WWW.IJRPR.COM

International Journal of Research Publication and Reviews

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

Spam Email Detection Using Ensemble Learning

Kasireddi Lahari

Student, Rajam, Vizianagaram, 532127, india.

A B S T R A C T

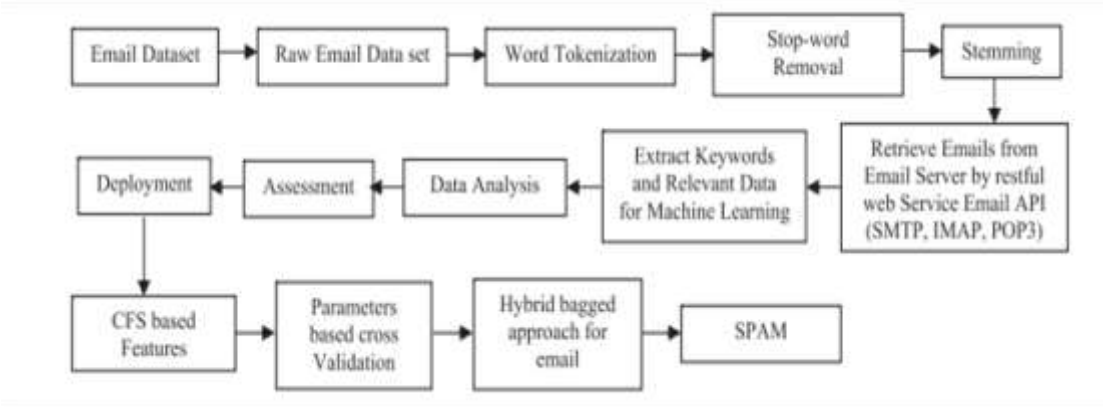
Social networking facilitates connection, cooperation, and contribution. Email is the most cost-effective communication tool used by business and general communication agents among all social networking interactions since sending an email is simple and inexpensive. This results in several attacks. such as link manipulation, phony websites, spamming, phishing emails, and many others. Determining whether of these spam emails were fraudulent is therefore crucial. This study uses machine learning techniques to detect spam emails, creating a secure avenue for social network participation. Several machine learning techniques, including supervised and unsupervised learning, are used in this work to detect spam. Naive Bayes and Support Vector Machines (SVM) are used in supervised learning. The analysis found that Naive Bayes provided 87% and 90% accuracy was obtained by the Support Vector machine, in contrast. The accuracy rate for identifying spam emails will rise with the use of ensemble models like random forest.

Keywords: Spamming, supervised, Naïve Bayes, Accuracy rate, Support vector machine, ensemble learning.

Introduction

In the last several years, machine learning models have found numerous applications in computer science as a result of the Internet's rapid advancement and the increasing popularization of intelligent terminals. When it comes to digital communications, email is the main channel used worldwide. Email is necessary for all commercial, social, and personal communications. This results in other assaults, such as spamming.

The act of sending unsolicited communications in large quantities via email is sometimes referred to as email spam. Conversely, emails that are sent for real, authorized, lawful, and legitimate reasons are referred to as Ham. Spammers employ the act of spamming not just for commercial gain but also for more malevolent objectives like financial disruption and harm to one's reputation on an institutional and personal level.



In order to jointly provide superior predictions, ensemble learning makes use of the combined intelligence of several algorithms, including decision trees, random forests, and gradient boosting. Higher accuracy is the outcome of each model's contribution of its special strengths and compensating for its own shortcomings.

Literature Survey

In Paper [1]: In order to increase performance, the current work uses hyperparameter tweaking to optimize the baseline models of the random forest and XG Boost algorithms for spam email detection. The results showed that both models' performance was considerably enhanced by hyperparameter adjustment However, it was discovered that the created XG Boost model was successful and efficient in identifying spam emails. The dataset utilized in this study has a complementarily balanced class distribution. These models will behave differently when the dataset is noticeably unbalanced. This study

made a distinction between two kind of spam emails: semi-spam, or spam that is seen as spam by some users but not by others, and complete spam, or spam that is regarded as spam by all users. They created a technique for detecting spam that combines crowdsourcing for the identification of semi-spam and Bayesian filtering for complete spam. The method's crowdsourcing component entails asking acquaintances or reliable users with comparable interests to submit reports of spam. The model's accuracy rate was 95.1%.

In Paper [2]: This research presents a new hybrid bagging technique that combines the random forest and J48 (decision tree) algorithms with machine learning for the detection of email spam. To increase the efficacy 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 accuracy 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 may be increased by combining more complex methods including dataset processes and evolutionary algorithms. The advantage of marketing stems from the capacity to pinpoint a specific target market thanks to self-characteristics like age and sexual orientation that are displayed on profile sites.

In Paper [3]: The study suggests a spam email filtering system that classifies emails using two distinct feature selection techniques: rough set theory and TF-IDF. They used machine learning techniques and got results that were comparatively decent. The study suggests a hybrid bagging strategy based on machine learning for spam email detection, utilizing J48 (decision tree) and Naïve Bayes algorithms. The hybrid bagged approach's overall accuracy of 87.5% shows how successful it is in identifying spam emails. The accuracy of the J48 method is 91.5%, compared to 83.5% for the individual Naïve Bayes algorithm. Filtering is necessary for the email's categorization in order to determine if it is spam or ham. Two distinct feature selection techniques are used in the spam email filtering system that Mohamad and Selamat have suggested to categorize the emails.

In Paper [4]: The study employs a hybrid bagging strategy for spam email identification that combines the J48 (decision tree) and Naïve Bayes machine learning algorithms. The findings of the comparison study indicated that the hybrid bagged technique performed better in terms of accuracy, recall, and precision than the J48 and Naïve Bayes decision tree algorithms. The suggested method's total accuracy in identifying spam emails was 88.12%. The utilization of machine learning algorithms in data science for spam email identification is a crucial aspect of email security enhancement and well-organized receipt of emails. The electronic mail, or email, communication technology is the most extensively used and popular one. Numerous organizations around have been dedicating their efforts to detecting spam emails. The writers who were the subject of a discussion on the identification of spam or ham are further detailed. The categorization of emails as spam or ham requires the use of filtering techniques

In Paper [5]: The growing issue of email spam is covered in the article, along with the importance of spam detection. It examines many machine learning methods for spam filtering and offers a thorough analysis of them based on recall, accuracy, and precision. Supervised machine learning techniques form the foundation of the majority of suggested strategies for email spam detection. When it comes to spam identification, supervised learning algorithms like SVM and Naive Bayes perform better than other models. In addition to offering thorough explanations of these algorithms, the study makes recommendations for future lines of inquiry into email spam filtering and detection. According to estimates from social networking specialists, 40% of social network accounts are exploited for spam.

Methodology

Dataset:

The Enron dataset, which is comprised of six main directories, each of which has several subdirectories, each containing emails as a single text file, was used in this study because it is the only significant collection of public emails and because researchers use it extensively.

Preprocessing phase:

Eliminating unnecessary characters or characteristics that make up noise from the data is crucial to improving the classification's quality. lists the cleaning procedures and tools that were utilized.

Splitting of dataset:

After the data cleaning procedure was finished, the data was separated into two groups: the train set and the test set. Seventy percent of the original dataset made up the train set. Thirty percent of the original dataset makes up the test set.

Machine Learning Algorithms:

Random Forest:

For both classification and regression, the supervised ensemble classifier Random Forest is employed. It is an ensemble learning technique that builds a collection of decision trees and aggregates them to get a final forecast. The random forest does the following actions to provide a prediction:

Step 1: Select a random sample of data from the dataset.

Step 2: Build a decision tree using the sample data.

Step 3: Repeat the process a certain number of times, creating a new decision tree each time.

Step 4: By averaging their projections, combine the decision trees. In a random forest, every decision tree produces a forecast; the ultimate prediction is calculated by averaging the predictions from each decision tree.

XG Boost:

An approach for supervised ensemble machine learning is called Extreme Gradient Boost (XGBoost). The method uses the boosting approach as an ensemble technique.

The XG Boost algorithm's seven-step process is as follows:

Step 1: First, a collection of decision-tree models is trained to identify whether a limited number of emails are spam or not.

Step 2: Boosting- By concentrating on incorrectly categorized emails, new models are added to the ensemble and taught to fix errors produced by earlier models.

Step 3: Gradient Descent- Gradient descent is used to minimize the ensemble's loss function in order to optimize the parameters of new models.

Step 4: Regularization- To avoid overfitting, regularization is applied.

Step 5: Pruning- To reduce overfitting and enhance generalization, low-weight leaves on the decision tree are removed.

Step 6: Repeat- After a predetermined number of iterations or until a predetermined stopping condition is satisfied (such as achieving a predetermined accuracy level), steps 2 through 5 are repeated.

Step 7: Return- By using a majority vote among the base models, the final ensemble of base models is returned as the final model and may be used to categorize fresh emails as spam or not.

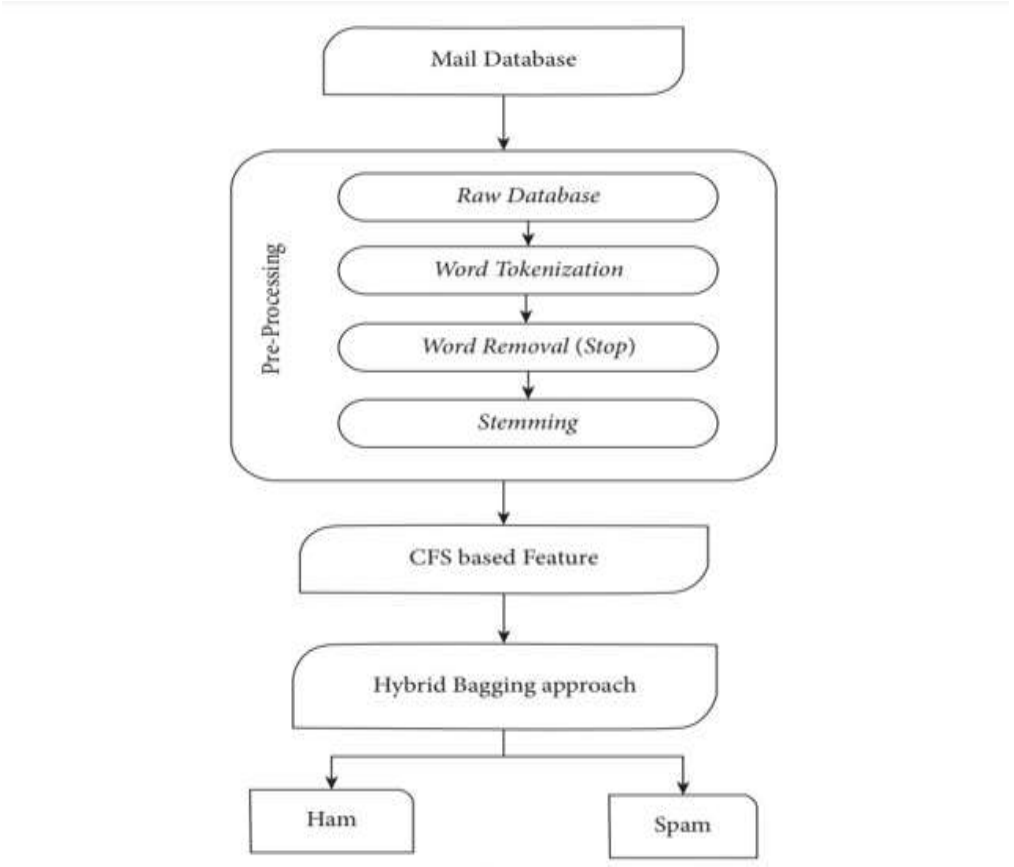


Fig. 2 – Spam mail Identification.

Results and discussions:

Users need to know which emails are spam and which are not, from a security standpoint. The study yielded a variety of observations, particularly in the area of machine learning-based propositions. First, real-world problems make it challenging to detect spam solely through machine learning; second, involving experts in the process of spam detection can result in more time-consuming or expensive expenses that may pose a problem; and third, a framework that combines machine learning techniques with expert judgment performed well in detecting spam on social networks. Most of the spam-

email detection methods currently in use rely on a single model, which can cause overfitting and errors. Ensemble models have not been applied as much in spam email detection, while being widely used in other machine learning applications. There is evidence to suggest that employing ensemble models improves consistency. In contrast to other machine learning algorithms, Random Forest demonstrated good performance. Using the spam corpus dataset, random forest achieved 99.9% accuracy.

Spam Corpus	**Spam base**
Marketing	Word_freq_address
Credit	Word_freq_remove
Offer	Word_freq_internet
Money	Word_freq_order
Loans	Capital_run_length_longest

Table 1: Useful features of the datasets

4. Conclusion:

From a security perspective, users value the classification of emails as spam and ham above all else. Before being utilized, all classification algorithms must first be taught to distinguish spam emails from regular emails. These approaches 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 sifted and tagged. Updating the training materials with knowledge on the latest forms of spam is one way to find a solution. Should it be successful in doing so, the spam message 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 get better outcomes, further research is needed to compare machine learning models with deep learning models.

References

1. Ghosh, Argha, and A. Senthilrajan. "Comparison of machine learning techniques for spam detection." Multimedia Tools and Applications (2023): 1-28.
2. Omotehinwa, T. O., & Oyewola, D. O. (2023). Hyperparameter optimization of ensemble models for spam email detection. Applied Sciences, 13(3), 1971
3. 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.
4. Sharma, P., & Bhardwaj, U. (2018). Machine Learning based Spam E-Mail Detection. *International Journal of Intelligent Engineering & Systems*, *11*(3).
5. Rakesh Nayak and Salim {Amirali Jiwani} and B. Rajitha, " Spam email detection using machine learning algorithm",in Sciencedirect,Materials Today: Proceedings,ISSN 2214-7853,https://doi.org/10.1016/j.matpr.2021.03.147,2021
6. Naeem Ahmed, Rashid Amin, Hamza Aldabbas, Deepika Koundal, Bader Alouffi, Tariq Shah, "Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges", Security and Communication Networks, vol. 2022, Article ID 1862888, 19 pages, 2022.