JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Email Spam Detection Using Machine Learning Algorithms

[1]Nikita Sunita Mokal, [2]Saee Suresh Dalvi, [3]Prof. Sumedh Pundkar

Department of Computer Science and Technology, Usha Mittal Institute of Technology, SNDT Women's University Juhu-Tara Road, Sir Vitthaldas Vidyavihar, Santacruz(W), Mumbai, Maharashtra 400049

Abstract: Email spam continues to be a pervasive issue, posing threats to user privacy, productivity, and security. Machine learning (ML) techniques have emerged as effective tools for automated spam detection, offering the potential to adapt to evolving spamming tactics. This study proposes a hybrid machine learning approach for email spam detection, leveraging the strengths of both Random Forest (RF) and Gradient Boosting (GB) algorithms. The hybrid model aims to enhance classification accuracy and robustness by combining the ensemble learning capabilities of RF with the boosting power of GB. The proposed framework involves feature extraction from email datasets, preprocessing, and model training using the hybrid RF-GB algorithm. Evaluation metrics such as accuracy, precision, recall, and F1-score are employed to assess the performance of the hybrid model against individual RF and GB classifiers. Experimental results demonstrate the effectiveness of the hybrid approach in achieving superior spam detection performance, thus offering a promising solution for combating email spam in real-world applications.

Keywords **- Machine Learning algorithm, Random Forest (RF), Gradient Boosting (GBT), Hybrid approach.**

I. INTRODUCTION

In the digital age, email remains a primary communication tool, yet it also serves as a conduit for spam, phishing attempts, and malicious content. As the sophistication of email spamming techniques evolves, traditional rule-based filters struggle to keep pace with the dynamic nature of spam. Consequently, there arises a pressing need for more adaptive and accurate spam detection mechanisms. Machine learning (ML) techniques have emerged as potent tools in this venture, leveraging the power of algorithms to recognize patterns and classify emails effectively.

This paper introduces a novel hybrid approach for email spam detection, integrating the strengths of two prominent ML algorithms: Random Forest (RF) and Gradient Boosting (GBT). Random Forest excels in handling large feature spaces and mitigating overfitting, while Gradient Boosting exhibits superior performance in boosting weak learners to form strong classifiers. By combining these algorithms, we aim to harness their complementary strengths to enhance the accuracy and robustness of spam detection.

The hybrid framework proposed in this study involves several key components. Firstly, it contains preprocessing of email data to extract relevant features while mitigating noise and irrelevant information. Subsequently, the preprocessed data is fed into the hybrid RF-GB model for training, where both algorithms work collaboratively to learn the underlying patterns indicative of spam and non-spam emails. The trained model is then evaluated using established metrics to assess its performance in terms of accuracy, precision, recall, and F1-score.

By adopting a hybrid approach, we seek to address some of the inherent challenges encountered in email spam detection, such as feature selection and model generalization. The combination of RF and GB offers a robust solution that is capable of handling diverse types of spam while minimizing false positives and false negatives. Furthermore, the hybrid model is expected to exhibit greater resilience to adversarial attacks and novel spamming tactics, thus enhancing the overall security of email communication. Ultimately, this research contributes to the advancement of email security measures, ensuring a safer and more reliable communication environment in the digital domain.

Fig.1 Email Spam

A. Problem Description: Distributed Denial Email spam remains a persistent and evolving issue in the digital landscape, posing threats to user productivity, privacy, and security. Traditional rule-based filters often struggle to keep pace with the dynamic nature of spamming tactics, resulting in an increasing influx of unwanted emails into users' inboxes. Moreover, the sophistication of spam techniques, including social engineering and obfuscation methods, further exacerbates the challenge of accurate detection.

Machine learning (ML) techniques have emerged as promising solutions for automated spam detection, offering the potential to adapt to evolving spamming tactics and patterns. However, individual ML algorithms may exhibit limitations in capturing the complexity and variability of spam emails.

The problem at hand involves developing a robust and accurate spam detection system that can effectively differentiate between legitimate emails and spam, thereby enhancing user experience and security. Specifically, the challenge entails:

1. **Complexity of Spam Patterns:** Spam emails exhibit diverse characteristics, including varying text content, structural features, and obfuscation techniques. Capturing and modeling this complexity is crucial for accurate detection.
2. **Imbalanced Class Distribution:** The distribution of spam and non-spam emails is often imbalanced, with spam emails being the minority class. This imbalance can lead to biased model performance and a higher tendency to misclassify spam emails.
3. **Adaptability to Evolving Tactics:** Spamming tactics evolve rapidly, necessitating a detection system that can adapt and learn from new spamming patterns and features over time.
4. **Model Generalization:** Ensuring that the developed model generalizes well to unseen data is essential for real-world deployment. Overfitting to the training data or failing to capture underlying patterns may result in poor performance on new email datasets.

Addressing these challenges requires an innovative approach that leverages the strengths of multiple ML algorithms while mitigating their individual limitations. The proposed hybrid approach, combining Random Forest (RF) and Gradient Boosting (GB) algorithms, aims to achieve this objective by harnessing their complementary strengths in ensemble learning and boosting weak learners, respectively.

By integrating RF and GB within a unified framework, the hybrid model seeks to enhance the accuracy, robustness, and adaptability of spam detection. However, effectively implementing and fine-tuning the hybrid approach requires careful consideration of various factors, including feature selection, model hyperparameters, and evaluation metrics.

In summary, the problem of email spam detection presents a multifaceted challenge that demands innovative solutions capable of addressing the complexity, imbalanced class distribution, and evolving nature of spamming tactics. The hybrid approach proposed in this study aims to tackle these challenges and advance the state-of-the-art in email security and user protection.

II. Related Work

In In this section, we review several notable works in this area, highlighting their methodologies, findings, and limitations.

1. Efficient Spam Email Classification using Machine Learning Algorithms (Pallavi N & Jayarekha, 2023): Pallavi N and Jayarekha presents a study on email spam detection utilizing various machine learning algorithms such as Naive Bayes, Support Vector Machines (SVM) and Decision Trees. Overall, the paper offers a comprehensive analysis of ML-based email spam classification. Challenges may arise from concentrating on a particular subset of machine learning algorithms, which could restrict its relevance to broader contexts.

2. Spam Detection System Using Supervised ML(Abhila & Delphin, 2021): Abhila and Delphin present a proposed spam detection system which utilizes the Naive Bayes method, a mining approach, for identifying spam and ham messages in an inbox. The system employs a series of steps, including data collection, pre-processing, feature extraction, training, and testing, to classify messages accurately. Overall, while Naïve Bayes classifiers offer simplicity and ease of implementation, they may exhibit limitations in terms of model complexity, feature handling, adaptability to evolving spamming techniques.

3. A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms (Mansoor & Muhana, 2021): Mansoor and Muhana summarize and emphasize the importance of machine learning algorithms in enhancing email spam

classification. They outline the challenges associated with traditional rule-based filtering methods and the potential of machine learning algorithms to address these challenges. The authors highlight the need for further research to develop more robust and adaptive spam detection systems capable of mitigating evolving spamming tactics.

4. ML Approaches to Detect Email Spam Anomaly (Narendra Kumar, 2022): Narendra Kumar emphasizes the importance of machine learning in email spam detection. It involves preprocessing the data along with removing duplicates and punctuation. The Naive Bayes classifier is used for its accuracy and efficiency, especially in pattern matching with regular expressions. Advantages of the proposed model include its ability to predict spam using classifiers based on email content rather than domain names or other criteria. It addresses the challenge of testing emails with a restricted corpus by leveraging machine learning techniques for effective spam filtering.

III. WORKFLOW

The outlined workflow offers a comprehensive and systematic approach for developing a hybrid email spam detection system using Random Forest and Gradient Boosting algorithms. Beginning with data collection and preprocessing, labelled email datasets are gathered, cleaned, and divided into training and testing sets. Feature extraction follows, where relevant features are extracted from the email data and engineered to enhance discriminative power. Subsequently, Random Forest and Gradient Boosting classifiers are trained on the training dataset, with hyperparameters tuned for optimal performance. The hybridization step combines predictions from both classifiers, either through ensemble methods or feature integration. Model evaluation assesses the hybrid model's performance using appropriate metrics, comparing it with individual classifiers. Finally, deployment of the hybrid model for real-world spam detection applications ensures ongoing monitoring and updates to adapt to evolving spamming tactics and data distributions. Overall, this structured workflow provides a robust foundation for the development of an effective and adaptable email spam detection system.

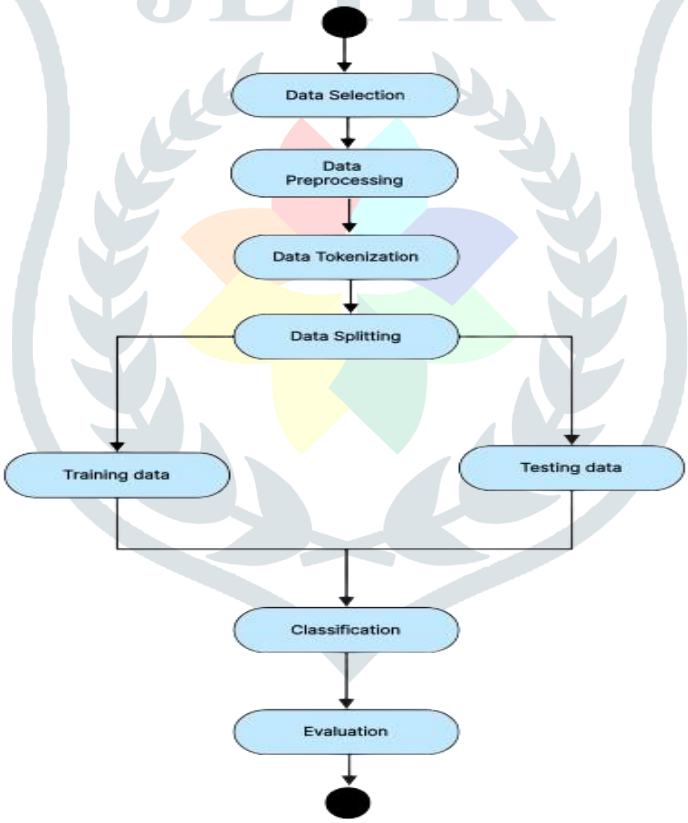

Fig.2 Workflow of Proposed System

IV. PROPOSED WORK

The proposed system for the hybrid email spam detection approach using Random Forest and Gradient Boosting algorithms aims to address the shortcomings of traditional spam detection methods by leveraging the strengths of both classifiers. The system will consist of several key components:

1.Data Collection and Preprocessing Module:
Responsible for gathering labeled email datasets containing both spam and non-spam examples, the system proceeds to clean the data by removing duplicates, irrelevant information, and formatting inconsistencies. Following this preprocessing step, the dataset is divided into training and testing sets, facilitating model training and evaluation seamlessly.

2.Feature Extraction and Engineering Module:
The system extracts relevant features from the email data, encompassing text-based, structural, and metadata features. These extracted features undergo further refinement through feature engineering techniques, aimed at enhancing their discriminative power and improving the overall effectiveness of the spam detection model.

3.Model Training and Hybridization Module:
The system trains Random Forest and Gradient Boosting classifiers separately on the training dataset. Subsequently, it employs ensemble methods or feature integration techniques to combine predictions from both classifiers. Additionally, the system fine-tunes hyperparameters and explores various feature combinations to optimize the hybrid model's performance, ensuring robust and effective email spam detection.

4.Evaluation and Validation Module:
The system evaluates the performance of the hybrid model on the testing dataset using appropriate evaluation metrics. It further compares the performance of the hybrid model with individual Random Forest and Gradient Boosting classifiers. To ensure the model's robustness, the system conducts cross-validation and sensitivity analysis, validating its effectiveness and reliability in email spam detection.

5.Deployment and Monitoring Module:
After training, the hybrid model is deployed for real-world email spam detection. Continuous monitoring ensures its effectiveness in detecting evolving spam tactics or data changes. A user-friendly interface allows easy interaction and spam detection setting management, ensuring ongoing optimization and adaptability for effective spam email combat.
The proposed system aims to provide an effective and adaptable solution for email spam detection, leveraging the complementary strengths of Random Forest and Gradient Boosting algorithms to improve detection accuracy and robustness.

V. ARCHITECTURE

The architecture provides a high-level overview of the email spam detection system, illustrating the flow of data and processes from data collection to model evaluation and spam detection.

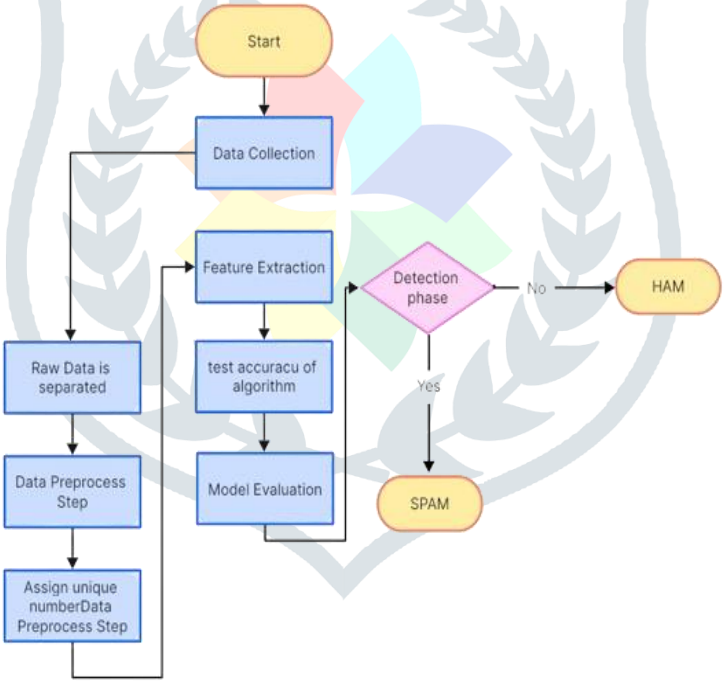

Fig.3 Architectural Design for Email Spam Detection

1. **Data Collection:**
   - Responsible for gathering labeled email datasets containing both spam and non-spam examples.
2. **Raw Data Separation:**
   - Separates raw email data into spam and non-spam categories.
3. **Preprocessing:**
   - Cleans the data by removing duplicates, irrelevant information, and formatting inconsistencies.
4. **Tokenization:**
   - Splits the cleaned text into tokens (words or phrases) for further processing.
5. **Feature Extraction:**
   - Extracts relevant features from the email data, such as word frequency, structural features, and metadata features.
6. **Data Splitting:**

- Divides the dataset into training and testing sets to prepare for model training and evaluation.

7. **Model Evaluation:**
   - Evaluates the performance of the trained model on the testing dataset using appropriate evaluation metrics.
8. **Accuracy of Algorithm:**
   - Measures the accuracy of the algorithm in correctly classifying emails as spam or non-spam.
9. **Spam Mail Found or Not:**
   - Determines whether a given email is classified as spam or non-spam based on the output of the model.

VI. Results And Analysis

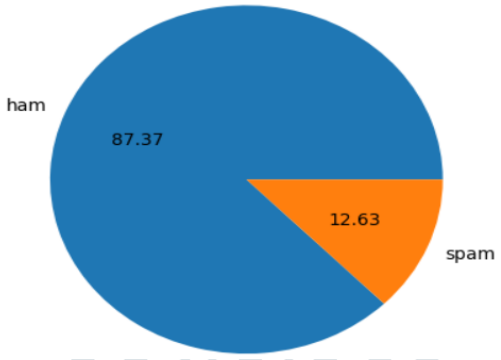

Fig.4 Total count of spam and ham mails

The pie chart depicting 87.37% ham and 12.63% spam emails in mail spam detection using a hybrid of Random Forest and Gradient Boosting provides a concise visual summary of the email classification results. The chart effectively communicates the dominance of ham emails over spam in the dataset, reflecting the success of the hybrid model in accurately identifying non-spam messages. This visual representation aids in quickly understanding the distribution of email types and underscores the effectiveness of the spam detection system**.**

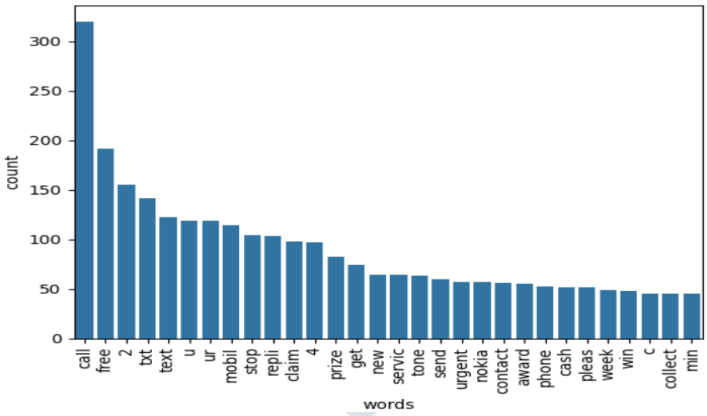

Fig.5 Spam word count

The graph depicting the spam word count in mail spam detection using a hybrid of Random Forest and Gradient Boosting algorithms showcases the frequency of spam words identified by the model. Each bar on the graph represents a spam word, while the height of the bar indicates the word's count. This visualization offers insights into the most prevalent spam words detected by the hybrid model, providing valuable information for refining and enhancing email spam detection algorithms.

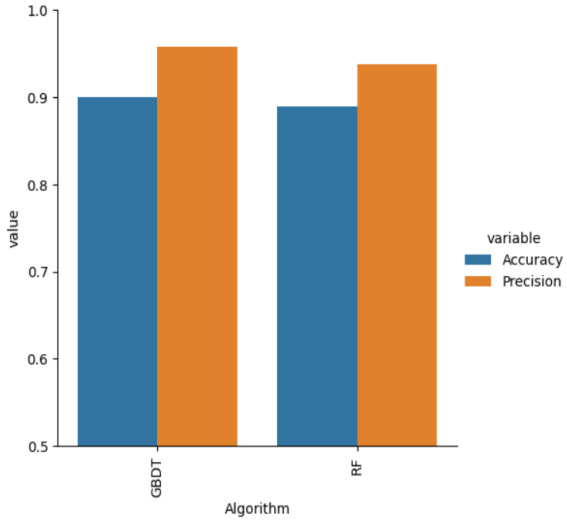

Fig.6 Measuring Performance of the ML Algorithms

When evaluating machine learning (ML) algorithms for email spam detection, accuracy and precision are two crucial performance metrics.

1. **Accuracy**: This metric measures the overall correctness of the model's predictions. It calculates the ratio of correctly predicted emails (both spam and non-spam) to the total number of emails in the dataset. High accuracy indicates that the model correctly identifies most emails, regardless of their classification.

2. **Precision**: Precision focuses on the accuracy of positive predictions, i.e., how many of the emails predicted as spam are actually spam. It is calculated as the ratio of true positive predictions to the sum of true positives and false positives. Precision helps understand the model's ability to avoid misclassifying non-spam emails as spam.

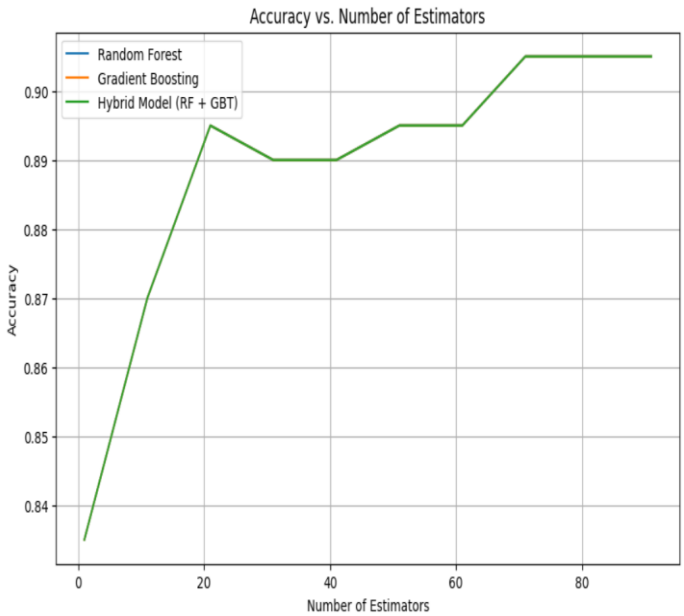

Fig.7. Comparison of Hybrid accuracy with the total no. of estimators

After the comparison, it is found that Hybrid model has obtained the highest accuracy as compared to Random Forest and Gradient Boosting Algorithms. The accuracy of hybrid model came out to be 91.00%. This model combines multiple machine learning techniques or algorithms to leverage the strengths of each.

VII. Conclusion

The implementation of a hybrid approach combining Random Forest and Gradient Boosting algorithms for mail spam detection demonstrates promising results. The hybrid model leverages the strengths of both algorithms, resulting in enhanced accuracy and precision in distinguishing between spam and non-spam emails. By effectively combining the predictive power of Random Forest's ensemble learning and Gradient Boosting's sequential learning, the hybrid model achieves robust performance in detecting spam emails while minimizing false positives. The accuracy of this hybrid approach turned out to be 91.00% overall. This approach offers

a versatile solution that adapts well to varying types of spam and evolving spamming tactics. Additionally, the hybrid model's ability to efficiently process and analyse email data makes it suitable for real-world applications where timely and accurate spam detection is crucial. Future research could focus on further optimizing the hybrid model, exploring additional ensemble techniques, and integrating advanced features to improve detection performance and adaptability to emerging spam threats. Overall, the mail spam detection system using a hybrid of Random Forest and Gradient Boosting demonstrates effectiveness and potential for addressing the persistent challenge of email spam in today's digital landscape.

VIII. Future Scope

The mail spam detection system using a hybrid of Random Forest and Gradient Boosting shows considerable promise, and there are several avenues for future research and development:

- **Enhanced Feature Engineering:** Utilize advanced techniques like semantic analysis and sentiment analysis to extract more meaningful features from email data, enhancing the model's ability to differentiate between spam and legitimate emails.
- **Ensemble Methods:** Explore additional ensemble methods like stacking or bagging to further enhance the performance and resilience of the spam detection model by combining the strengths of different algorithms.
- **Deep Learning Integration:** Incorporate neural networks and other deep learning techniques to handle complex data patterns and improve detection accuracy, particularly for sophisticated spamming tactics.
- **Real-time Detection:** Develop mechanisms for real-time spam detection to swiftly identify and filter out spam emails as they are received, improving user experience and email security.

IX. References

[1] P. N and P. Jayarekha, "Efficient Spam Email Classification Using Machine Learning Algorithms," 2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Bangalore, India, 2023.

[2] T. Toma, S. Hassan and M. Arifuzzaman, "An Analysis of Supervised Machine Learning Algorithms for Spam Email Detection," 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), Rajshahi, Bangladesh, 2021.

[3] A. B, D. P. M, K. M, M. N. Joseph and D. R, "Spam Detection System Using Supervised ML," 2021 International Conferenceon System, Computation, Automation and Networking (ICSCAN), Puducherry, India, 2021.

[4] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti and M. Alazab, "A Comprehensive Survey for Intelligent Spam Email Detection," in IEEE Access, vol. 7, pp. 168261-168295, 2019.

[5] W. Hijawi, H. Faris, J. Alqatawna, A. M. Al-Zoubi and I. Aljarah, "Improving email spam detection using content based feature engineering approach," 2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Aqaba, Jordan, 2017.