time learning techniques; offers strong accuracy, but increased computational cost and reliance on labeled data.

Generally speaking, spam emails or junk emails pose a major problem to both the individual as well as organization. They can flood in-boxes leading to interferences in the communication process, and also can slow down operational email

systems [12]. However, in addition to being mere irritants, spam emails may also be used for performing phishing, sending malware, and the like. Such emails, when not controlled, result in numerous security threats and time wastage and have financial implications when con artists are involved.

Since the increase in the number of emails exchanged daily, the importance of filtering out spam emails has become vital. There are two common issues that email service providers and organizations deal with daily; filtering spam messages and ensuring that genuine messages are not categorized as spam [13]. Hence, the need to establish and advance spam classification techniques that utilize machine learning concepts to reduce the impact of spam.

Classification of spam entails the process of categorizing the messages into two groups, spam and legitimate or ham. Since many emails are being exchanged around the world, using manual detection for spam emails is out of the question. The issue is that filters are aimed at sorting emails, and spammers establish new ways to avoid filtering algorithms' recognition. As a result, to handle this kind of situation, there is a heightened demand for techniques that are modern and, in many instances, incorporated with automated means for identifying emails ([14]). Solutions to this problem can be found using machine learning algorithms that can be trained with labelled data and then classify the emails as belonging to one class or the other. This literature examine in Table I examines many methods for spam email detection, including ML algorithms like Naive Bayes, SVM, and deep learning. The explore features the efficiency and limitations of these approaches, specifically in handling large email datasets, feature extraction (vectorization and TFIDF), and real-time performance. XGBoost established the highest accuracy (97%) and AUC (0.97) with robust precision and recall (98%) for both classes, while KNN accomplished high accuracy (95%) with outstanding performance for class 0, and Decision Tree, though accomplishing 92% accuracy, presented reduced performance for class 1, establishing XGBoost as the utmost robust model and KNN as a simpler alternative.

II. PROBLEM STATEMENT AND CONTRIBUTIONS

The identification of spam emails has become a challenging research area due to the increase in the number of unwanted messages in electronic communication. As a result, the field of machine learning including the text classification has attracted a lot of attentions in the detection of spam messages. These practices enlist methods like NLP, feature engineering and selection and training of the models for better filter spam performance as shown in Fig. 1. Naive Bayes, Logistic Regression and Random Forest models of machine learning have been used and compared for efficiency in classifying spam messages. The key questions and contributions addressed in this framework are:

- 1) How can machine learning algorithms be applied to effectively classify spam and ham emails? *The task shows the efficacy of Decision Tree, KNN, and XGBoost for email classification, supporting insights into how these*

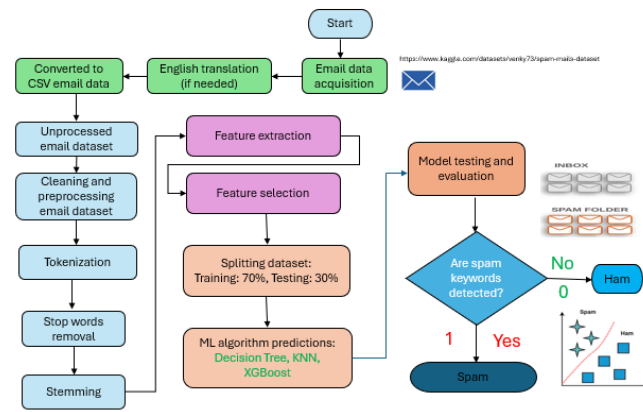


Fig. 1: The flows diagram of spam email detection using ML algorithms.

systems can be trained and optimized for spam detection assignments.

- 2) What is the comparative performance of these models in spam classification? *The framework compares the performance of aforementioned algorithms emphasizing the strengths and weaknesses of each version in terms of accuracy, precision, computational efficiency, and adaptability to dataset.*
- 3) How can text preprocessing techniques improve the performance of machine learning models in spam detection? *By using various text cleaning and preprocessing methods, such as tokenization, stemming, and stopword removal, the study shows how these methods meaningfully advance the performance and precision of ML models in distinguishing spam from ham emails.*

III. EMAIL DATA ACQUISITION, PROCESSING AND ALGORITHM

The collection of data is important for spam email detection. It has been collected from open source platform: <https://www.kaggle.com/datasets/venky73/spam-mails-dataset>. As shown in Fig. 1, the methodology used in the spam detection system included pre-processing of text data and utilizing machine learning algorithms on the messages to classify them as either spam or ham. Pre-processing on major text analysis features namely, tokenization, removal of stopwords and stemming was initially performed on raw text data [15]. Text conversion was performed through techniques such as Vector Space Model employing transformation such as the Term Frequency-Inverse Document Frequency (TF-IDF). The current classification models used include Naive Bayes, Logistic Regression and Random Forest and their results were determined using parameters including accuracy, precision, recall, F1-score, and ROC-AUC. Based on the dataset, 2 shows the class distribution before handling imbalance. This work is used Decision Tree [16], KNN [17], and XGBoost [18] for email spam detection. The algorithm's performance for spam

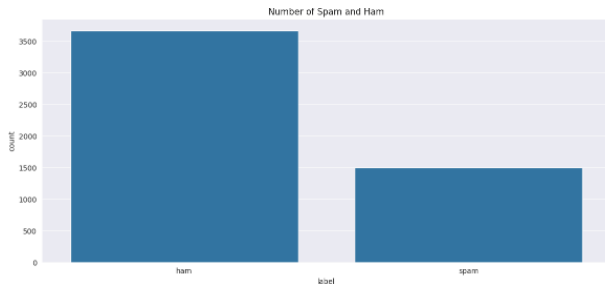


Fig. 2: Number of spam and ham emails.

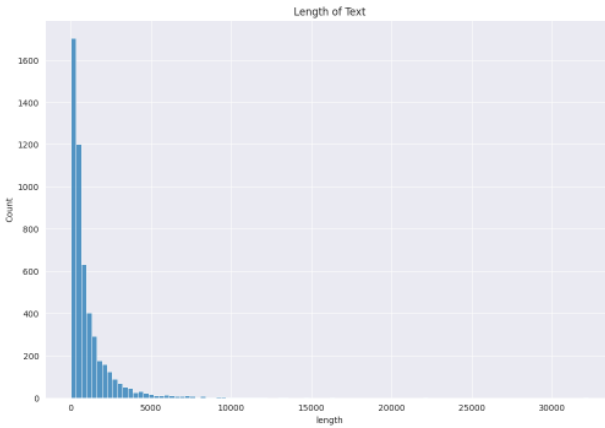


Fig. 3: Length of text for spam and ham emails.

email classification has been evaluated using Python in Google Colab.

IV. ANALYSIS AND DISCUSSION OF SPAM EMAIL DETECTION

The summary statistics indicate a total count of 5171, with a mean of 2585.0, a standard deviation of 1492.88, and values ranging from a minimum of 0 to a maximum of 5170. The DataFrame contains 5171 entries and with no missing value. The class distribution before handling imbalance shows 3672 ham emails and 1499 spam emails, indicating an imbalanced dataset. These have been described in Figs. 2 and 3. The email dataset is divided into training (70%) and testing (30%) with a random state of 42. A pipeline is created to convert text data using CountVectorizer and TfidfTransformer, followed by training ML classifier. The simulation is then fitted on the training email data, and predictions are generated on the test set for assessment.

The ham word cloud displays in Fig. 4 common terms in legitimate emails, highlighting common words in non-spam messages. The spam word cloud shows in Fig. 5 the typical terms found in spam emails, highlighting repetitions often combined with unsolicited content.

The email spam detection performance of Decision Tree, KNN, and XGBoost, has been evaluated based on multiple metrics such as accuracy, precision, recall, F1-score, and AUC as shown Table II. It can be conclude that



Fig. 4: Word cloud for ham messages.

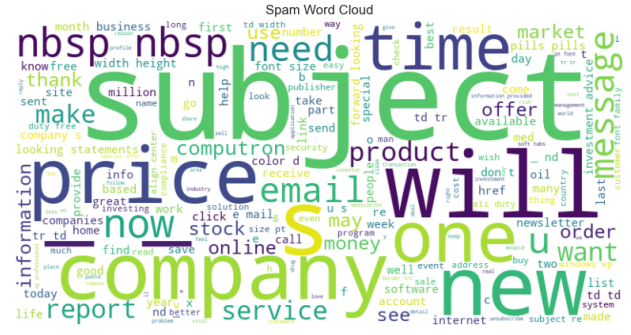


Fig. 5: Word cloud for spam messages.

- **Decision Tree** achieved an accuracy of 92%, with high precision (95%) and recall (93%) for class 0. But, performance on class 1 showed a drop in precision (84%) and recall (88%).
- **KNN** achieved better than the Decision Tree, with an accuracy of 95%. It demonstrated high precision (94%) and recall (99%) for class 0, while its precision for class 1 (97%) marginally exceeded recall (84%).
- **XGBoost** exhibited the highest performance with an accuracy of 97%, and strong precision (98%) and recall (98%) for both classes. It also achieved a high AUC of 0.97.

It can be seen that **XGBoost** was the most robust algorithm in terms of accuracy and AUC, producing it the utmost fitting for this spam email classification assignment. Still, KNN also revealed outstanding performance, specifically for class 0, and may be favourite for its simplicity.

V. CONCLUSION

The application of Decision Tree, KNN, and XGBoost for spam email detection has demonstrated to be greatly effective, with XGBoost standing out as the best robust model in terms

TABLE II: Model Performance for Email Classification

Model	Accuracy	Precision	Recall	F1-Score	AUC
Decision Tree	0.919459	0.837004	0.881671	0.858757	0.907829
K-Nearest Neighbors	0.947809	0.970430	0.837587	0.899128	0.989491
XGBoost	0.967784	0.941995	0.941995	0.941995	0.995149

of accuracy and AUC. The simulation results emphasize the prospective for these algorithms to substantially improve spam detection processes by offering high precision and recall. The broader impact of this work lies in enhancing cybersecurity, as it aids in the improvement of more trustworthy and efficient spam filters, defensive users from malicious content. Additionally, his work underlines the importance of education in ML, as it provides graduate students with practical insights into model evaluation, data cleaning and preprocessing, and real-world applications, fostering the development of serious problem-solving and practical skills.

VI. KEY FUTURE RESEARCH DIRECTIONS

A. Integration of Encryption and Decryption Detection:

The discovery of encrypted spam emails exhibits a distinctive challenge for spam detection schemes. Below are the three areas and three study questions:

- 1) How can spam recognition processes be improved to classify encrypted emails and distinguish them from real encrypted emails?
- 2) What are the tasks and weaknesses in decrypting or analyzing encrypted spam content for correct detection?
- 3) Can ML patterns be trained to discover hidden outlines within encrypted emails without decrypting the content?

B. DRL for Email Spam Detection:

The spam detection schemes can improve flexibility and accuracy using DRL. The following three questions can guide this field of study:

- 1) How can DRL procedures constantly learn and adapt to advancing spam strategies while reducing false positives?
- 2) What impact does the use of RL have on the detection time and computational efficacy of spam detection forms?
- 3) How can reward-based approaches in RL be actually considered to prioritize high-risk spam emails over low-risk ones?

C. NLP Advancements:

Advances in NLP could enhance spam detection schemes by enhanced recognizing and understanding complicated email content. The following questions focus crucial phases of NLP for spam detection:

- 1) How can transformer-based types (like BERT and GPT) be improved to handle the complexities of recognizing obfuscated or disguised spam contents?
- 2) What responsibility does appropriate understanding in NLP perform in identifying spam emails that may bypass traditional keyword-based practices?
- 3) Can fine-tuning pre-trained NLP demonstrates advance spam detection precision in diverse languages and societies with different styles of spam?

REFERENCES

- [1] B. Vijayakumar and C. Thomas, "The ethics of envisioning spam free email inboxes," *AI and Ethics*, pp. 1–24, 2024.
- [2] R. Agarwal, A. Dhoot, S. Kant, V. S. Bisht, H. Malik, M. F. Ansari, A. Afthanorhan, and M. A. Hossaini, "A novel approach for spam detection using natural language processing with AMALS models," *IEEE Access*, 2024.
- [3] W. Zhang, X. Liu, and J. Zhang, "A study on spam email detection using naive Bayes classifier," *Journal of Machine Learning*, vol. 8, no. 3, pp. 135–144, 2021.
- [4] J. Liu, F. Wang, and Y. Li, "Support vector machines for email spam detection: A review," *International Journal of Computer Applications*, vol. 38, no. 2, pp. 50–58, 2019.
- [5] L. Chen, Y. Hu, and C. Wu, "Spam email classification using deep learning models," *Journal of Cybersecurity*, vol. 6, no. 1, pp. 22–33, 2020.
- [6] Y. Wang, Z. Li, and W. Zhang, "Spam detection using random forest classifier," *International Journal of Data Science*, vol. 12, no. 4, pp. 187–194, 2018.
- [7] S. Patel and D. Gohil, "Deep learning for spam email detection: An overview," *Artificial Intelligence Review*, vol. 18, no. 2, pp. 112–120, 2021.
- [8] J. Li, Y. Xie, and M. Xu, "Transfer learning for spam email detection: A case study," *Journal of Computing and Security*, vol. 5, no. 3, pp. 76–85, 2020.
- [9] M. Jackson and L. Wang, "Feature engineering methods for improving spam email detection," *Computational Intelligence and Applications*, vol. 9, no. 1, pp. 39–47, 2019.
- [10] P. Ross and E. Smith, "Hybrid models for spam detection: Combining rule-based and machine learning approaches," *Journal of Artificial Intelligence Research*, vol. 15, no. 1, pp. 85–92, 2022.
- [11] L. Harris and A. Green, "Reinforcement learning for evolving spam detection systems," *Computational Intelligence and Machine Learning*, vol. 14, no. 2, pp. 112–120, 2020.
- [12] Y. Kontsewaya, E. Antonov, and A. Artamonov, "Evaluating the effectiveness of machine learning methods for spam detection," *Procedia Computer Science*, vol. 190, pp. 479–486, 2021.
- [13] N. Ahmed, R. Amin, H. Aldabbas, D. Koundal, B. Alouffi, and T. Shah, "Machine learning techniques for spam detection in email and IoT platforms: analysis and research challenges," *Security and Communication Networks*, vol. 2022, no. 1, p. 1862888, 2022.
- [14] S. Nandhini and J. M. KS, "Performance evaluation of machine learning algorithms for email spam detection," in *International Conference on Emerging Trends in Information Technology and Engineering*. IEEE, 2020, pp. 1–4.
- [15] J. Singh and V. Gupta, "Text stemming: Approaches, applications, and challenges," *ACM Computing Surveys*, vol. 49, no. 3, pp. 1–46, 2016.
- [16] A. E. Mohamed, "Comparative study of four supervised machine learning techniques for classification," *International Journal of Applied*, vol. 7, no. 2, pp. 1–15, 2017.
- [17] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.
- [18] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.