WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Enhancing Email Security: A Comprehensive Approach with Ensemble Learning

Killamsetty Sreya

Student, Rajam, Vizianagaram, 535127, Andhra Pradesh, India.

ABSTRACT

Social network is prevalent that all people across the globe are visible to anyone and anywhere. Increase in growth of interest in various social network platforms lead to the huge number of interactions between the users to users or users to websites. Among all, most of the business and general communication agents are working through email because of its cost effectiveness as sending an email is easy and cheap. This leads to various attacks like Spamming. Therefore, detecting of these spam mails that were fraud is more important. We can detect spam emails using machine learning techniques which improves a way in social network analysis. In this work, spam detection includes different Machine Learning Techniques such as supervised and unsupervised learning. This research proposes a bagging method (ensemble learning) for email spam detection by combining two machine learning baseline models of random forest and J48 (decision tree).

According to the study, the proposed bagging method produced an accuracy rate of 95 percent.

Keywords: Spamming, Machine learning, Supervised, Bagging, Ensemble learning, Accuracy rate

1. Introduction

Over the past few years, the Internet has been leaping forward, and the intelligent terminals have been progressively popularized and in this form, Machine learning models have been utilized for multiple purposes in the field of computer science from resolving a network traffic issue to detecting a malware. Coming to digital communications, Email is an primary medium throughout the world. Every personal, social and business communication needs email. No one wants to receive emails that are not related to their interest because they waste receiver's time and resources. Besides, these emails can have malicious content hidden in the form of attachments or URLs that may lead to the host system's security breaches. Email spamming is generally defined as the act of dispersing messages that are unsolicited sent in bulk, using the medium of email. On the other side, emails that are communicated for genuine, lawful and authorised and legitimate purposes are defined as Ham. Spam is any irrelevant and unwanted messages or emails sent by the attacker to a significant number of recipients by using emails or any other medium of information sharing. Spammers use the act of spamming for not only marketing purposes, but also to achieve more malicious goals such as reputational damage and financial disruption, both in institutional and personal front.

In this journal, we delve into the realm of spam email detection, exploring the efficacy of ensemble learning techniques. By combining the strengths of multiple algorithms, our approach aims to enhance the accuracy and robustness of spam classification, paving the way for more reliable email filtering systems.

2. Literature Survey

In paper [1]. This research presents a new hybrid bagging technique that combines the random forest and J48 (decision tree) algorithms with machine learning for the identification of email spam. To increase the efficiency of the suggested approach, the research discusses the usage of tokenization, stemming, stop word removal, and correlation feature selection (CFS) during the preprocessing phase. The J48 approach yielded precision and recall values of 94 percent and 90 percent, respectively, whereas the random forest classifier's values were 86 percent and 82 percent, respectively. According to the research, the efficacy of the suggested strategy can be increased by combining more complex methods including dataset procedures and evolutionary algorithms.

In paper [2]. A sophisticated framework is designed in which experts and machine learning algorithms collaborate worked together to detect spam filed effectively. The author has used tree-based J48, random forest (RF), rule-based PART, Naïve Bayes network algorithms in Machine Learning. The paper have shown by experiments that it is difficult to detect spam using only machine learning because of various problems encountered in reality. So, author proposed a collaborative system including both machine learning and experts for spam detection. While reducing the time-consuming or costly expenses that can be problematic when experts are involved.

In paper [3]. This paper aimed to detect the spam emails with machine learning algorithms that are optimized with bio-inspired methods. The paper has implementation of various Machine Learning algorithms that are Naive Bayes, Support Vector Machine, Random Forest, decision tree along with feature extraction and pre-processing combined with bio-inspired algorithms. Various ratios of training and testing were performed to get the best ratio among 60:40, 70:30, 75:25, 80:20 and 80:20 was chosen as it got good F1-score, Precision and Recall in comparison to Accuracy.

In paper [4]. The paper trains the logistic regression model for spam email detection using the Teaching-Learning-Based Optimization (TLBO) algorithm. The proposed solution is evaluated on two benchmark spam email datasets (CSDMC2010 and TurkishEmail) and compared against seven other metaheuristics algorithms commonly used in the same experimental setup. On both English and Turkish datasets, the suggested LR-ITLBO approach demonstrated exceptionally high levels of accuracy in spam email detection. The paper also mentions the need for additional experiments with more real-world email datasets to further enhance the credibility of the approach.

In paper [5]. In order to overcome the shortcomings of current methods, the paper suggests a novel approach to email spam detection that combines machine learning models with an improved sine cosine swarm intelligence algorithm. The study makes use of a hybrid machine learning-metaheuristics framework that combines XGBoost models trained with an improved sine cosine swarm intelligence algorithm and logistic regression. The hybrid approach that was suggested performed better in terms of recall, accuracy, precision, and f1 score. The study also emphasizes the difficulty of solving NP-hard problems when adjusting the XGBoost hyperparameters and the requirement for metaheuristic algorithms to deal with them.

3. Methodology

Data Collection and Preprocessing

1. Gather Email Data: Collect labelled email data, including both spam and legitimate emails. This data should be representative of the types of emails you encounter in your environment.
2. Engineering features:

Textual Elements:

- Take words, phrases, and n-grams out of subject lines and email body content.
- Use text cleaning methods to cut down on noise, such as lemmatization, stop word removal, and stemming.
- To capture the tone and themes of emails, think about using sentiment analysis or topic modeling.

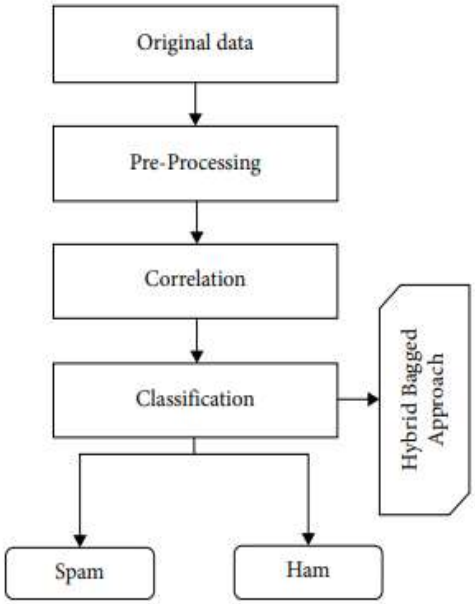
Structural Features:

- Extract information such as sender, recipient, date, time, and subject from email headers.
- Examine email structure (content type, HTML vs. plain text, attachments, etc.).

Sender/Recipient Features:

- Take into account email address patterns, domain details, and sender reputation scores.
- Examine the actions of the recipient (such as previous correspondence

3. Data Cleaning:
    - Missing Values: Use the appropriate methods (e.g., mean, median, mode, predictive models) to impute missing values.
    - Outliers: To avoid model bias, identify and manage outliers using statistical techniques or domain expertise.
    - Noise Reduction: Use methods like feature selection or dimensionality reduction to cut down on errors and noise in the data.
4. Data Transformation: Apply appropriate transformations (e.g., TF-IDF for text features, one-hot encoding for categorical features) to make data suitable for machine learning algorithms.
5. Data Normalization: Scale features to a common range (e.g., using min-max scaling or standardization) to improve model convergence and performance.



Model training

1. Bootstrap sampling: Generate multiple random samples with replacement from the original dataset. Each sample should be approximately the same size as the original data. This creates diverse training sets for individual models.
2. Train base learners: For each bootstrap sample, train a base learner using your chosen algorithms. In this case, we would train both a Random Forest and a J48 decision tree model on each sample.
3. Hyperparameter tuning: Optimize the hyperparameters of each base learner individually for each bootstrap sample. This can significantly improve model performance.

Prediction

For each new email:

- Apply the same pre-processing steps (cleaning, tokenization, feature extraction) as used for the training data.
- Run the new email through each of the trained base learners (Random Forest and J48 models from each bootstrap sample).
- Aggregate the predictions from each base learner. For classification tasks like spam detection, you can use majority voting, where the most frequent prediction (spam or ham) from the individual models wins.

Evaluation

1. Measure performance: Evaluate the effectiveness of the bagging ensemble using traditional metrics like accuracy, precision, recall, F1-score, and AUC-ROC curve for binary classification.
2. Compare with individual models: Compare the performance of the bagging ensemble with the performance of the individual Random Forest and J48 models trained on the original dataset. Bagging usually improves performance by reducing overfitting and variance.
3. Analysis and improvement: Analyze the results and identify potential areas for improvement. Consider applying different feature engineering techniques, hyperparameter tuning strategies, or even different base learners within the bagging ensemble.

4. Results and Discussion

Here is a comparison of the results and accuracy of J48 (decision tree), random forest and Bagging approach of both the algorithms.

Algorithm	**Accuracy**
Random Forest	93%
J48	91.5%
Bagging Approach	95%

Users need to understand which emails are deemed spam and which are not, from a security standpoint. The study yields a variety of observations, particularly in the field of propositions based on machine learning. Because of the different issues that arise in real life, it is challenging to identify spam solely through machine learning. When spam detection is left to the experts alone, it can result in more expensive or time-consuming expenses. A

framework that combined machine learning techniques with the expertise of experts worked well together to detect spam on social networks. Nowadays, a lot of spam-email detection techniques rely on a single model, which increases the risk of overfitting and errors. While ensemble models are widely applied in other areas of machine learning, their use in spam email detection is less common. It has been noted that employing ensemble models improves consistency. In this study, we came to know that by using bagging approach we get better results.

5. Conclusion

From a security perspective, users value the classification of emails as spam and ham above all else. Before being used, all classification techniques must first be trained to distinguish spam emails from regular emails. These techniques are trained on a training set of data. However, spam mail continues to exist despite all of this effort. They continue because a new type of spam email is introduced every day. Because of this, new spam messages continue to arrive even if the old ones are sorted and marked. Updating the training materials with information about the latest forms of spam is one way to find a solution. It should be successful in doing so, the spam mail will be handled before it gets to our mailbox. Additionally, this will save us time because our inbox will be less cluttered and it will be simpler to locate important emails. In conclusion, machine learning, particularly supervised learning, is crucial to the classification process used to identify spam mail in real life. In order to achieve better results, more research is needed to compare machine learning models with deep learning models.

REFERENCES

1. Alanazi Rayan, "Analysis of e-Mail Spam Detection Using a Novel Machine Learning-Based Hybrid Bagging Technique", *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 2500772, 12 pages, 2022. https://doi.org/10.1155/2022/2500772
2. J. Choi and C. Jeon, "Cost-Based Heterogeneous Learning Framework for Real-Time Spam Detection in Social Networks With Expert Decisions," in IEEE Access, vol. 9, pp. 103573-103587, 2021, doi: 10.1109/ACCESS.2021.3098799.
3. S. Gibson, B. Issac, L. Zhang and S. M. Jacob, "Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms," in IEEE Access, vol. 8, pp. 187914-187932, 2020, doi: 10.1109/ACCESS.2020.3030751.
4. Berrou, B. K., Al Kalbani, K., Antonijevic, M., Zivkovic, M., Bacanin, N., & Nikolic, B. (2023, January). Training a Logistic Regression Machine Learning Model for Spam Email Detection Using the Teaching-Learning-Based-Optimization Algorithm (Vol. 104, p. 306). Springer Nature.
5. Bacanin, N., Zivkovic, M., Stoean, C., Antonijevic, M., Janicijevic, S., Sarac, M., & Strumberger, I. (2022). Application of natural language processing and machine learning boosted with swarm intelligence for spam email filtering. *Mathematics*, *10*(22), 4173.