IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

"A Machine Learning Approach To Spam Email Classification"

Shalu Pal
Computer Science & Engineering (Mewar University)

Abstruct- One of the most significant problems in the online world is spam emails. Spam emails irritate individual email users in addition to having a negative financial impact on organizations. By applying the two machine learning algorithms—Naïve Bayes, Support Vector Machines, and Decision Trees—for the detection of spam emails, this research seeks to suggest a hybrid bagging strategy based on machine learning. The dataset is split up into many sets and supplied as input to each algorithm in this procedure. Three trials are conducted in all, and the outcomes are compared with respect to the spam filter's precision, recall, accuracy, efficiency, and suitability. Separate Naïve Bayes and decision tree methods are used for the two studies. The third experiment uses a hybrid bagged technique to construct the suggested SMD system. The hybrid bagged approach based SMD system attained an overall accuracy of 87.5%.

Key words: Correlation based feature selection, Spam filtering, Hybrid bagged approach, J48 & DT algorithm, Naïve bayes, Text mining.

1. INTRODUCTION

Unwanted, uninvited, and frequently unrelated emails sent in mass to a large number of recipients are known as spam. Although they are usually delivered for business-related reasons, such promoting goods or services, they can also be used maliciously, like phishing or malware distribution. In essence, they are junk mail on the internet.

One of the most popular and efficient forms of communication is the email system. The cost-effectiveness and speed of email communication are the main factors contributing to its widespread use. Unfortunately, spam emails are posing a threat to the email system. Uninvited emails sent by unsanctioned users, commonly referred to as spammers, with the intention of profiting are known as spam emails. Sorting through these spam emails takes up the majority of the email users' valuable time . Multiple copies of the same communication are delivered repeatedly, which annoys the recipient in addition to having a negative financial impact on the company.

Spam emails not only invade users' inboxes but also generate significant amounts of unnecessary data, consequently impacting the network's capacity and utilization. This paper presents a Spam Mail Detection (SMD) system designed to categorize email data into spam and legitimate messages. Spam filtering operates on three primary levels: the sender's email address, the email subject, and the message content. Every email follows a universal format, which includes the email's subject and its main content. A common spam email can be identified by analyzing its content. The method of identifying spam emails relies on the premise that the content of spam differs from that of legitimate or Ham (non-spam) emails.

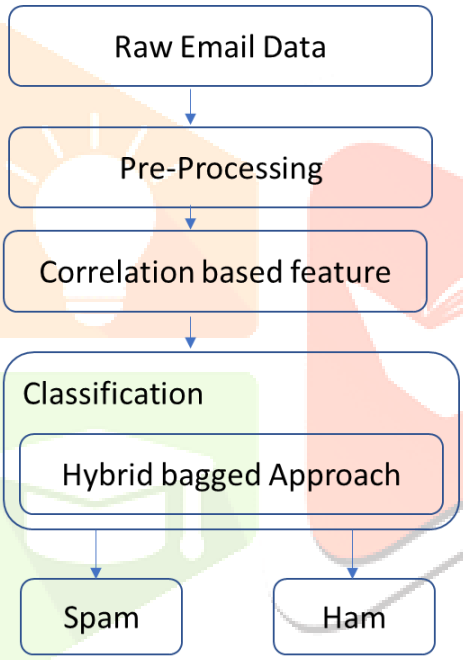
For instance, terms associated with product advertising, service endorsements, dating content, etc. Spam email detection can generally be divided into two methods: knowledge engineering and machine learning approaches. Knowledge engineering utilizes a network-oriented method where the IP (internet protocol) address, network address, and a specified set of rules are taken into account for classifying emails. The method has demonstrated encouraging outcomes, yet it requires a significant amount of time

The upkeep and responsibility of refreshing rules is not suitable for every user. Conversely, the machine learning method does not rely on predefined rules and is more effective than the knowledge engineering

method. The classification algorithm categorizes the email according to its content and various other features. In many classification issues, the steps of feature extraction and selection hold significant importance. Attributes are essential in the classification process. This study employs a correlation based feature selection (CFS) technique for feature extraction. The CFS method identifies the optimal features from the feature set for effective classification outcomes. To address the limitations of the existing model, a new hybrid bagging approach is presented in the suggested spam mail detection (SMD) system.

The suggested spam email detection system is motivated by the efficiency of the machine learning method. In a spam email detection system, the first step involves gathering email data. The gathered email data is unrefined and lacks structure. To minimize computations and achieve precise outcomes, it is essential to pre-process email data. The data undergoes pre-processing by eliminating stop words, conducting stemming, and performing word tokenization to extract useful information. Subsequently, correlation-based feature selection (CFS) is executed to identify the most suitable features from the feature pool. The data dimensionality is reduced during the pre-processing phase, and features are subsequently extracted in the form of a bag of words.

A bagged hybrid method (which combines Naïve Bayes classifier and J48) is employed for classification to enhance strength and accuracy. The dataset is split randomly into various groups and acts as input for each classification algorithm. The bagging method merges the classification outcomes of the two machine learning algorithms to assess the ultimate classification result



The remaining part of the paper is organized as follows: Section II presents the contributions of authors pertaining to email filtering. Section III contains the initial details of the suggested model. In section IV, the components of the suggested Spam Mail Detection are detailed with a functioning example. Section V presents the computed experimental results, while Section VI wraps up the paper

Spam email detection mainly depends on machine learning and natural language processing methods to differentiate between unsolicited emails (spam) and legitimate ones (ham), employing principles such as feature extraction, classification techniques, and assessment metrics. Important concepts encompass: spam (unsolicited mass emails, frequently commercial or harmful), ham (valid emails), feature extraction (determining significant email traits such as keywords, sender details, and format), pre-processing (sanitizing and arranging email data), and classification algorithms (e.g., Naive Bayes, Support Vector Machines, Decision Trees) that learn from annotated data to differentiate spam from ham.

2. Key Concepts and Terminologies:

- **Spam vs. Ham:**

Spam denotes unwanted and typically mass emails, whereas ham signifies genuine, welcome email

- **Machine Learning (ML):**

A fundamental method for spam detection, in which algorithms identify patterns from data to categorize emails without established guidelines

- **Natural Language Processing (NLP):**

Methods employed to allow machines to comprehend and handle human language in emails, essential for content evaluation

- **Feature Extraction:**

The task of recognizing and choosing key features from emails to serve as input for classification models. Esto incluye:

• Keywords: Terms commonly associated with spam, including those pertaining to marketing, financial scams, or dating platforms.

• Email Headers: Data such as IP addresses, sender information, and routing details, which may reveal fraudulent origins.

• Email Format and Data: The size of emails, existence of attachments, or particular formatting may serve as indicators.

- **Preprocessing:**

Actions performed to ready email data for analysis, including tokenization (dividing text into terms), stemming (simplifying words to their base form), and stop word removal (discarding frequently used words like "the" or "a").

- **Term Frequency-Inverse Document Frequency (TF-IDF):**

A common method for feature weighting that reflects how important a word is to a document in a collection..

3. METHODOLOGY

The Research Methodology in the development of a spam email detection system using machine learning algorithms

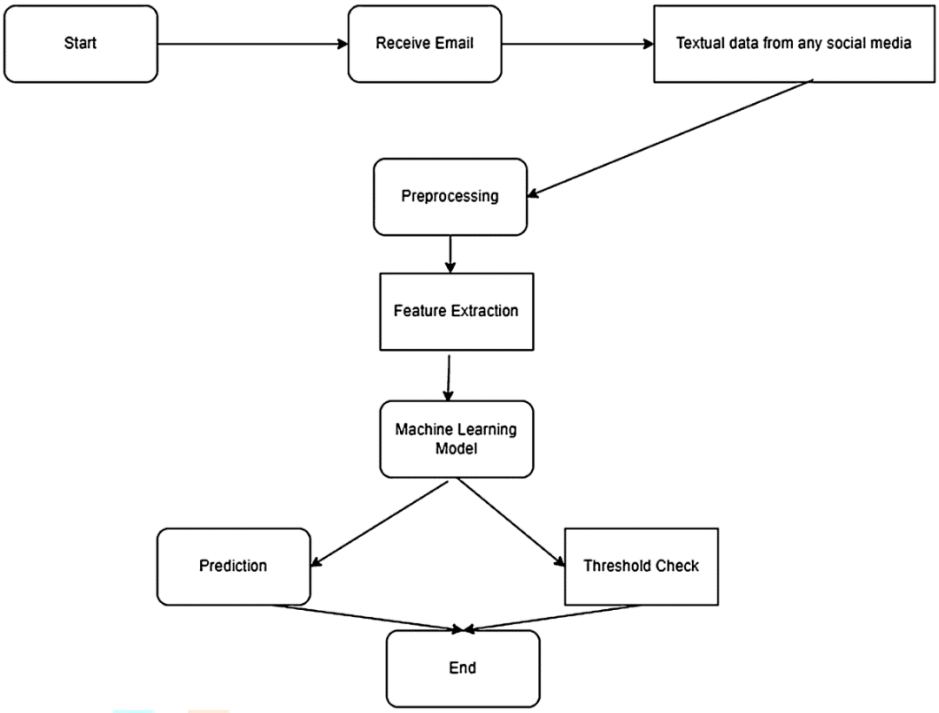
- **Naive Bayes:** . A probabilistic classifier that assumes independence between features. It's often used in spam detection due to its simplicity and efficiency.

- **Support Vector Machines (SVM):** A powerful algorithm that finds an optimal hyperplane to separate data points into classes..

- **Decision Trees (DT) & J48 Algorithm:** Tree-like structures that represent decision rules for classification.

- **K-Nearest Neighbors (KNN):** Classifies emails based on the majority class of their nearest neighbors in the feature space.

- **Artificial Neural Networks (ANN):** Can be used, sometimes in combination with optimization techniques like Particle Swarm Optimization (PSO).



ML based spam email detector

4. Evaluation Metrics:

A model's performance is evaluated using a number of evaluation metrics in machine learning for spam email detection. Accuracy, precision, recall, F1-score, and AUC-ROC are important metrics. These metrics, which take into account both correctly identified instances and misclassifications, measure how well the model differentiates between spam and legitimate emails, or "ham" and spam.

Below is a summary of these metrics:

1. Accuracy: This evaluates the overall accuracy of the model's predictions, reflecting the percentage of accurately classified emails (spam and ham) relative to the total emails

- **Formula:** Accuracy = (TP + TN) / (TP + TN + FP + FN)

2. Precision: Precision emphasizes the correctness of positive forecasts. It specifically calculates the percentage of emails identified as spam that are genuinely spam. A high precision indicates the model effectively reduces false positives (accurately detecting spam)

- **Formula:** Precision = TP / (TP + FP)

3. Recall: Recall, referred to as sensitivity, assesses the model's capability to detect all genuine spam emails. It determines the ratio of genuine spam emails that are accurately recognized as spam. A high recall suggests the model effectively reduces false negatives (failing to identify real spam)

- **Formula:** Recall = TP / (TP + FN)

4. F1-Score: The F1-score offers a proportionate assessment of precision and recall, reflecting their harmonic average. It's beneficial in situations where class distribution is imbalanced (more spam or more ham) or when precision and recall hold equal significance

- **Formula:** F1-Score = 2 * (Precision * Recall) / (Precision + Recall)

Where:

- TP = True Positives (spam emails correctly identified as spam)
- TN = True Negatives (legitimate emails correctly identified as legitimate)
- FP = False Positives (legitimate emails incorrectly classified as spam)
- FN = False Negatives (spam emails incorrectly classified as legitimate)

5. Conclusion

The detection of spam emails using machine learning has proven to be an effective and scalable approach in addressing the growing challenges posed by unsolicited and malicious email content. By leveraging various ML algorithms such as Naive Bayes, Support Vector Machines, Random Forests, and deep learning models, spam filters have significantly improved in accuracy, adaptability, and efficiency compared to traditional rule-based systems.

This project demonstrates how machine learning can analyze and classify large volumes of email data based on features such as word frequency, sender behavior, and message patterns. The ability of ML models to learn from data and adapt to new spam tactics makes them highly suitable for dynamic environments like email systems.

However, while current solutions offer substantial benefits, continuous refinement is necessary to stay ahead of evolving spam techniques. Ensuring low false positives, handling adversarial examples, and preserving user privacy are ongoing challenges that require future exploration.

In conclusion, machine learning provides a powerful foundation for building robust spam email detection systems. With ongoing research, real-time processing improvements, and integration with advanced security frameworks, ML-powered spam detectors can offer even greater protection and efficiency in the future.

6. Future Scope

The development of spam email detectors using machine learning (ML) has made significant progress in improving the accuracy and efficiency of spam filtering. However, the ever-evolving tactics of spammers necessitate ongoing research and enhancement of current models. The future scope of spam email detection using ML can be outlined as follows:

1. **Integration of Deep Learning Techniques**
   Future systems can leverage advanced deep learning models such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformers (e.g., BERT) for improved understanding of the semantic content of emails. These models are particularly effective at detecting sophisticated phishing attacks and obfuscated spam messages.
2. **Real-Time Spam Detection**
   Optimizing models for real-time inference and low-latency detection can enhance user experience and security. This involves lightweight model architectures or efficient deployment using edge computing or serverless cloud functions.
3. **Multi-Language Spam Detection**
   Most spam filters are currently optimized for English. Expanding the scope to support multilingual spam detection will help serve a global audience more effectively, particularly in regions with high internet usage but limited local-language spam protection.
4. **Adversarial Attack Resistance**
   Spammers increasingly use adversarial techniques to evade detection. Future models must incorporate adversarial training and robust feature engineering to counteract these methods and improve model resilience.

References

1. Priti Sharma, Uma Bhardwaj 'Machine Learning based Spam E-Mail Detection' International Journal of Intelligent Engineering and Systems, Vol.11, No.3, 2018
2. Isra'a AbdulNabi, Qussai Yaseen' Spam Email Detection Using Deep Learning Techniques' Volume 184, 2021, Pages 853-858
3. Nikhil Kumar; Sanket Sonowal; Nishant 'Email Spam Detection Using Machine Learning Algorithms' IEEE 15-17 July 2020
4. Zeeshan Bin Siddique, Mudassar Ali Khan, Ikram Ud Din, Ahmad Almogren, Irfan Mohiuddin, Shah Nazir' Machine Learning-Based Detection of Spam Emails' 2021
5. W.A. Awad, S.M. ELseuofi' Machine LearningMethods for SpamE-MailClassification' International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011
6. Nikhil Kumar, Sanket Sonowal, Nishant 'Email Spam Detection Using Machine Learning Algorithms' July 2020 IEEE
7. TA Almeida, JMG Hidalgo, A Yamakami "Contributions to the study of SMS spam filtering: new collection and results" Proceedings of the 11th ACM symposium on Document engineering, 2011
8. B Biggio, G Fumera, F Roli "Security evaluation of pattern classifiers under attack" - IEEE transactions on knowledge,2013
9. E Blanzieri, A Bryl "A survey of learning-based techniques of email spam filtering" - Artificial Intelligence Review, 2008
10. T Ellman "Explanation-based learning: A survey of programs and perspectives" – ACM Computing Surveys (CSUR), 1989
11. TS Guzella, WM Caminhas" A review of machine learning approaches to spam filtering" Expert Systems with Applications, 2009
12. Sarah Jane Delany [a], Mark Buckley [b], Derek Greene" SMS spam filtering: Methods and data" Volume 39, Issue 10, August 2012, Pages 9899-9908
13. Shafi'I Muhammad Abdulhamid; Muhammad Shafie Abd Latiff; Haruna Chiroma; Oluwafemi Osho; Gaddafi Abdul-Salaam; Adamu I. Abubakar" A Review on Mobile SMS Spam Filtering Techniques" Volume 5, IEEE ,February 2017 15650 – 15666
14. Tiago Almeida José María Hidalgo Tiago Silva "Towards SMS Spam Filtering: Results under a New Dataset" Year 2013, Volume: 2 Issue: 1, 1 - 18, 31.03.2013
15. W Li, GG Wang, AH Gandomi "A survey of learning-based intelligent optimization algorithms" Archives of Computational Methods in …, 2021
16. Kuldeep Yadav;Swetank K. Saha;Ponnurangam Kumaraguru;Rohit Kumra" Take Control of Your SMSes: Designing an Usable Spam SMS Filtering System" 2012 IEEE 13th International Conference
17. Shaghayegh Hosseinpour;Mohammad Reza Keyvanpour" A Comprehensive Approach to SMS Spam Filtering Integrating Embedded and Statistical Features" 2023 13th International Conference on Computer and Knowledge Engineering (ICCKE),IEEE
18. Pavas Navaney;Gaurav Dubey;Ajay Rana" SMS Spam Filtering Using Supervised Machine Learning Algorithms" 2018 8th International Conference on Cloud Computing,IEEE
19. Hyun-Young Lee;Seung-Shik Kang" Word Embedding Method of SMS Messages for Spam Message Filtering" 2019 IEEE International Conference
20. Agustinus Theodorus;Tio Kristian Prasetyo;Reynaldi Hartono;Derwin Suhartono" Short Message Service (SMS) Spam Filtering using Machine Learning" 2021,IEEE