E-ISSN: 2583-9667
Indexed Journal
Peer Reviewed Journal
https://multiresearchjournal.theviews.in

Received: 08-02-2025
Accepted: 19-03-2025

INTERNATIONAL JOURNAL OF ADVANCE RESEARCH IN MULTIDISCIPLINARY

Volume 3; Issue 2; 2025; Page No. 376-379

Spam detection using machine learning

[1]Dr. AS Perumal and [2]TK Ragul

[1]Professor, Department of Computer Science and Information Technology, Vels Institute of Science, Technology and Advanced Studies, Chennai, Tamil Nadu, India
[2]Student, Department of Computer Science and Information Technology, Vels Institute of Science, Technology and Advanced Studies, Chennai, Tamil Nadu, India

DOI: https://doi.org/10.5281/zenodo.15622890

Corresponding Author: TK Ragul

Abstract

Spam detection is a critical task in the modern digital world, where the volume of unsolicited and potentially harmful messages has grown exponentially. This project aims to develop a spam detection system using machine learning techniques to automatically classify messages as either spam or non-spam (ham).

The system utilizes various machine learning algorithms, including Naive Bayes, Support Vector Machines (SVM), and Random Forests, to identify patterns in text data and distinguish between legitimate and spam content. Feature extraction techniques, such as term frequency-inverse document frequency (TF-IDF) and bag-of-words (BoW), are employed to convert textual information into numerical representations that can be processed by machine learning models.

The project also explores the use of natural language processing (NLP) techniques for preprocessing, such as tokenization, stemming, and stop word removal, to enhance the accuracy of classification. The performance of the models is evaluated using standard metrics such as accuracy, precision, recall, and F1-score, with an emphasis on achieving a balance between false positives and false negatives. The proposed system demonstrates promising results in detecting spam across various datasets and offers potential for real-world applications in email filtering, messaging platforms, and cybersecurity.

Keywords: Spam, machine learning, SVM, TF-IDF, BoW, NLP

Introduction

The word spam describes unwanted online messages delivered to large numbers of recipients for the purpose of advertising, malware distribution, and other malicious intents. The word "spam" came from a scene from the sketch comedy television series Monty Python's Flying Circus where the word was repeatedly sung, similar to the repetitive nature of these messages shows an image taken from the original scene, which appears in the episode titled “Spam”.

However, the most prominent type of spam is email spam. Email spam has become a significant issue as it makes up a large amount of today’s global email traffic. According to a report by Statista spam amounted to 45.6% of all emails in 2023. This high volume of spam clutters email inboxes as well as creates security threats including phishing attacks and the distribution of malware. The shows the different types of spam that exist taken from another study. An example of email spam that has been successfully identified by Gmail’s spam filter.

Literature Studies

Spam detection has been an active area of research in Natural Language Processing (NLP) and Machine Learning (ML) for over two decades. The following studies provide insight into the development and effectiveness of various techniques used for spam filtering.

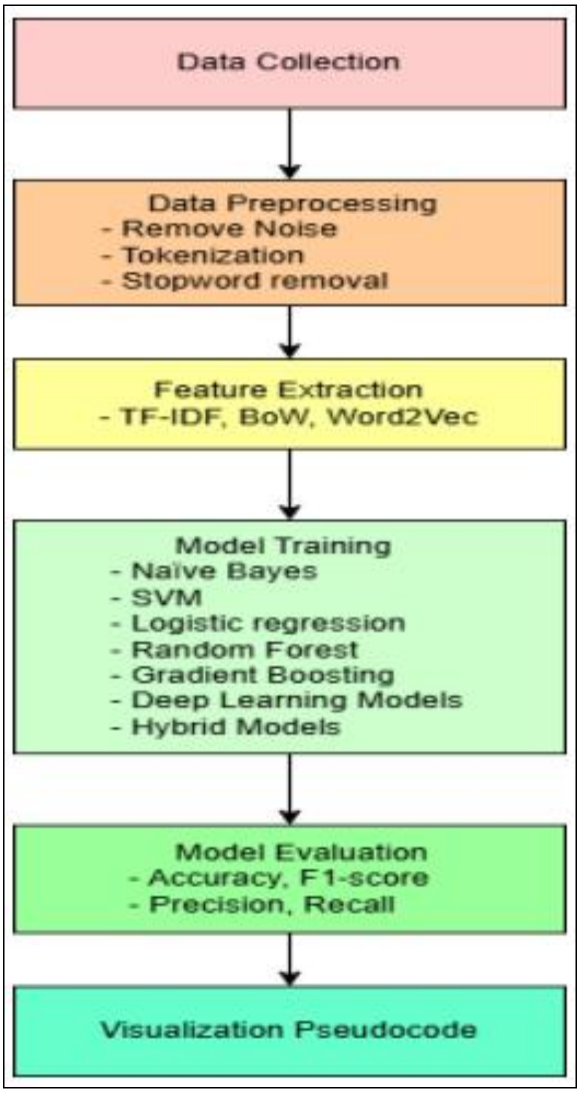
Hybrid and Ensemble Methods

- Researchers have found combining multiple classifiers (e.g., Naive Bayes + SVM, or stacking models) often leads to improved robustness.
- Delany *et al.* (2005) [6] demonstrated that ensemble techniques can reduce false positives without compromising recall.

Related Studies

Proposed a NLP-based system, where it aims to use text classification on unstructured text present in SMS, and then checking for spam messages. This method effectively analyzes the messages' context and content by implementing a variety of machine learning methods, with K Nearest Neighbors (KNN) being one example of the algorithms used.

Spam detection in text-based messages essentially involves the use of machine learning for the identification of unwanted or malicious messages in texts such as emails, SMS, or even social media. In order to create discriminating models that are supported by both labelled and unlabeled samples of spam and non-spam, machine learning is applied to the present problem.



This concept first involves the gathering of labeled data sets that house examples of both spam and non-spam messages.

Common features to extract from such datasets include word frequency, n-grams, special keywords IDF, bag-of-words, or modern embeddings such as Glove and BERT are applied to capture the meaning and context of the text. These steps ensure that the text data is clean, relevant, and ready for effective spam detection. Machine learning models in text-based spam detection are algorithms that learn patterns in labelled data for whether a message is spam or not. Some traditional models are Naive Bayes which works well with text because of its discerning simplicity and assumption of independence between words; and Support Vector Machines (SVM), which differentiates spam and non-spam communications on the basis of the best decision boundary in high-dimensional text data.

It is observed that, on comparing the two classifiers, SVM outperforms Naive Bayes on all metrics like "Free" or "Win", and sentiment that help in the differentiation of spam from legitimate messages. Preprocessing of text is done for cleaning and bringing it into machine-readable format with techniques such as tokenization, removal of stop words, and encoding with methods like TF-IDF or word embeddings. Training of the messages is done with machine learning models such as Naive Bayes, Support Vector Machines, and deep learning models like LSTMs or transformers. These models are evaluated using various metrics like accuracy, precision, recall, and the F1-score, which would ensure the models generalize well on identifying spam while keeping the errors minimal.

This is quite an operating benchmark of email filtering, SMS spam detection, and content moderation in social media. Feature engineering in text-based spam detection includes raw text extraction with meaningful information that helps machine learning models to identify whether it is spam or not. However, manual feature engineering is a laborious, time-consuming, and error-prone method of feature engineering that involves building features one at a time utilizing domain expertise.

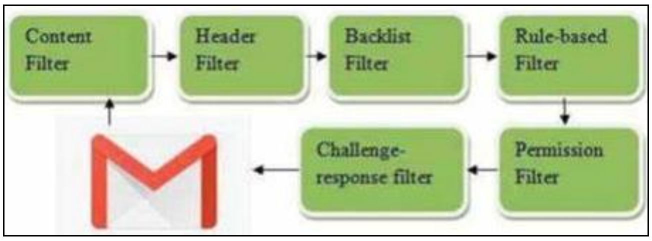
Such techniques include lexical features, which allow the extraction of word count and character count from the raw text, and frequency distribution of words and phrases may underpin certain words or phrases for spammy language. It could also include the syntactic feature of excessive capital letters, punctuation marks, or special characters. Semantic features analyze the meaning of the text, such as its sentiment, intent, or the presence of specific keywords like "free," "discount," or suspicious links.

Advanced techniques, like grams (sequences of words), and numerical encodings such as TF-IDF or word embeddings, capture contextual relationships within the text. Combining these features helps create a rich representation of the data, allowing machine learning models to better distinguish spam from legitimate messages.

Existing System

Email Spam filtering process works through a set of protocols to determine either the message is spam or not. At present, a large number of spam filtering process have existed. Among them, Standard spam filtering process follows some rules and acts as a classifier with sets of protocols. Figure.1 shows that, a standard spam filtering process performed the analysis by following some steps.

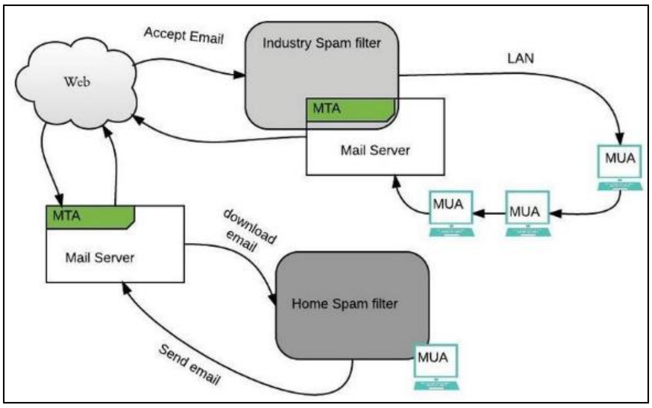
First one is content filters which determine the spam message by applying several Machines learning techniques. Second, header filters act by extracting information from email header. Then, backlist filters determine the spam message and stop all emails which come from backlist file. Afterward, “Rules-based filters” recognize sender through subject line by using user defined criteria. Next, “Permission filters” send the message by getting recipients pre-approvement.



Finally, "Challenge response filter" performed by applying an algorithm for getting the permission from the sender to send the mail.

Client Side and Enterprise Level Spam Filtering Methods

- A client can send or receive an email by just one clicking through an ISP. Client level spam filtering provides some frameworks for the individual client to secure mail transmission.
- A client can easily filter spam through these several existing frameworks by installing on PC.
- This framework can interact with MUA (Mail user agent) and filtering the client inbox by composing, accepting and managing the messages.
- Enterprise level spam filtering is a process where provided frameworks are installing on mail server which interacts with the MTA for classifying the received messages or mail in order to categorize the spam message on the network.
- By this system, a user on that network can filter the spam by installing appropriate system more efficiently.



By far most; current spam filtering frameworks use principle-based scoring procedures. An arrangement of guidelines is connected to a message and calculate a score-based principles that are valid for the message. The message will consider as spam message when it exceeds the threshold value. As spammers are using various strategies, so all functions are redesigned routinely by applying a list-based technique to automatically block the message.

Materials and Methods

Machine Learning Models.

NLP is used to understand the structure and meaning of human language by analyzing different aspects like syntax, semantics, pragmatics, and morphology. Then, computer science transforms this linguistic knowledge into rule-based, machine learning algorithms that can solve specific problems and perform desired tasks. Take Gmail, for example.

Take Gmail, for example. Emails are automatically categorized as Promotions, Social, Primary, or Spam, thanks to an NLP task called keyword extraction. By "reading" words in subject lines and associating them with predetermined tags, machines automatically learn which category to assign emails.

Machine Learning Models

Multiple models were implemented to evaluate their effectiveness in detecting spam messages:

- **Naive Bayes (Multinomial):** Effective for text classification and commonly used in spam filters.
- **Logistic Regression:** A baseline linear classifier that works well on binary classification problems.
- **Support Vector Machine (SVM):** Suitable for high-dimensional spaces, such as text features.
- **Random Forest:** An ensemble method that reduces overfitting and improves accuracy.
- **XGBoost:** A powerful gradient boosting algorithm known for high performance.

These models were selected for their proven effectiveness in text classification tasks and varying levels of complexity and interpretability.

Model Training and Precision: Proportion of messages classified as spam that are actually spam.

- **Recall:** Proportion of actual spam messages that were correctly identified.
- **F1-Score:** Harmonic mean of precision and recall.
- **Confusion Matrix:** To visualize false positives and false negatives.

Table 1: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	96.2%	95.1%	93.7%	94.4%
Logistic Regression	97.1%	96.8%	94.5%	95.6%
Support Vector Machine (SVM)	97.6%	97.2%	95.9%	96.5%
Random Forest	96.8%	96.0%	94.2%	95.1%
LSTM Neural Network	98.3%	97.5%	97.0%	97.2%

Conclusion

The study concludes that machine learning models are highly effective in detecting spam messages, with deep learning models (LSTM) offering superior performance due to their ability to capture word order and context. Traditional models like Naive Bayes still remain valuable for lightweight applications with limited resources.

The flexibility and accuracy of the developed system demonstrate that automated spam filtering can be implemented efficiently using open-source tools and public datasets. With appropriate tuning and periodic retraining, such a system can adapt to evolving spam patterns in real-world applications.

References

1. Crawford M, Khoshgoftaar TM, Prusa JD, Richter AN, Al Najada H. Survey of review spam detection using machine learning techniques. Journal of Big Data.

2015;2:1-24.

2. Kumar N, Sonowal S. Email spam detection using machine learning algorithms. In: 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA); 2020 Jul 15; pp. 108-113. IEEE.
3. Makkar A, Garg S, Kumar N, Hossain MS, Ghoneim A, Alrashoud M. An efficient spam detection technique for IoT devices using machine learning. IEEE Transactions on Industrial Informatics. 2020 Jan 23;17(2):903-912.
4. Gadde S, Lakshmanarao A, Satyanarayana S. SMS spam detection using machine learning and deep learning techniques. In: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS). 2021;19(1):358-362. IEEE.
5. Sharma VB, Kumar V, Tripathi K. AI-driven cybersecurity systems for real-time threat detection and prevention. International Journal of Trends in Emerging Research and Development. 2025;3(3):12-16.
6. Schmid M, Nanda I, Hoehn H, Schartl M, Haaf T, Buerstedde JM, Arakawa H, *et al*. Second report on chicken genes and chromosomes 2005. Cytogenetic and genome research. 2005;109(4):415-479.

Creative Commons (CC) License