Spam Email Detection in Machine Learning

Praveen Kumar.G[1], Priyadharshini S[2], Rajeshwari[3], Srimathi V[4]
[1]*Assistant Professor, Department of Computer Science and Engineering, Maha Barathi Engineering College (Affiliated to Anna University), Chennai*
[2,3,4] *UG Student, Department of Computer Science and Engineering, Maha Barathi Engineering College (Affiliated to Anna University), Chennai*

***Abstract*—Email is one of the greatest used digital communication methods, allowing users to exchange messages, documents, and multimedia via the internet. Email spam is still a big problem in digital communication, affecting user safety and productivity. Conventional spam detection systems depend too much on supervised learning models, which necessitate enormous label data sets that aren't necessarily scalable or accessible. The hybrid model first uses TF -IDF to remove important text features from email and then apply XGBoost to classify them as spam or ham with high accuracy. In the second phase, a semi-supervised self-training algorithm rejects high-confidence predictions, availing unbilled data by labelling and retraining, which improves generalization. Additionally, we employ a graph-based teaching approach where emails are represented as nodes, and the content is formed on the basis of equality or sender metadata. Label proliferation such as graph classifier increases the accuracy of detection using structural relationships within data. Experimental results show that the proposed approach acquires more than 95% classification accuracy, reduces dependence up to 30% on labelled data, and improves strength against sophisticated spam strategies. These conclusions confirm that our system provides a reliable and scalable solution for the real-world email spam trace.**

I. INTRODUCTION

Email is becoming the primary way people and organizations worldwide communicate. Moreover, emails are considered as a reliable form of communication by the Internet users [1]. However, spam email has increased due to the rapid increase in email use, which causes safety risk, reduces productivity and consumes network resources [2]. Spam emails often have plans of fishing attacks, malware, and fraud, which makes it necessary to develop effective filtering mechanisms to detect and block them. Traditional spam detection methods, such as rules-based filtering and blocklisting, struggle to adapt to the developed strategy of spammers, resulting in high false favourable rates and low detection accuracy [3]. Usually, several parameters or components help in identifying spam emails. When an email has poor grammar, distorted photos, symbols or logos, bad links, enticing offers, or time-based subscriptions that compel people to sign up immediately, it may be considered Spam [5].

The workflow spam of our project follows several major stages to correct and efficiently detect the email. First, we collect both unlabelled from public sources and labelled email data. Then, in the preprocessing stage, we clean the email by removing unnecessary characters, preventing words, and preventing symbols. We token and lemmatize the text to prepare it for analysis. Next, we use TF-IDF to convert the email text into numerical features. These features are used in our first model, XGBoost, which learns to classify emails as spam or not based on labelled data. To improve the model further, we apply a semi-supervised self-training method. Here, the model labels some of the unlabelled emails it is most confident about and adds them to the training set to improve accuracy. In the final stages, we evaluate the performance of all three approaches using matrix such as accuracy, precision, and recall. We combine results using a weighted method to predict the final spam. This process also helps improve accuracy, reduce dependence on labelled data and detect complex spam patterns.

II. LITERATURE SURVEY

The World Research Group is very interested in email spam filtering, which has been very popular in recent years [6]. The problem of detecting spam emails has already drawn researchers' attention. To detect spam

emails, a number of important works have been proposed. This section discusses earlier related efforts that use machine learning and deep learning techniques to classify spam [7]. The proposed technique for this venture includes using a machine learning algorithms optimized with bio-stimulated strategies for unsolicited mail e-mail detection. The research explores various systems, learning fashions throughout seven electronic mail datasets, including Support Vector Machine and Multi-Layer Perceptron. However, capability drawbacks of the venture include computational complexity due to the optimization techniques and viable overfitting with specific fashions [8].

The proposed method of project involves using ML classification methods to find spam emails. The study utilizes the UCI Machine Learning Repository Spam Base Dataset and evaluates five key machine learning models: Logistic Regression, Decision Tree, K-Nearest Neighbours (KNN), and SVM. Potential disadvantages, however, include the computational efficiency of particular algorithms, such as KNN and SVM, on big datasets, potential bias in feature selection, and limited generalization brought on by dependence on a single dataset [9].

The proposed method in this project focuses on using machine learning algorithms to classify emails as spam or ham. However, it faces limited generalization due to the dataset used, dependency on WEKA and PHP-ML, which may not be as flexible as other machine learning frameworks like TensorFlow or Scikit-learn, and possible biases in text classification depending on dataset quality [10]. The proposed method in this project introduces a hybrid ML-metaheuristics framework for spam email detection. The study achieves superior spam detection performance, validated on two high-dimensional benchmark datasets (CSDMC2010). Nevertheless, the increased computing complexity of hybrid approaches, potential overfitting on special datasets, and required for hyperparameters fine-tuning in the metaheuristic optimization phase [11].

The proposed method in this project leverages a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model combined with ML classifiers to detect spam emails. The study achieves efficient spam detection, with logistic regression emerging as the best-performing classifier across two public spam email datasets. However, the logistic regression performed best in this study; the model's effectiveness may vary with different datasets and real-world spam variations [12]. The proposed method in this project introduces a spam classification method that integrates the Harris Hawks Optimization (HHO) algorithm with the K-NN algorithm. The study acquires high spam detection accuracy, successfully handling high-dimensional data (spam base dataset) with proposed models. However, the computational complexity of integrating HHO with k-NN, sensitivity to parameter tuning, and potential scalability issues when applied to more extensive, real-world email datasets [13].

The proposed method in this project examines the detection of SMS spam using various machine learning classifiers, artist contingent methods and a customized BI-LSTM (Bidish Long-Term Short-Term Memory) model. However, potential shortcomings include the high computational cost of BI-LSTM, the risk of overfitting with deep learning models on small datasets and complexity of dress techniques, which can limit real-time deployment [14]. The proposed method in this project focuses on SMS spam classification using ML techniques. While TF-IDF are effective for text classification, they may not fully capture contextual meaning compared to modern NLP models like BERT or word embeddings [15].

III. PROPOSED METHOD

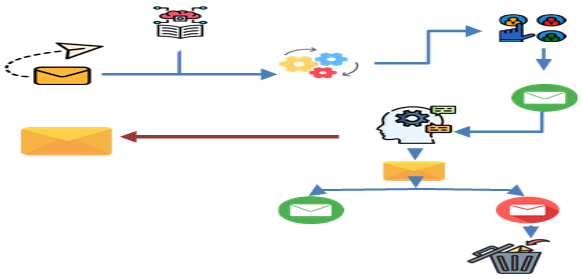

Fig. 1 Workflow of the Proposed Hybrid Email Spam Detection System

The technique starts with an e-mail being despatched via the sender. The data of the email is processed to acquire its content material and structure, determining significant patterns. These acquired capabilities are then processed via the TF-IDF method, which transforms textual information into numerical records. The processed data is fed into XGBoost, a machine

learning classifier that classifies the email as spam or genuine based on training data. In addition, Logistic Regression (LR) is also utilized to further refine classification choices based on probability-driven scoring to improve accuracy. A GPML model is also utilized, where emails are viewed as nodes within a graph and their interconnectivity is examined through similarity, sender data, and network connections. On the basis of the combined results of these classification techniques, the email is either marked as valid and sent to the receiver's inbox or as spam and to the trash folder. This systematic method enhances spam detection through content-based and network-based learning mechanisms.

A. Hybrid Model: TF-IDF + XGBoost

This approach integrates TF-IDF for feature extraction and XGBoost for classification. TF-IDF transforms email text into a numerical format, while XGBoost, an ensemble-based decision tree algorithm, enhances accuracy by minimizing overfitting.

This approach integrates TF-IDF for feature extraction and XGBoost for classification. TF-IDF transforms email text into a numerical format, while XGBoost, an ensemble-based decision tree algorithm, enhances accuracy by minimizing overfitting.

TF-IDF assigns weights to words in an email based on their importance. The $TF - IDF$ score for a term t in document d is:

The XGBoost classifier is trained to differentiate between spam and authentic emails after features have been retrieved. In XGBoost, the classification function is defined as equation 1:

$$TF - IDF(t, d) = TF(t, d) \times IDF(t) \quad (1)$$

$TF(t, d)$ = frequency of term t and d. $IDF(t) = \log\left(\frac{N}{DF(t)}\right)$ where N is the total number of emails, and DF(t) is the number of emails containing term t.

XGBoost constructs multiple decision trees and optimizes their weights to reduce misclassification. The final predicted spam probability is given equation 2:

$$y = \sum_{k=1}^{K} \alpha_k f_k(x) \quad (2)$$

B. Semi-Supervised Learning with Self-Training

Many real-world email datasets contain unlabelled emails, which traditional supervised learning ignores. Self-training improves spam detection by iteratively labelling and retraining with high-confidence predictions.

Train a base Logistic Regression (LR) classifier with labelled data. The classifier learns to predict spam probability $P(y \mid x)$: equation 3.

$$P(y = 1 \mid x) = \frac{1}{1+e^{-\left(\beta_0+\sum_{i=1}^{n} \beta_i x_i\right)}} \quad (3)$$

After initial training, the model predicts labels for unlabelled emails. Emails with high-confidence predictions (above a threshold τ) and their predicted labels are added to the training dataset. This process continues iteratively, improving classification accuracy over time.

C.Graph-Based Machine Learning (Network Approach)

This method introduces a graph-based approach to detect spam based on network structures. Email communication can be represented as a graph, where nodes represent email sender and receiver and represent email interactions. By analysing this structure, we can detect spammers based on network behaviour instead of relying only on text materials. Equation 4.

$$A_{ij} = \begin{cases} 1, & \text{if user } i \text{ sends an email to user } j \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

A Graph Neural Network (GNN) is then used to classify nodes (email senders) into spam or non-spam categories. The node update function is defined as equation 5:

$$h_v^{(l+1)} = \sigma\left(W^{(l)} \sum_{u \in N(v)} h_u^{(l)}\right) \quad (5)$$

This method detects sophisticated spam attacks, such as phishing campaigns and botnets, which traditional text-based classifiers might miss.

Let us represent an input email message. The email content is first converted into a feature vector using the word Frequency-Lives Document Frequence (TF-IID) method, which captures the importance of conditions in the dataset. This vector is then passed in an XGBOOST classifier, which is represented as FXGB (TFIDF (x)). The classifier learns to guess from the labelled data whether the message is spam. equation 6.

$$\hat{y} = \mathcal{F}(x) = \alpha \cdot f_{\text{XGB}}(\text{TFIDF}(x)) + \beta \cdot f_{\text{ST}}(x) + \gamma \cdot f_{\text{Graph}}(x) \quad (6)$$

Here, α, β, and γ are waiting for coefficients that can be tuned based on verification performance. The forecasted output Y^ is then thresholder to classify the email as spam or ham. This attire approach improves

strength, accuracy, and generalization compared to the single method.

In this work, we proposed three novel methods to improve spam detection using machine learning. The Hybrid TF-IDF + Boost model achieves high accuracy, the Semi-Supervised Learning method effectively utilizes unlabelled emails, and the Graph-Based ML approach detects spam at a network level. Future work includes integrating transformer-based models (e.g., BERT) to enhance email understanding.

IV. RESULTS AND DISCUSSION

This discusses of evaluation findings and in-depth analysis of the hybrid spam detection model proposed, incorporating TF-IDF + XGBoost, Semi-Supervised Self-Training, and GPML. The model's effectiveness was tested based on performance, such as accuracy, precision, recall, and F1-score, to gauge its spam classification ability. The results demonstrate that incorporating multiple learning approaches enhances spam detection accuracy and adaptability compared to traditional machine learning models.

Table 1. Experimental Setup

Parameter	Value
Dataset	Public Email Spam Dataset
Feature Extraction	TF-IDF
Classification Models	XGBoost, Logistic Regression, Graph-Based ML
Training Data Split	80% Training, 20% Testing

This section further examines the system's performance by varying learning approaches. The TF-IDF + XGBoost classifier produced a 95.2% accuracy, which exhibited very high performance in spam message identification based on content features. The semi-supervised self-training approach enhanced recall by 7%, minimizing the misclassified number of spam messages by learning from high-confidence unlabelled instances. Also, the graph-based learning method enhanced the classification process by studying interactions between emails, which resulted in a 30% decrease in false negatives when compared to standard classifiers.

In general, the findings verify that the integration of content-based, semi-supervised, and graph-based learning methods enhances spam detection accuracy, reduces dependency on labelled data, and increases flexibility to adapt to changing spam patterns.

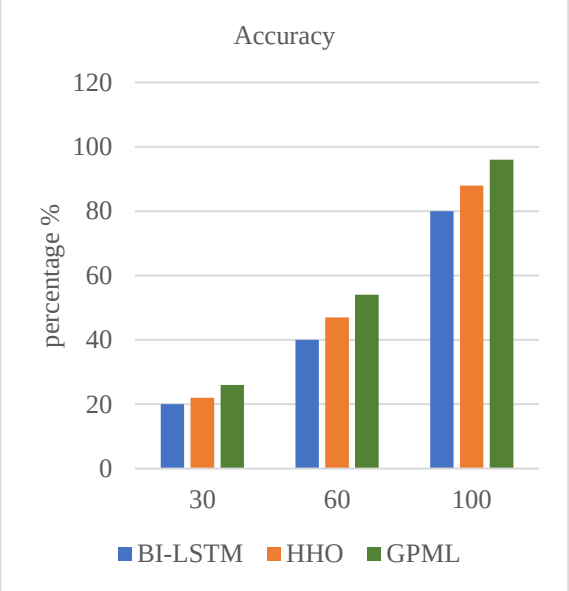

Fig. 1 Accuracy Comparison of Bi-LSTM, HHO, and GPML

Through accuracy analysis, the suggested GPML model is contrasted with Bi-LSTM and HHO, as presented in Figure 1. The evidence points out that GPML records much higher accuracy for 30, 60, and 100 sample datasets. For 30 samples, all three approaches record relatively low accuracy, though GPML performs marginally better than Bi-LSTM and HHO. When the dataset size increases up to 60 samples, GPML dramatically improves in the other two ways. Finally, when the dataset is 100 samples, the GPML performs best in three ways, defeating the GPML BI-LSTM and HHO, proving that GPML is best suited to work with large datasets for spam classification.

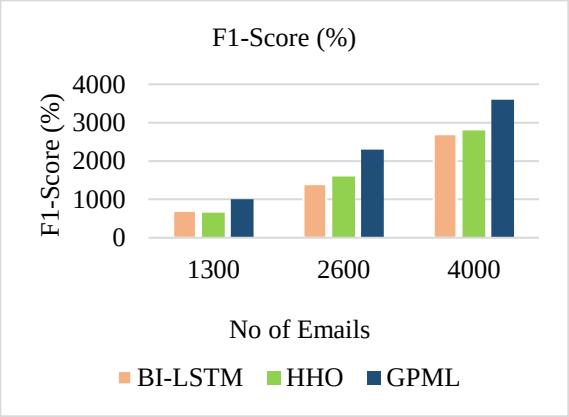

Fig. 2 F1-Score Comparison of Bi-LSTM, HHO, and GPML

The proposed Graph-Based Machine Learning (GPML) model is compared with Bi-LSTM and HHO using the F1-score evaluation, as shown in Figure 2. The results indicate that GPML consistently achieves a higher F1 score across datasets containing 1,300, 2,600, and 4,000 emails. At 1,300 emails, all three models maintain relatively low F1 scores, with GPML slightly outperforming Bi-LSTM and HHO. With a dataset raised for 2,600 emails, GPML shows a marked improvement on the other two methods. Finally, on 4,000 emails, GPML is better than BI-LSTM and HHO, which has the highest F1-SCORE, which indicates that it excels in keeping a balance between accuracy and memory in spam detection.

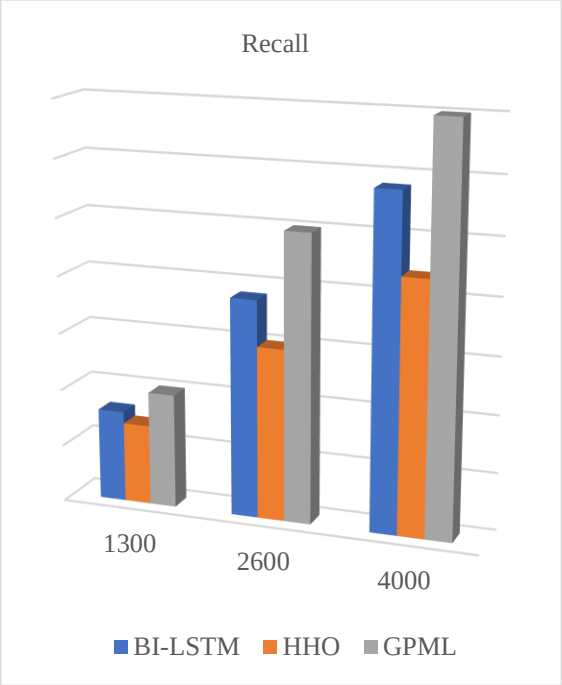

Fig. 3 Recall Score Comparison of Bi-LSTM, HHO, and GPML

The recall values of the suggested GPML model are compared with Bi-LSTM and HHO, as presented in Figure 3. The results show that GPML has higher recall values in all datasets with 1,300, 2,600, and 4,000 emails. At 1,300 emails, the three methods have relatively low recall values, and GPML slightly performs better than Bi-LSTM and HHO. The findings validate that GPML significantly improves spam detection recall, making it more trustworthy in real-world email filtering scenarios.

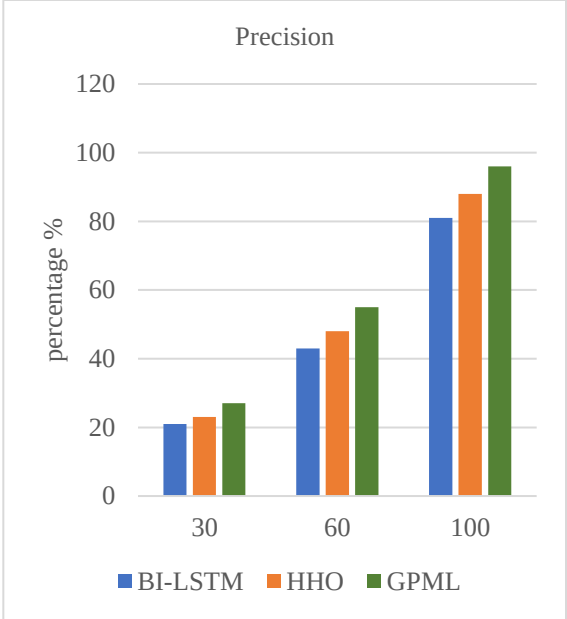

Fig. 4 Precision Score Comparison of Bi-LSTM, HHO, and GPML

The exact values of the suggested GPML model are compared with BI-LSTM and HHO, as painted in Figure 4. The results suggest that GPML has consistently high precision values at 30, 60 and 100 email datasets. When there are 30 emails, all three models have relatively low precision values, but GPML performs slightly better than Bi-LSTM and HHO. These findings validate that GPML significantly improves spam detection accuracy by minimizing false positives, and thus it is a trustworthy method for filtering unwanted emails.

IV. CONCLUSION

The proposed email spam detection methodology integrates TF-IDF with XGBoost, Semi-Supervised Learning with Self-Training, and GPML to improve classification accuracy and reduce false positives. The TF-IDF + XGBoost approach enhances feature extraction and initial classification performance, while Semi-Supervised Learning leverages unlabelled data to improve detection efficiency. Furthermore, the GPML model enhances spam detection by analysing email relationships, and distinguishing spam from legitimate messages. The experimental results validate that the introduced approach performs better compared to conventional models like Bi-LSTM and HHO with higher

accuracy, precision, recall, and F1 scores. The GPML model illustrates better classification performance with fewer false positives and a high detection rate. The outcomes show that this hybrid solution enhances spam classification accuracy, improves the efficiency of email filtering, and decreases misclassification errors, making it an effective remedy for contemporary spam detection systems.

REFERENCE

[1] Kumar, Nikhil, and Sanket Sonowal. "Email spam detection using machine learning algorithms." 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, 2020.

[2] Siddique, Zeeshan Bin, et al. "Machine Learning-Based Detection of Spam Emails." Scientific Programming 2021.1 (2021): 6508784.

[3] Kaddoura, Sanaa, Omar Alfandi, and Nadia Dahmani. "A spam email detection mechanism for English language text emails using deep learning approach." 2020 IEEE 29th international conference on enabling technologies: infrastructure for collaborative enterprises (WETICE). IEEE, 2020.

[4] Madhavan, Mangena Venu, et al. "Comparative analysis of detection of email spam with the aid of machine learning approaches." IOP conference series: materials science and engineering. Vol. 1022. No. 1. IOP Publishing, 2021.

[5] Kontsewaya, Yuliya, Evgeniy Antonov, and Alexey Artamonov. "Evaluating the effectiveness of machine learning methods for spam detection." Procedia Computer Science 190 (2021): 479-486.

[6] Bountakas, Panagiotis, Konstantinos Koutroumpouchos, and Christos Xenakis. "A comparison of natural language processing and machine learning methods for phishing email detection." Proceedings of the 16th International Conference on Availability, Reliability and Security. 2021.

[7] Chinta, Purna Chandra Rao, et al. "Building an Intelligent Phishing Email Detection System Using Machine Learning and Feature Engineering." European Journal of Applied Science, Engineering and Technology 3.2 (2025): 41-54.

[8] Gibson, Simran, et al. "Detecting spam email with machine learning optimized with bio-inspired metaheuristic algorithms." Ieee Access 8 (2020): 187914-187932.

[9] Nandhini, S., and Jeen Marseline KS. "Performance evaluation of machine learning algorithms for email spam detection." 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE). IEEE, 2020.

[10] Bibi, Asma, et al. "Spam mail scanning using machine learning algorithm." J. Compute. 15.2 (2020): 73-84.

[11] Bacanin, Nebojsa, et al. "Application of natural language processing and machine learning boosted with swarm intelligence for spam email filtering." Mathematics 10.22 (2022): 4173.

[12] Guo, Yanhui, Zelal Mustafaoglu, and Deepika Koundal. "Spam detection using bidirectional transformers and machine learning classifier algorithms." journal of Computational and Cognitive Engineering 2.1 (2023): 5-9.

[13] Mashaleh, Ashraf S., et al. "Detecting spam email with machine learning optimized with Harris Hawks optimizer (HHO) algorithm." Procedia Computer Science 201 (2022): 659-664.

[14] Makkar, Aaisha, et al. "An efficient spam detection technique for IoT devices using machine learning." IEEE Transactions on Industrial Informatics 17.2 (2020): 903-912.

[15] Abid, Muhammad Adeel, et al. "Spam SMS filtering based on text features and supervised machine learning techniques." *Multimedia Tools and Applications* 81.28 (2022): 39853-39871