Shivam Devidas Gawade, 2025, 13:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Spam Email Detection Using Machine Learning

Shivam Devidas Gawade , Professor Nishant Rathod
Department of Master of Computer Application
Anantrao Pawar College of Engineering and Research Pune

Abstract- The rapid development of digital communication has resulted in a huge volume of email including unsolicited spam, which can cause serious problems such as criminal fraud, time wastage and difficulty in identifying useful emails The aim of this study is to develop pattern-based machine learning that accurately detects and filters spam emails It can do that. By leveraging algorithms to analyze email content, sender information, and metadata attributes, we address the growing need for an efficient, scalable solution to this problem. Our approach involves pre-processing email data through tokenization, stopword extraction, stemming, and vectorization, followed by feature extraction focusing on content-based, metadata, behavioral attributes. We look at how different machine learning models some including Naive Bayes, Random Forest, Gradient Boosting are performed Model performance is evaluated using , and F1-scores The study concludes that clustering methods, especially random forests, provide solutions that are difficult for, balances accuracy and computational efficiency. Although deep learning models such as CNN and NLP-based transformers provide good detection capabilities, their inherent robustness limits their practical application in small-scale applications Future work should focus on nature further integration of advanced language processing techniques to improve the effectiveness and efficiency of spam email detection.

Keywords- Spam email, Detection, Machine Learning.

I. INTRODUCTION

Spam mail is defined as irrelevant and unwanted messages that are received in large numbers by the user and are risky. Such emails may also be designed with phishing schemes, which aim at tricking victims to divulge personal/critical information, or they may be used in the distribution of viruses and other malware. The evolution of communication due to the increased availability of internet and email available has also come as a boon for the growth of email abuse that is spam mails at an alarming rate across the global perspective. While the inconvenience caused by such spam emails may be minimal, other more massive direct and indirect losses, including security threats posed by cyber créateurs and attacks to corporations or individuals, are highlighted by others. Given that both business and personal communication has increasingly become dependent on the use of e-mail, concerns on anti spam measures are at their highest. Besides, such emails take up the bandwidth and time which are precious in the first place and center quite a lot of attention which deters the quality and safety of the online interaction. To this end, many organizations and investigators have been taking the initiative in creating approaches for recognizing spam mail and preventing its further delivery to mailboxes.

One such traditional method over the years has been the use of a rulebased filtering approach e-mails regarding spam, which has now begun to be inadequate. Despite their effectiveness to some

extent, they have lagged behind as spammers are becoming more creative and altering their techniques in order to defeat the filters. Spammers have been rapidly evolving and outsmarting filters that have been put in place to minimize spam which is a nuisance to most users of the Internet. Unfortunately competition in email spam detection has become more challenging as new spenters have entered into the fray. Machine learning (ML) has evolved as one of such solutions to counter and minimize spam towards email inboxes. While the limits of rule based approaches have been recognized, such systems developed using MF administer dominance in spam detection and filtering over a period. These algorithms look for patterns in the emails and decide whether the e-mail belongs in the category of spam or in a structure other than spam. There is hope in trojan possibilities in monetizing spam but encouraging married subscribers' spam only leaves em financial helps. Spammers have an advantage that as spammers' ingenuity grows, opposing marketers' technology also advances. To be precise effort must be made to ensure that spam information search is worth or corresponds to the intended enrichment and improvement of a study. Email became an important integral part of business communication and therefore deserved special attention for its protection against unwanted advertising. In this research we describe a pattern recognition based machine learning approach for spam detection instead of static rule based techniques.

Our approach centers toward improving classification accuracy through tapping a number of email features in turn enhancing the system that systems administrator managers in the face of changing tendencies of spam emails. In particular, while this is certainly quite limited and thus surface-focused as it anchors the analysis of mails on keywords or sender addresses only, our technique is broader than this. This frees up new possibilities for the analysis of the spam and the legitimate email by providing additional insights into the structural and language elements that separate the two classes of emails.

As noticing an invasion of certain words or an attack message format within an email would provide an idea if such an email is a spam message or not. The subject line analysis provides another useful tool to detect phishing emails where one may look for extended subjects urging and pleading people to provide sensitive information or some other kind of information in return for money. The accuracy of identifying spam messages also can be improved by looking at the sender information, the domain and the reputation of the domain. Last, emails come with headers giving details regarding the path taken by the email across various mail servers and this information is very sensitive as it is associated with ordain fraud the email header contains certain pf medicine that would be useful to c hoose cardiovascular diseases medications judiciously by deliving imformation of the patients comorbidity diseases and conditions with cardio vascular implications to the systems har.

Our Approach has combined such various features to enable machine learning systems convert more email correctly and precisely.Integrating the Webb resource using the strategy of sender verification, message content exploration, and subjects of the message, we aim at beating the spam detection mechanisms with the efficiency of other authors. This entails that this approach is emphasized more where spam is present, which means high accuracy and low false positive rates, an important concern in spam fighting.

In any case, more and more sophisticated spam email will demand an improved detection. As such, machine learning has emerged as one of the most effective means to achieve this goal since it changes along with spammers' practices. Here, we aim in further development towards using machine learning by building a system based on patterns in emails that combines different email parameters in order to increase spam detection. This response aims to not only enhance the security of emails communications but also improve the efficiency in email communication management systems including the email filtering and sorting for both the users and the organizations.

II. PROBLEM STATEMENT

Due to the high volume of emails received daily, it becomes challenging to manually detect and filter out spam, which can lead to criminal fraud, waste of time, and difficulty in finding useful emails. To address these issues, we need to develop a Machine Learning model capable of accurately identifying and filtering spam emails. This model will help safeguard users against fraud, enhance productivity by reducing time wasted on irrelevant emails, and ensures that important emails are not overlooked.

III. SOLUTION

We propose to develop this model using a combination of Machine
Learning (ML) and Natural Language Processing (NLP) techniques. By leveraging NLP, the model can analyze the content of emails to identify suspicious keywords and patterns that are typically associated with spam. The model will be trained iteratively using a trial-and-error approach, allowing it to learn from previous spam keywords and adapt to new threats. Over time, the model will evolve to become more robust in distinguishing between legitimate and spam emails, thereby reducing the likelihood of fraud and enhancing users experience.

IV. LITERATURE REVIEW

Many studies have focused on spam detection, from rule-based algorithms to advanced machine learning. Rule-based systems rely on fixed rules to categorize emails but are less applicable to other forms of spam. Over the past decade, machine learning models have gained popularity in spam detection due to their scalability and ability to learn from data.

The *Naive Bayes Classifier* has become a popular method for spam detection, mainly due to its simplicity and effectiveness in dealing with big data [1]. However, Naive Bayes models usually assume independence between items, which is not always true in real-world email datasets [2]. Support vector machines (SVM) have also been widely used for email classification. SVMs map data in high-dimensional spaces and have been shown to be effective in binary classification problems including spam detection . Although SVMs provide high accuracy, they can be computationally expensive, especially for large data sets [3] .

Recent research has focused on *Deep Learning* techniques, including
Convolutional Neural Networks (CNN) And Recurrent Neural Networks (RNN).
These models have shown improved performance over traditional methods by extracting features from raw data [4][5]. However, deep learning models require significant computational resources and large datasets, which can be difficult for small applications.

Promising results were also obtained using ensemble methods such as Random Forest and Gradient Boosting. These methods combine several weak classifiers into more complex classifiers, which improve classification performance [6] . Even if accurate, cluster models can be slow to train and difficult to interpret [7] .

Recent advances in *Natural Language Processing (NLP)* have led to the use of BERT and GPT transformers for spam detection.

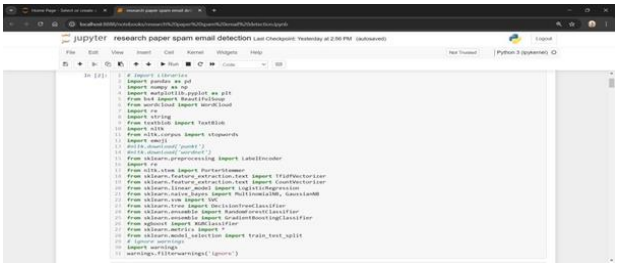
This model is able to understand the context and semantics of emails, resulting in better detection rates [8]. However, they need higher computational power and more training data [9].Researchers have explored hybrid models, such as combining Naive Bayes and SVMs, to balance performance and computational efficiency in spam detection [10].The impact of feature selection techniques, such as chi-square and information gain, has been highlighted in improving the performance of classifiers in spam email detection [11][12].Several studies have demonstrated the importance of using behavioral features such as user interaction patterns and email sending frequency, which can complement content-based features to enhance detection accuracy [13][14].

V. METHODOLOGY

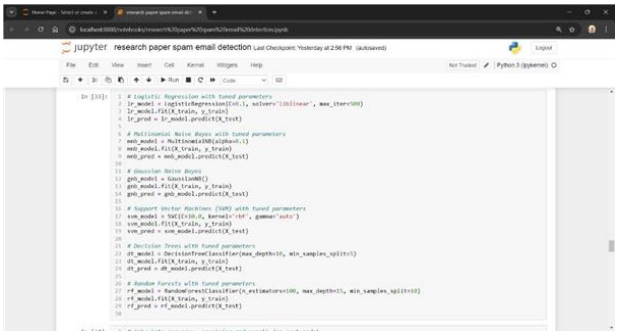
This study focuses on the development and evaluation of spam email detection model using machine learning method. We use a supervised learning approach, using labeled data of spam and non-spam (ham) emails. The process has several steps: data collection, preprocessing, feature extraction, model training, analysis, and deployment.

- **Summary :**
  The dataset used in this study is the publicly available spam email dataset, which contains thousands of emails labeled as spam or ham. The data structure includes email content, subject line, sender information, and subject line information. **Pre-processing data :**
  Preprocessing is necessary to clean the data and prepare it for modeling. Preliminary steps include:
- **Tokenization :** Email text is broken down into individual words or tokens.
- **Stopword Removal :** Removal of common words that do not contribute to spam detection, such as "the", "is", and "at".
- **Stemming/Lemmatization :** Reducing words to their original form (e.g. "run" becomes "run").
- **Vectorization :** Conversion of text to numerical representation using techniques such as Term Frequency-Inverse Document Frequency (TFIDF) or word processing. Filtering : We extract special features from the emails, e.g.
- **Content-based :** the frequency with which particular words, phrases, or connections are used.
- **Metadata attributes :** Information from the email header, such as the sender's domain, sent time, and subject.
- **Behavioral features :** Patterns of sending behavior, such as multiple emails being sent from the same IP address over a relatively short period of time.
- **Sample Training :** We explore various machine learning algorithms including Naive Bayes, Random Forest, and Gradient Boosting to determine the best performing model. Each model is trained on a subset of pre-processed data, using cross-validation to ensure robustness. Sample analysis .

- **Models are evaluated using common classification criteria:**
- **Accuracy :** Percentage of emails that are correctly shared. - Specific : Proportion of genuine spam emails among those classified as spam.
- **Recall :** Part of the correctly identified genuine spam emails. - F1score : harmonic mean of precision and recall, providing a balanced measure **of performance. System Required :**
- Raw data for training and testing : A diverse and comprehensive dataset of emails, including a balanced mix of spam and legitimate messages, is crucial for the model's accuracy.
- **Various software tools : > Software tools :**



Libraries such as pandas, sklearn, and nltk for data processing and model implementation. We have to fit machine learning model for that we use python and popular library numpy, pandas matplotlib and seaborn. For implementation strategy, using python library and get a most accurate result. We can see the result step-by-step by below –



- **Algorithms:**

Data Training and Testing For Various Algorithms :

Various classifiers, including Logistic Regression, Naive Bayes, SVM, and Random Forest, with tuned parameters.

1. LR : Logistic regression is used in statistics for predicting the categorical result like yes / no.

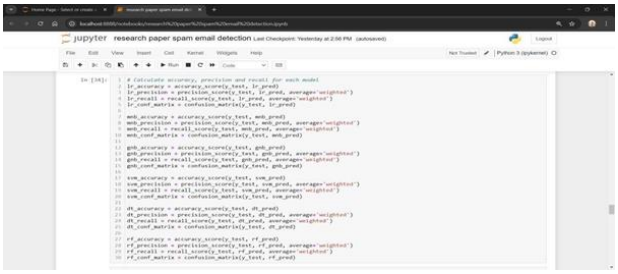
2. Multinomial Naive Bayes : This is mainly used for discrete data, such as word counts in text classification. It assumes that the features (like words in a document) follow a multinomial distribution, meaning that each feature can take on multiple discrete values.

3. Gaussian Naive Bayes : This variant is typically used for discrete data, especially for text classification problems like document classification, spam detection, or sentiment analysis. It is effective for data where the features are counts or frequencies of events.

4. SVM : SVM is used for the sorting the data into classification and regressions but mostly used for classifications.

5. Decision Tree : This algorithm is used to sort things into groups. it is look like tree and it is quite similar to the humans thinking ability.

6. Random forest : It is the based on ensemble method. in random forest we use boosting as well also used for both classification and regression but mostly used for classification.



Accuracy :

- **LR :** 0.971899224806205
- **MNB :** 0.9563953488372093
- **GNB :** 0.8691860465116279
- **SVM :** 0.8953488372093024
- **DT :** 0.9525193798449613
- **RF :** 0.9147286821705426

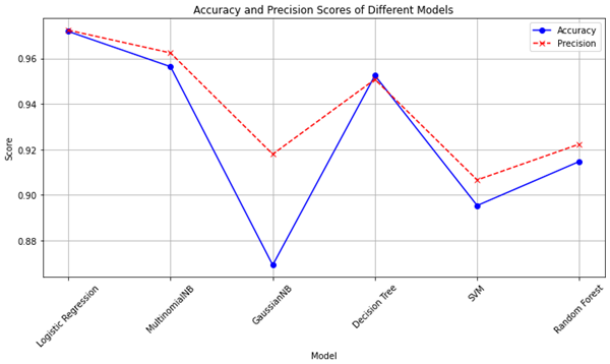
Precision Score :

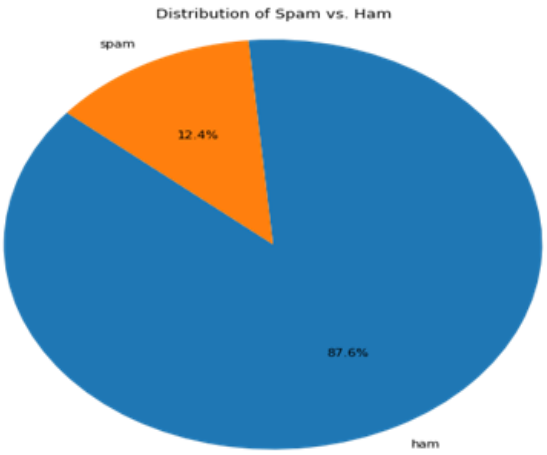
- **LR :** 0.9724528699938659
- **MNB :** 0.9624374427464473
- **GNB :** 0.9179176426684793
- **SVM :** 0.9066061336051144
- **DT :** 0.9508188032736695
- **RF :** 0.9223545723829332

Recall Score :

- **LR :** 0.9718992248062015
- **MNB :** 0.9563953488372093
- **GNB :** 0.8691860465116279
- **SVM :** 0.8953488372093024
- **DT :** 0.9525193798449613
- **RF :** 0.9147286821705426

Process Diagram :





VI. CONCLUSION

This study concluded that the pattern-matching machine learning method developed to identify spam emails proved to be practical and effective in reducing the flood of junk and harmful emails The growth of spam emails is a significant challenge because it introduces phishing malware. Introduce the risk of attacks and wasted valuable time, as email is a primary form of communication Traditional methods of filtering spam through pre-defined rules have become ineffective when spammers develop their methods using modern techniques such as machine learning (ML).

In this study, we have highlighted how Naive Bayes, Random Forest and Gradient Boosting techniques can improve spam classification. Considering the primary metadata of the email, this model is able to prevent spam with greater and greater accuracy. Tokenization, stopword extraction, stemming, and vectorization as preprocessing methods focused on important features to be extracted. The addition of root such as sender behavior and email header also supports the functionality of any search engine. Random forests and growth rates in particular have achieved a fine trade-off between accuracy and speed that has led to an advantage in optimizing spam detection algorithms.

It was also observed that now, there is increased interest in more sophisticated models e.g., CNN and transformer based models (BERT, GPT deeplearning) to increase detection rate but those models have challenges in terms of computing power requirements and data amount in and this limits use to small systems.

Consequently, although looking at traditional machine learning devices is a viable solution for most organizations or individuals on a budget, lowcost equipment is not it is easy to use.

The analysis identified the strengths and weaknesses of different models in terms of accuracy, precision, recall, F1-score, and other metrics with random forests being better at recall but none better at motion smoothness in the Neve Bayes forest model or performs universally better than others . Instead, the best model depends on several factors, including the amount of data and the need to limit false positives. In order to advance meaningful research for the future, it is important to integrate different technical approaches to enable the development of new areas and new approaches to reduce the need for experimental design. They are even more adept at spearing deeply on smaller systems although there are limitations and they are particularly promising for exploring and developing hybrid systems using deep learning and machine learning techniques works well.

Ultimately, this study confirms that machine learning is an extremely useful technology that can be used to help increase the security of emails, as well as strategically manage spam.
This unique approach taken in this study supports the need to build this research by incorporating various methods towards intelligent spam detection and can be expanded in the future Due to the continuous changes in communication technology, policies should also be redesigned to protect email users and ensure they are not abusiveUltimately, this study confirms that machine learning is an extremely useful technology that can be used to help increase the security of emails, as well as strategically manage spam .This unique approach taken in this study supports the need to build this research by incorporating various methods towards intelligent spam detection and can be expanded in the future Due to the continuous changes in communication technology, policies should also be redesigned to protect email users and ensure they are not abused.

REFERENCES

1. Metsis , V. , Androutsopoulos , I. , & Palioras , G. (2006). Naive Bayes cum Spam Filtering - Who is Naive Bayes?. CEAS.
2. Hidalgo, J. M. G. (2002). Evaluation of cost-effective unwanted bulk email classification. Proceedings of the International Conference on Machine Learning and Online Learning.

3. Bhowmik, K. (2015). A comparative analysis of machine learning algorithms for spam email classification. International Journal of Computer Science and Information Security.
4. Kim, Y. (2014). Convolutional neural networks for sentence segmentation. EMNLP is the operator.
5. Yang , Z. , Yang , D. , Dyer , C. , et al. (2016) and its results. Hierarchical attentional networks for document classification. Proceedings of the Annual Meeting of the Association for Computational Linguistics.
6. Sahmi , M. , Dumais , S. , Heckerman , D. , & Horwitz , E. (2006).
   (1998) and the author. A Bayesian approach to junk e-mail filtering.
   Proceedings of the AAAI-98 Workshop on Teaching for Text Classification.
7. Amin, A., Anwar, S., Nawaz, M., and Wahid, S. (2017). Spam detection by ensemble classification. Journal of computer science.
8. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Burt: Pretraining deep bidirectional transformers for speech understanding. Proceedings of the NAACL-HLT.
9. Wu, L., Liu, S., & Yang, Y. (2020). A comprehensive review on deep learning-based spam detection. IEEE Access.
10. Carreras , X. , & Marquez , L. (2001). Growing trees for anti-spam email filtering. ACM SIGKDD Research Journal, 3(2), 65-76.
11. Bratko , A. , Filipich , B. , Cormack , G. V. , Lynam , T. R. , & Zupan , B. (2006). Spam filtering using statistical data compression models. Journal of Machine Learning Research, 7 (Dec), 2673-2698.
12. Androutsopoulos , I. , Koutsias , J. , Chandrinos , K. V. , & Spyropoulos , C. D. (2000). Comparative testing of naive Bayesian and keyword-based anti-spam filtering and personalized e-mail messaging. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.
13. Zhou, L., & Chaovlit, P. (2008). An empirical study of sentiment analysis for spam filtering. Journal of Organizational Computers & Electronic Commerce, 19(1), 1-23.
14. Nana , E. G. , Bassi , J. S. , Chiroma , H. , Abdulhamid , S. M. , and Abubakar , A. I. . (2019) no. Machine learning for email spam filtering: Research, conceptualization, and open research problems.
    Helion, 5(6), e01802.