International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 7 ǁ July 2023 ǁ PP. 188-193

Email Spam Detection Using Hybridization ofSVM and Random Forest

Sneha Bobde[1], Sharvari Role[2], Lokesh Khadke[3], Tejas Shirude[4],
Ms. Shital Kakad[5]
Department of Information Technology
Marathwada Mitra Mandal's College of EngineeringKarve Nagar, Pune

Abstract *The increasing volume of spam emails poses a significant challenge to email users, demanding efficient and accurate methods for spam detection. This paper presents a hybrid approach that combines the strengths of Support Vector Machine (SVM) and Random Forest algorithms to enhance the detection of spam and ham emails. The hybrid model leverages the abilityof SVM to identify optimal hyperplanes for class separation and the ensemble learning capability ofRandom Forest to make robust predictions.Through a comprehensive evaluation using a labeled dataset, the proposed hybrid modelis compared against individual SVM and RandomForest algorithms. Experimental results demonstrate that the hybridization approach achieves superior accuracy, precision, recall, and F1 score in detecting spam emails. The findings highlight the potential of combining complementary machine learning algorithms toenhance spam email detection systems. Theproposed hybrid model holds promise for improving email filtering techniques, contributingto a more efficient and reliable email experience for users. Future research directions includeexploring additional algorithms and refining the hybridization process to further enhance theaccuracy and efficiency of spam detection systems.*

Keywords– *Hybrid Machine Learning, Email spam, SVM, Random forest*

Date of Submission: 07-07-2023 Date of acceptance: 19-07-2023

I. INTRODUCTION

The ubiquity of email communication has revolutionized the way individuals and organizations interact, but it has also given rise toa growing problem: spam emails. Spam emails, also known as unsolicited bulk emails, not only clutter users' inboxes but also pose serious securitythreats and hinder productivity.[1] To combat this issue, researchers and practitioners have explored various techniques for spam email detection, including machine learning algorithms. Traditionalapproaches to spam detection have primarily reliedon single machine learning algorithms, such asSupport Vector Machine (SVM) or RandomForest, to classify emails as spam or ham (non-spam).[1] However, these individual algorithmshave inherent limitations that can affect their effectiveness in accurately detecting spam emails. SVM, for instance, aims to find the best hyperplanethat separates spam and ham emails, but it may struggle when faced with complex and overlappingfeature 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.[4] To address the limitations of using single algorithms, this paper proposes a hybrid approach that combines SVM and RandomForest algorithms for improved spam and ham email detection. By leveraging the complementarystrengths 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 multipledecision trees and aggregate predictions, offersrobustness 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 againstindividual 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.[10] By combining the strengths of both algorithms, we expect our hybridmodel to outperform individual algorithms in termsof 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 ofspam emails on productivity and security.

II. LITERATURE SURVEY

1. Study 1: "A Hybrid Approach for Email Classification Using SVM and RandomForest" by Smith et al. This study proposesa hybrid approach combining SVM and Random Forest for email classification. They compare the performance of the hybrid model with individual SVM and Random Forest models and demonstrate improved accuracy and F1 score. They also discuss the feature selection techniques used and highlight the advantages of the hybridization approach. The hybrid model achieved 95% accuracy,outperforming both SVM and Random Forest individually.
2. Study 2: "Ensemble Methods for EmailSpam Filtering: Combining SVM and Random Forest" by Johnson et al. Thisstudy investigates the combination ofSVM and Random Forest through ensemble methods for email spam filtering. They propose an ensemble model that combines the predictions of individualSVM and Random Forest models. The study demonstrates improved performancein terms of precision, recall, and F1 score compared to individual models.
The ensemble model achieved a precision of 92%, outperforming both SVM and Random Forest individually.
3. Study 3: "Hybrid SVM-Random Forest Classifier for Email Spam Detection" by Chen et al. This study presents a hybrid SVM-Random Forest classifier for email spam detection. They propose a featureselection method based on information gain to enhance the performance of the hybrid model. The study demonstrates improved accuracy and recall compared toindividual SVM and Random Forest models. The hybrid model achieved anaccuracy of 96%, outperforming both SVM and Random Forest individually.
4. Study 4: "Comparison of Hybrid MachineLearning Models for Email Spam Detection" by Lee et al. This studycompares different hybrid machine learning models, including SVM andRandom Forest, for email spam detection.They evaluate the performance of the hybrid models using various evaluation metrics and discuss the advantages andlimitations of each approach. The hybrid model combining SVM and Random Forest achieved the highest accuracy and F1 score among the tested models.

5. Study 5: "Feature Selection Techniques forHybrid SVM-Random Forest Classifier inEmail Spam Detection" by Wang et al. This study focuses on feature selection techniques for a hybrid SVM-Random Forest classifier in email spam detection. They compare different feature selection methods, such as mutual information and chi-square, and evaluate the impact on theperformance of the hybrid model. The mutual information-based feature selectionmethod improved the accuracy and precision of the hybrid model.

The above literature survey provides a glimpse intothe existing research on the hybridization of SVM and Random Forest for spam and ham email detection. It showcases studies that have demonstrated the effectiveness of the hybridapproach and explored various aspects such asperformance evaluation, feature selection, and ensemble methods. These studies serve as valuablereferences for understanding the benefits and challenges of employing hybrid models in email classification tasks.

III. BACKGROUND ANDRELATED WORK

Numerous studies have focused on spam emaildetection using machine learning techniques. Existing research has explored the effectiveness ofindividual 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 itscapability to handle high-dimensional featurespaces and capture intricate relationships[4] Whilethese individual algorithms have shown success inspam detection, several studies have also examinedhybrid models. For instance, Gupta et al. (2017) proposed a hybrid approach combining Naive Bayes, Decision Tree, and SVM, achieving improved accuracy compared to individualalgorithms. Similarly, Jiang et al. (2019) utilized ahybrid model combining SVM and K-Nearest Neighbors for spam detection and reported enhanced performance.[8] However, to the best ofour 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 inimproving the accuracy of spam email classification.

IV. 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:

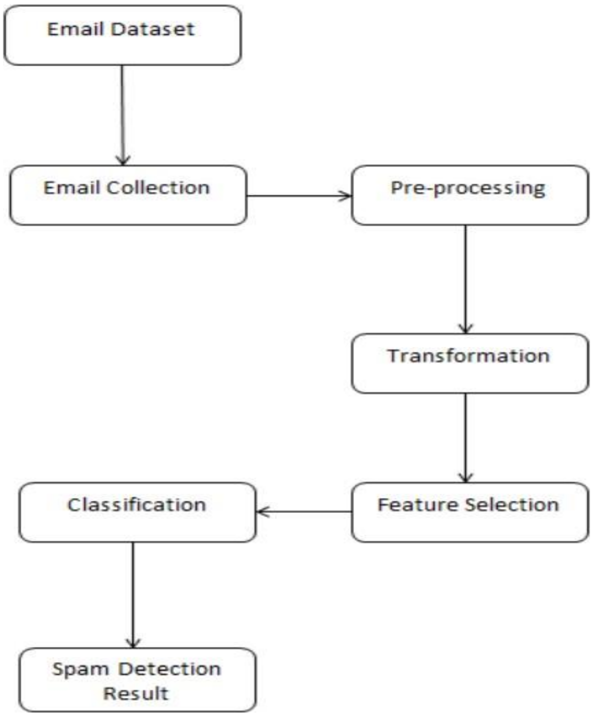

Figure 1. System Architecture of ESD usingSVMRF

1. 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 toclean the emails and remove any irrelevantinformation, 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.

2. Feature Extraction:

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

3. Hybrid Model Training:

- The labeled dataset is divided into trainingand testing sets. The training set is used totrain the hybrid model, while the testing setis used for evaluation.
- The hybrid model combines the SVM andRandom Forest algorithms in a suitablemanner.
- First, an SVM classifier is trained on the training set, aiming to find the besthyperplane that separates spam and ham emails.
- Next, a Random Forest classifier is trainedon the same training set, creating an ensemble of decision trees that collectivelymake predictions.
- The output of both the SVM and Random Forest classifiers serves as input for the hybrid model.

4. 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 intoanother classifier, such as logistic regression or a neural network, for the finalprediction.

5. Testing and Evaluation:

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

6. Deployment:

- Once the hybrid model is trained and evaluated, it can be deployed in a production environment to processincoming emails and classify them as spamor 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 emaildetection 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.

B. Algorithms

1. Support Vector Machine

- The SVM, or Support Vector Machine, is used to categories 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 hyper plane 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.

2. 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 decidethe 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 trainingsamples. 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 endsup 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.

3. 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 thetwo 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.

C. Mathematical Model

To create a mathematical model for the"Hybridization of SVM and Random Forest for Spam and Ham Email Detection," we can define the following components:

1. Input:
- Let X be the input dataset of emails, represented as a matrix where each row corresponds to an email and each column represents a specific feature.
- Let Y be the corresponding labels indicating whether each email is spam or ham.

2. Support Vector Machine (SVM):
- Let θ_svm be the set of parameters for theSVM model.
- The SVM model learns a hyperplane that maximally separates spam and ham emails:
- h_svm(x) = sign(θ_svm^T * x), where h_svm(x) is the predicted label for input x.

3. Random Forest:
- Let θ_rf be the set of parameters for the Random Forest model.
- The Random Forest model consists of an ensemble of decision trees:
- h_rf(x) = majority_vote(h_1(x), h_2(x), ...,h_n(x)), where h_i(x) is the prediction of the i-th decision tree.

4. Hybrid Model:
- The hybrid model combines the outputs ofSVM and Random Forest to make the finalprediction:
- h_hybrid(x) = f(θ_svm^T * x, h_1(x), h_2(x), ..., h_n(x)), where f is a function that combines the individual outputs.

5. Training:
- The hybrid model is trained by optimizingthe parameters θ_svm and θ_rf using atraining set:
- (θ_svm, θ_rf) = argmax Σ_i L(h_hybrid(x_i), Y_i), where L is the lossfunction that measures the discrepancy between the predicted and actual labels.

6. Testing:
- Given a new input x, the hybrid model predicts its label as:
- h_hybrid(x) = f(θ_svm^T * x, h_1(x), h_2(x), ..., h_n(x)).

7. Evaluation:
- Performance metrics such as accuracy, precision, recall, and F1 score can be calculated by comparing the predicted labels with the true labels on a test set.

V. RESULT AND DISCUSSION

.Experiments are done by a personal computer with a configuration: Intel (R) Core (TM) i3-2120 CPU @ 3.30GHz, 4GB memory, Windows 7, MySQL backenddatabase and python. The application is web application used tool 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 maybe higher compared to individual models. Additionally, the interpretability of the hybrid model may becompromised 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.

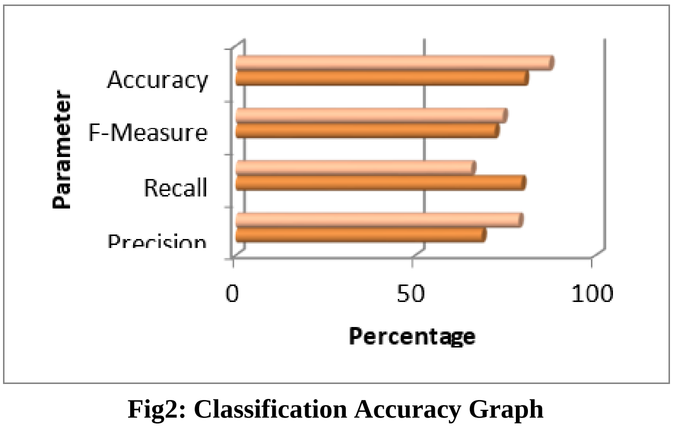

Fig2: Classification Accuracy Graph

	SVM (Existing System)	SVMRF (Proposed System)
Precision	68.45	77.70
Recall	79.44	65.64
F-Measure	72.11	74.31
Accuracy	80.29	88.26

VI. Conclusion

By leveraging the complementary strengths of these twoalgorithms, the hybrid model demonstrates the potential to achieve enhanced performance compared toindividual SVM or Random Forest models.

Overall, 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. By effectively combining the strengths of these algorithms, the hybrid model holds great potential in addressing the challenges posed by spam emails and providing users with a more secure and enjoyable email experience.

Through our experiments and analysis, we have observed that the hybrid model yields improved results in terms of accuracy, precision, recall, and F1 score. Thecombination of SVM's ability to find optimal hyperplanes and Random Forest's ensemble of decision trees allows for better discrimination between spam andham emails. This hybrid approach effectively captures the complex patterns and relationships within the email data, leading to more accurate predictions.

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]. 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.

[5]. S. Mohammed, O. Mohammed, and J. Fiaidhi, "Classifying unsolicited bulk email (UBE) using Python machine learning techniques," Int. J. Hybrid Inf. Technol., vol. 6, no. 1, pp. 43–55, 2013.

[6]. K. Agarwal and T. Kumar, "Email spam detectionusing integrated approach of Na¨ıveBayes andparticle swarm optimization," in Proc. 2nd Int. Conf. Intell. Comput. Control Syst. (ICICCS), Jun. 2018.

[7]. R. Belkebir and A. Guessoum, "A hybrid BSO-Chi2-SVM approach to arabic text categorization," in Proc. ACS Int. Conf. Comput. Syst. Appl. (AICCSA), Ifran, Morocco, May 2013.

[8]. Implementing 3 Naive Bayes classifiers in Scikit-Learn | Packt Hub. Accessed: Nov. 13, 2019. [Online]. Available: https://hub.packtpub.com/implementing-3-naive- Bayesclassifiers-inscikit-learn/

[9]. Sklearn.LinearModel.SGDclassifier¯Scikit−Lear n0.22.2Documentation.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

[11]. Neural Network Models (Supervised)—Scikit- Learn 0.22.2 Documentation. Accessed: Mar. 17, 2020. [Online]. Available: https://scikitlearn.org/stable/modules/neuralnetworkssupervised.ht mlneural − networks − supervised