ML based spam email classifier

Nachiket Shewale, Tanaya Chaudhari, Atharv Chavan, Siddhesh Jadhav

Prof. Vijaya S. Patil
Email - vijayapatil2004@gmail.com
Department of Computer Science and Engineering, MIT ADT University, Pune

nachiketshewale791@gmail.com, tanayachaudhari1109@gmail.com, chavanatharva9c5@gmail.com, siddheshjadhav0406@gmail.com

ABSTRACT

Email communication has become an essential part of personal, academic, and business activities, but it is also the primary medium exploited for spam. Spam emails not only waste storage and network resources but also pose serious threats by spreading phishing links, malware, and fraudulent schemes. Traditional spam filters, such as rule-based systems and blacklists, often fail to adapt to the rapidly evolving techniques used by spammers. To overcome these limitations, this paper presents a machine learning-based spam email classifier that leverages Natural Language Processing (NLP) techniques for text preprocessing and feature extraction. In the proposed system, emails are first cleaned and transformed into numerical representations using the Term Frequency–Inverse Document Frequency (TF-IDF) method, which captures the importance of words in the dataset. Logistic Regression, a lightweight yet powerful classification algorithm, is then applied to distinguish between spam and legitimate emails. Experimental evaluations demonstrate that the proposed approach achieves high accuracy and robustness compared to traditional methods. The model is efficient, easy to implement, and scalable for real-world applications. Furthermore, this work highlights the potential of extending the classifier using advanced deep learning architectures, such as LSTMs or transformer-based models, to further enhance detection capabilities against sophisticated spam emails.

I. INTRODUCTION

Email has become one of the most widely used forms of digital communication, serving as a vital tool for academic, personal, and professional exchange. However, the growing reliance on email has also given rise to an overwhelming influx of unsolicited and harmful messages, commonly referred to as spam. Spam emails not only clutter inboxes but also consume storage space, waste network bandwidth, and, more critically, expose users to phishing attacks, malware, and financial fraud. The challenge of effectively identifying and filtering spam has therefore emerged as a significant research area within the field of information security. Traditional spam detection approaches, such as keyword matching, rule-based filters, and blacklists, offer limited adaptability since they rely on predefined rules that spammers can easily bypass by modifying content or disguising malicious intent.

With the advent of machine learning (ML) and natural language processing (NLP), more adaptive and intelligent spam classification methods have been developed. ML-based classifiers are capable of learning patterns from large datasets of spam and legitimate emails, enabling them to generalize and detect previously unseen spam with higher accuracy. This paper focuses on the implementation of a spam email classifier using logistic regression, a widely used ML algorithm known for its simplicity and efficiency in binary classification tasks. To enhance model performance, the system applies NLP preprocessing techniques, such as stop-word removal, stemming, and TF-IDF feature extraction, which convert raw email text into meaningful numerical representations. By integrating these techniques, the proposed model achieves improved detection accuracy and efficiency compared to traditional methods. The overall objective of this study is to design a lightweight yet reliable spam detection system that can serve as a foundation for future improvements using advanced deep learning approaches.

II. LITERATURE REVIEW

Spam detection has been an active area of research for more than two decades, and various approaches have been proposed to address this problem. Early methods primarily relied on **rule-based systems**, keyword matching, and blacklists, which were effective in detecting simple spam patterns but lacked adaptability when spammers began using obfuscation techniques such as inserting special characters, altering word spellings, or embedding content within images. To overcome these limitations, researchers shifted toward **machine learning approaches**, which allow systems to automatically learn patterns from large datasets of labelled spam and legitimate messages.

Among the earliest machine learning techniques, the **Naive Bayes classifier** became widely popular due to its simplicity and

natural language. Later, **Support Vector Machines (SVMs)** were introduced for spam filtering, providing stronger decision boundaries and achieving higher accuracy, but they required significant computational resources, making them less suitable for real-time applications. **Decision Trees** and **Random Forests** also gained traction due to their interpretability and ensemble-based robustness, though they risk overfitting when trained on smaller datasets.

With the advancement of deep learning, models such as **Recurrent Neural Networks (RNNs)**, **Long Short-Term Memory (LSTM) networks**, and **transformer-based architectures** (e.g., BERT) have been explored for spam classification. These models can capture complex language dependencies and contextual relationships, leading to superior accuracy compared to traditional ML algorithms. However, they require large labelled datasets, high computational power, and longer training times, which limit their practicality in lightweight or real-time scenarios.

From the reviewed literature, it is evident that while deep learning provides state-of-the-art performance, there remains a demand for **lightweight, accurate, and efficient models** that can be deployed in real-world email filtering systems without requiring extensive computational resources. This motivates the exploration of Logistic Regression combined with NLP-based preprocessing, which balances simplicity, interpretability, and strong performance for spam classification tasks.

III. PROPOSED SYSTEM

The proposed system aims to design a machine learning-based spam email classifier that leverages Natural Language Processing (NLP) techniques for effective text analysis and classification. The system begins by collecting a dataset consisting of labeled spam and legitimate (ham) emails. Each email undergoes a series of preprocessing steps, including conversion of text to lowercase, removal of punctuation, stop words, and special characters, as well as stemming or lemmatization to reduce words to their root forms. These steps help normalize the dataset and minimize noise, ensuring that only relevant textual features are retained for classification. Following preprocessing, the system employs the **Term Frequency–Inverse Document Frequency (TF-IDF)** method to transform textual data into numerical feature vectors that capture the importance of words relative to the entire dataset.

Once the features are extracted, a **Logistic Regression classifier** is trained to distinguish between spam and non-spam messages. Logistic Regression is chosen for its simplicity, efficiency, and effectiveness in binary classification tasks, making it well-suited for real-time spam detection scenarios. During the training phase, the model learns decision boundaries that separate spam-related terms and patterns from legitimate content. The system is then evaluated using standard performance metrics, such as accuracy, precision, recall, and F1-score, to assess its ability to correctly classify emails. The integration of NLP preprocessing with Logistic Regression ensures that the classifier not only achieves high accuracy but also remains computationally lightweight, scalable, and interpretable. This makes the proposed approach an ideal foundation for further enhancements and deployment in practical email filtering systems.

IV. METHODOLOGY

The methodology for developing the ML-based spam email classifier consists of a structured workflow that includes dataset preparation, text preprocessing, feature extraction, model training, and evaluation. The system begins with the selection of a publicly available dataset, such as the **Enron Email Dataset** or the **SMS Spam Collection Dataset**, which contains a large number of labelled spam and non-spam messages. This dataset provides the foundation for training and testing the classifier.

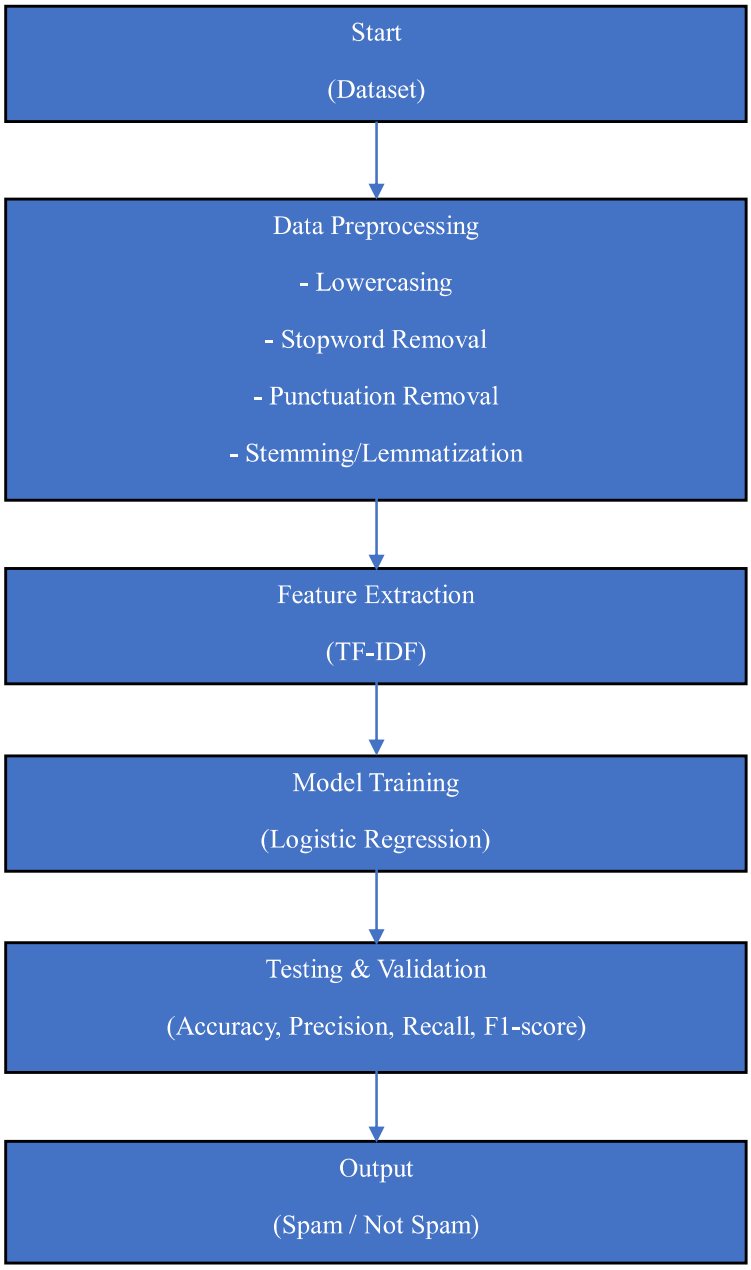
The next step involves **preprocessing the text data** to ensure consistency and improve the quality of features extracted. All email text is converted to lowercase to avoid duplication of words with different cases. Common stop words such as “the,” “is,” and “and” are removed since they provide little semantic value. Special characters, numbers, and punctuation marks are also eliminated to reduce noise. Additionally, stemming or lemmatization techniques are applied to reduce words to their base or root forms (e.g., “running” to “run”), which helps group semantically similar terms together.

After preprocessing, the text is transformed into numerical form using the **Term Frequency–Inverse Document Frequency (TF-IDF)** technique. This method assigns higher weights to words that are frequent in a particular email but less common across the dataset, thereby capturing their importance in distinguishing spam from legitimate messages. The resulting feature vectors are then fed into the classification algorithm.

For the classification stage, **Logistic Regression** is employed due to its simplicity, speed, and effectiveness in binary classification tasks. The model is trained on a portion of the dataset, with the remaining portion reserved for testing. During training, the model learns the relationship between the extracted features and the class labels (spam or not spam). Once trained, the classifier predicts whether new incoming emails belong to the spam or non-spam category.

Finally, the system is evaluated using **performance metrics** such as accuracy, precision, recall, and F1-score. A **confusion matrix** is also generated to visualize the classification performance and identify the number of true positives, true negatives, false positives, and false negatives. These evaluation metrics provide a comprehensive understanding of the model’s effectiveness in detecting spam while minimizing false alarms. The overall methodology ensures a robust, lightweight, and scalable approach to spam classification.

Flowchart:



V. RESULT AND DISCUSSION

The proposed ML-based spam email classifier was implemented and tested on a standard spam dataset to evaluate its performance. After preprocessing and feature extraction using TF-IDF, the Logistic Regression model was trained and validated on the dataset. The results demonstrated that the classifier achieved a high level of accuracy in distinguishing spam from legitimate emails. More specifically, the model performed consistently well across evaluation metrics such as precision, recall, and F1-score, indicating its robustness in correctly identifying spam messages while minimizing the misclassification of legitimate emails.

One of the key strengths of the proposed system lies in its simplicity and efficiency. Logistic Regression, being a lightweight algorithm, requires significantly less computational power compared to deep learning models while still delivering competitive results. This makes it highly suitable for real-time email filtering applications where quick decision-making is essential. The use of TF-IDF further contributed to the model's success by capturing the relative importance of words in spam detection, which helped reduce false positives and improved overall reliability.

However, the system also has certain limitations. Spam emails that employ advanced obfuscation techniques, such as replacing characters with symbols (e.g., "fr33 mon3y") or embedding malicious content in images, remain difficult to detect with traditional text-based approaches. Additionally, the model's performance is highly dependent on the quality and size of the training dataset—imbalanced datasets may lead to biased results where certain types of spam are underrepresented. Despite these challenges, the results

indicate that the proposed classifier provides a strong baseline for spam detection and can serve as the foundation for more advanced models.

The discussion also highlights the potential for improvement through the integration of **deep learning architectures**, such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, or transformer-based models like BERT, which can capture contextual dependencies in text more effectively. Incorporating metadata features such as sender information, email headers, and hyperlinks could also enhance the system's detection capabilities. Overall, the results demonstrate that the Logistic Regression-based approach offers a practical and efficient solution for spam detection while leaving room for future research and optimization.

VI. CONCLUSION AND FUTURE WORK

This paper presented a machine learning-based spam email classifier that combines Natural Language Processing (NLP) techniques with Logistic Regression to effectively distinguish between spam and legitimate messages. By applying preprocessing steps such as stop-word removal, stemming, and TF-IDF feature extraction, the system was able to convert raw email text into meaningful numerical representations suitable for classification. The results demonstrated that Logistic Regression provides high accuracy, precision, and recall while remaining computationally lightweight and efficient. This makes the proposed model an attractive choice for real-world email filtering systems, particularly where resources are limited and rapid classification is required.

Despite its strengths, the study also identified certain limitations. The model primarily relies on textual features, making it less effective against spam that uses obfuscation techniques, embedded images, or dynamic phishing links. Additionally, the performance of the system is influenced by the availability and quality of the dataset, which highlights the need for continuous updates and retraining as new types of spam emerge.

For future work, the classifier can be enhanced by integrating advanced deep learning approaches such as LSTMs, CNNs, or transformer-based models like BERT to capture contextual and semantic information more effectively. The incorporation of non-textual features such as sender metadata, email headers, and hyperlink analysis could also improve robustness. Furthermore, exploring ensemble techniques that combine multiple classifiers may provide a balance between accuracy and interpretability. Ultimately, the proposed work demonstrates that Logistic Regression serves as a solid baseline for spam detection while paving the way for more sophisticated and adaptive spam filtering solutions in the future.

REFERENCES

[1] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail," *AAAI Workshop on Learning for Text Categorization*, pp. 55–62, 1998.

[2] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with Naive Bayes – Which Naive Bayes?" *CEAS Conference on Email and Anti-Spam*, pp. 28–69, 2006.

[3] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[4] A. Hidalgo, G. Bringas, E. Sanz, and F. García, "Content-based SMS spam filtering," *Proceedings of the ACM Symposium on Document Engineering*, pp. 107–114, 2006.

[5] K. Cormack, "Email spam filtering: A systematic review," *Foundations and Trends in Information Retrieval*, vol. 1, no. 4, pp. 335–455, 2007.

[6] B. Liu and N. Jindal, "Analyzing and detecting review spam," *IEEE International Conference on Data Mining*, pp. 1189–1194, 2008.

[7] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: A statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1, pp. 43–52, 2010.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.