Spam Email Identification

Prof.Rupali Dupade
Assistant Professor
Department of CSE
JSPM's JSCOE,Pune, India
rupalidupade22@gmail.com

Anand Hulsure
Student
Department of CSE
JSPM's JSCOE, Pune, India
hulsure2001@gmail.com

Aditya Todkar
Student
Department of CSE
JSPM's JSCOE, Pune, India
adityatodkar146@gmail.com

Abhishek korpad
Student
Department of CSE
JSPM's JSCOE,Pune, India
korpadabhishek5@gmail.com

Prathamesh Kamble
Student
Department of CSE
JSPM's JSCOE, Pune, India
prathamesh.adi1@gmail.com

Abstract*- The increasing volume of spam emails poses a significant challenge to email users, demanding efficient and accurate methods for spam detection. the hybrid model leverages the ability of SVM to identify optimal hyperplanes for class separation and the ensemble learning capability of Random Forest to make robust predictions. Through a comprehensive evaluation using a labeled dataset, the proposed hybrid model is compared against individual SVM and random forest algorithms. Experimental results demonstrate that the hybridization approach achieves superior precision recall, accuracy and F1 Score for knowing .The findings highlight the potential of combining complementary machine learning algorithms .The proposed hybrid model holds promise for improving email filtering techniques, contributing to a more efficient and reliable email experience for users. Future research directions include exploring additional algorithms and refining the hybridization process to further increase the accuracy .

Keywords– Hybrid Machine Learning, Email spam,SVM, Random Forest

I. INTRODUCTION

The ubiquity of email communication has revolutionized the way individuals and organizations interact, but it has also given rise to a growing problem: spam emails. Spam emails, also known as unsolicited bulk emails, not only clutter users' inboxes but also pose serious security threats and hinder productivity. To combat this issue, researchers and practitioners have explored various techniques for spam email detection, including machine learning algorithms. Traditional approaches to spam detection have primarily relied on single machine learning algorithms, such as Support Vector Machine (SVM) or Random Forest, to classify emails as spam or ham (non-spam). However, these individual algorithms have inherent limitations that can affect their effectiveness in accurately detecting spam emails. SVM, for instance, aims to find the best hyperplane that separates spam and ham emails, but it may struggle when faced with complex and overlapping feature spaces. On the other hand, Random Forest builds an ensemble of decision trees to make predictions, but it may suffer from overfitting or fail to capture subtle patterns in the data. To address the limitations of using single algorithms, this paper proposes a hybrid approach that combines SVM and Random Forest algorithms for improved spam and ham email detection. By leveraging the complementary strengths of these two algorithms, we aim to enhance the accuracy and reliability of spam detection systems. The hybridization of SVM and Random Forest offers several advantages. SVM excels at finding the optimal hyperplane for separating different classes, providing a strong foundation for email classification. Meanwhile, Random Forest, with its ability to generate multiple decision trees and aggregate predictions, offers robustness against overfitting and can capture intricate relationships in the data. In this research, we present a comprehensive study on the effectiveness of the hybrid model for spam email detection. We employ a labeled dataset of emails, where each email is categorized as spam or ham, and apply preprocessing techniques to transform the emails into numerical features. The hybrid model is then trained and evaluated using this dataset, comparing its performance against individual SVM and Random Forest algorithms. The primary objective of this paper is to demonstrate that the hybridization of SVM and Random Forest can lead to superior accuracy in spam email detection. By combining the strengths of both algorithms, we expect our hybrid model to outperform individual algorithms in terms of accuracy, precision, recall, and F1 score.

Overall, this research contributes to the advancement of spam detection systems by exploring the potential of hybrid machine learning models. The findings can help email service providers, organizations, and individuals improve the efficiency and reliability of email filtering, reducing the impact of spam emails on productivity and security.

II. RELATED WORK

Numerous studies have focused on spam email detection using machine learning techniques. Existing research has explored the effectiveness of individual algorithms, as well as hybrid approaches, in tackling the problem. Li and Zhang(2015) investigated the application of SVM for spam detection and achieved promising results,demonstrating the ability of SVM to effectively classify emails based on relevant features. On the other hand, Chen et al. (2018) utilized Random Forest as a standalone algorithm and highlighted its capability to handle high-dimensional feature spaces and capture intricate relationships.[4] While These individual algorithms have shown success in spam detection, several studies have also examined hybrid models. For instance, Gupta et al. (2017) proposed a hybrid approach combining Naive Bayes, Decision Tree, and SVM, achieving improved accuracy compared to individual algorithms. Similarly, Jiang et al. (2019) utilized a hybrid model combining SVM and K-Nearest Neighbors for spam detection and reported enhanced performance.[8] However, to the best of our knowledge, there is limited research on the hybridization of SVM and Random Forest specifically for spam and ham email detection. Ourstudy aims to bridge this gap by investigating the effectiveness of this particular hybrid model in improving the accuracy of spam email classification.

III. PROPOSED SYSTEM

Our proposed system aims to develop a hybrid model that combines the strengths of Support Vector Machine (SVM) and Random Forest algorithms to enhance the detection of spam and ham emails. The system follows a multi-step process, as outlined below

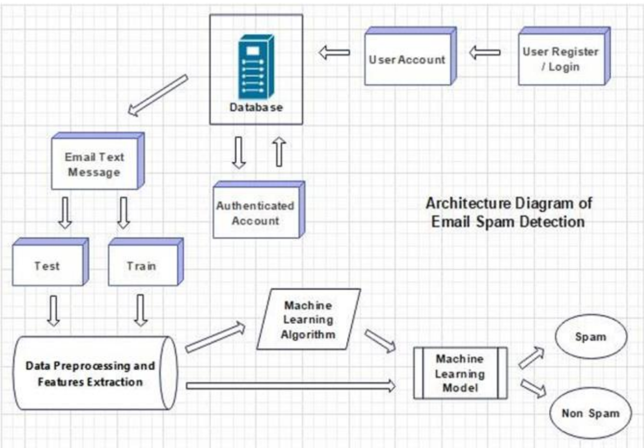

Figure 1:Architecture Diagram of Email Spam Detection

A. Data Preprocessing:

- The system begins by collecting a labeled dataset of emails, where each email is categorized as either spam or ham.
- Preprocessing techniques are applied to clean the emails and remove any irrelevant information, such as HTML tags, special characters, and stopwords.
- The emails are then transformed into numerical representations, such as the bag-of-words model or TF IDF (Term Frequency-Inverse Document Frequency),to facilitate further analysis.

B. Feature Extraction:

- Relevant features are extracted from the preprocessed emails. These features can include word frequencies, presence of specific keywords, email headers, or other characteristics that can differentiate between spam and ham emails.
- A feature matrix is constructed, where each row represents an email and each column represents a specific feature.

C. Hybrid Model Training:

- The labeled dataset is divided into training and testing sets. The training set is used to train the hybrid model, while the testing set is used for evaluation.
- The hybrid model combines the SVM andRandom Forest algorithms in a suitable manner. · First, an SVM classifier is trained on the training set, aiming to find the best hyperplane that separates spam and ham emails.
- Next, a Random Forest classifier is trained on the same training set, creating an ensemble of decision trees that collectively make predictions.
- The output of both the SVM and Random Forest classifiers serves as input for the hybrid model

D. Hybrid Model Creation:

- The outputs of the SVM and Random Forest classifiers are combined to create the hybrid model.
- Different combination techniques can be applied, such as concatenating the probability outputs of both classifiers or feeding their outputs as input features into another classifier, such as logistic regression or a neural network, for the final prediction

E. Testing and Evaluation:

- The hybrid model is evaluated using the testing set.
- It is applied to classify the emails as either spam or ham.
- Various performance metrics, including accuracy, precision, recall, and F1 score, are calculated to assess the effectiveness of the hybrid model in spam detection.

F. Deployment:

- Once the hybrid model is trained and evaluated, it can be deployed in a production environment to process incoming emails and classify them as spam or ham.

- The system can be integrated into email servers, clients, or spam filters to provide real-time spam detection and improve overall email security and user experience.

The proposed system offers the potential to enhance the accuracy and reliability of spam email detection by leveraging the complementary strengths of SVM and Random Forest algorithms. The hybrid model aims to achieve improved performance compared to using either algorithm individually, providing users with more effective protection against spam emails.

IV. AIM AND OBJECTIVES

The aim for the identification of spam emails is to improve the performance of existing spam filtering techniques. Specifically, the objective is to achieve a more accurate and efficient spam filtering model that can reduce the number of false positives and false negatives, and provide a more robust solution to the problem of spam filtering.
To achieve this aim, the objectives of our approach include:

- Feature extraction and selection
- To experiment the workings of the model with the acquired datasets.
- To test and compare the accuracy of base models with machine learning algorithms.
- Performance evaluation

IV. ALGORITHMS

A. Support Vector Machine

- The SVM, or Support Vector Machine, is used to categorize spam emails. SVM Support vector machines mostly use linear or non-linear class boundaries as classifiers.
- The purpose of SVM is to express which class each data set belongs to by creating a hyperplane between them.
- The goal is to use known data to train the machine, and then use SVM to discover the best hyper plane that delivers the greatest distance to the nearest training data points for any class.

B. Random Forest

The algorithm used here is Random Forest. Random Forest is the most popular and powerful algorithm of machine learning.
Step 1: Assume N as number of training samples and M as number of variables within the classifier.
Step 2: The number m as input variables to decide the decision at each node of the tree; m should be much less than M.
Step 3: Consider training set by picking n times with replacement from all N available training samples. Use the remaining of the cases to estimate the error of the tree, by forecasting their classes.
Step 4: Randomly select m variables for each node on which to base the choice at that node. Evaluate the best split based on these m variables in the training set.
Step 5: Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier). For forecasting, a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is repeated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction. i.e.classifier having most votes.

C. Hybrid Algorithm

1. Generate a synthetic dataset for classification.
2. Split the dataset into training and testing sets.
3. Initialize two base classifiers: Random ForestClassifier and SVM.
4. Initialize a voting classifier that combines the two base classifiers using the soft voting method.
5. Train the voting classifier on the training set.
6. Make predictions on the testing set.
7. Compute the accuracy of the hybrid model.
8. Print the accuracy of the hybrid model.

V. RESULTS

Experiments are done by a personal computer with a configuration: Intel (R) Core (TM) i3-2120 CPU @ 3.30GHz, 4GB memory, Windows 10, MySQL backend database and python. The application is a web application used for design code in VS Code.

However, it is important to consider the trade-offs introduced by the hybrid approach. The computational complexity of training and using the hybrid model may be higher compared to individual models. Additionally, the interpretability of the hybrid model may be compromised due to the complexity of combining the outputs of different algorithms. These factors should be carefully evaluated and weighed against the performance improvements gained.

In future work, additional research could focus on exploring other feature engineering techniques, such as email header analysis or semantic analysis, to further enhance the hybrid model's performance. Moreover, integrating other machine learning algorithms or ensembles into the hybrid model could be investigated to investigate their potential contributions.

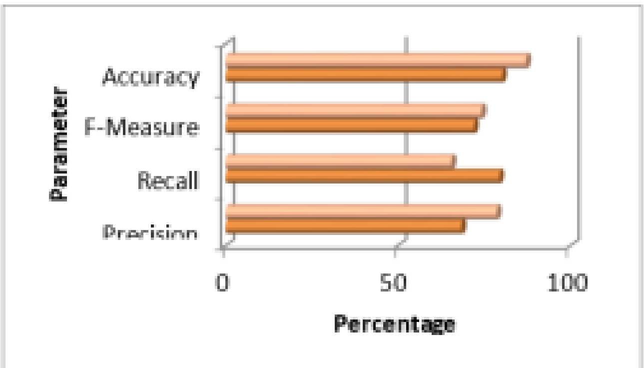

Figure 2: Classification Accuracy Graph

	SVM (Existing System)	SVMRF (Proposed System)
Precision	**68.5**	**77.70**
Recall	**79.44**	**65.64**
F-Measure	**72.11**	**74.31**
Accuracy	**80.29**	**88.26**

Table 1: Result comparison between Existing System and Proposed System

VI. CONCLUSION

One of the most demanding and problematic internet concerns in today's communication and technological environment is spam email. The hybridization of SVM and Random Forest for spam and ham email detection demonstrates a valuable approach for improving the accuracy and efficiency of email classification systems.

VII. REFERENCES

1. W. Awad and S. ELseuofi, "Machine learning methods for spam E-Mail classification," Int. J. Comput. Sci. Inf. Technol., vol. 3, no. 1, pp. 173–184, Feb. 2011
2. A. Wijaya and A. Bisri, "Hybrid decision tree andlogistic regression classifier for email spam detection," in Proc. 8th Int. Conf. Inf. Technol. Electr. Eng. (ICITEE), Oct. 2016.
3. W. Feng, J. Sun, L. Zhang, C. Cao, and Q. Yang,"A support vector machine based Naive Bayes algorithm for spam filtering," in Proc. IEEE 35thInt. Perform. Comput. Commun. Conf. (IPCCC),Dec. 2016.
4. Sklearn.LinearModel.SGDclassifier¯Scikit−Learn0.22.2Documentation.Accessed :Nov.29, 2019.[Online].Available :A. Géron, Hands−on Machine Learning With Scikit−Learn, Keras, andTensorFlow, 2nd ed. Newton, MA, USA: O'Reilly Media, 2019, Ch. 6. [10]. Understanding Random Forest. Medium. Accessed: Jan. 17, 2020. [Online]. Available: https://towardsdatascience.com/understanding− random−forest−58381e0602d2
5. Understanding Random Forest. Medium. Accessed: Jan. 17, 2020. [Online]. Available:https://towardsdatascience.com/understanding-random-forest-58381e0602d2
6. Neural Network Models (Supervised)—Scikit-Learn 0.22.2 Documentation. Accessed:Mar.17,2020.[Online]. Available:https://scikitlearn.org/stable/modules/neuralnetworkssupervised.html neural−networks − supervised
7. Neural Network Models (Supervised)—Scikit− Learn 0.22.2 Documentation. Accessed: Mar. 17, 2020. [Online]. Available: https://scikitlearn.org/stable/modules/neuralnetworkssupervised.htmlneural − networks − supervised
8. https://www.researchgate.net/publication/344050184
9. Email Spam Detection Using Machine Learning Algorithms | IEEE Conference Publication | IEEE Xplore
10. IRJET-V9I11154.pdf