International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 03 | Mar 2025 www.irjet.net p-ISSN: 2395-0072

Detecting Spam Email with Machine Learning and Writing Optimized Email

Siddharth More[1], Rajat Pandit[2], Yash Salunke[3], Rohit Jawale[4], Prof. Dr.Radhika Nanda[5]

[1,2,3,4] *B.E. Students Department of Computer Engineering*
[5] *HOD, Department of Computer Engineering, Bharat College of Engineering, Opp. Gajanan Maharaj Temple, Kanhor Road, Badlapur (West), Thane, Maharashtra - 421503*

Abstract – *In today's digital era, emails have become a fundamental mode of communication for both personal and professional purposes. However, with the growing number of email users, the volume of spam emails has also increased. Spam, often referred to as junk mail, consists of unwanted messages sent in bulk to multiple recipients, typically for commercial gain. These emails may include phishing attempts, image-based spam, malware, lottery scams, advertisements, and other unsolicited content.*

Spam not only clutters inboxes with irrelevant messages but also poses security risks. It can slow down internet performance and enable cybercriminals to extract sensitive information, such as personal details, professional contacts, and financial data. As a result, distinguishing spam from legitimate (ham) emails is crucial.

To combat this issue, spam filters are employed to identify and block unwanted, unsolicited, and potentially harmful emails before they reach the inbox. Various machine learning and deep learning techniques are used for spam detection, with some of the most effective being convolutional neural networks (CNN), support vector machines (SVM), and naïve Bayes (NB).

Spam Email, phishing scams, unwanted communications, machine learning, convolution neural network CNN, support vector machine SVM, naïve bayes NB.

1. INTRODUCTION

The Internet has become an integral part of modern society, enabling seamless communication and global connectivity at any time and from any location. One of the most widely used internet-based communication tools is email (electronic mail), which is utilized by students, professionals, business people, and government officials. Since sending emails is typically free, spammers exploit this feature to distribute large volumes of unwanted messages.

Initially, most spam emails contained plain text, however, as text-based spam filters improved by analyzing email headers, body content, and other features, spammers adapted by embedding text within images—a technique known as image spam. They began incorporating backgrounds, noise, and other distortions into images to evade detection further. To counter this, Optical Character Recognition (OCR) technology was developed to extract text from images, allowing traditional text-based filtering techniques, such as Naïve Bayes and other classification methods, to be applied. Despite these advancements, spammers continually evolve their strategies to bypass security measures, making spam detection an ongoing challenge.

Fig. 1: Spam Filter

1.1 Purpose

This discussion focuses on how machine learning can enhance spam email detection while improving email optimization. As email usage continues to grow, spam emails have become a major concern, leading to security threats and overcrowded inboxes. Machine learning techniques, including convolutional neural networks (CNN), support vector machines (SVM), and naïve Bayes (NB), play a crucial role in accurately identifying and filtering spam messages. Furthermore, crafting well-structured and optimized emails helps prevent legitimate messages from being mistakenly marked as spam.

1.1.1 Enhancing Security:

Spam emails frequently include harmful elements like malware, ransomware, or phishing links, which can jeopardize both personal and organizational data security. Implementing effective spam detection and filtering mechanisms is essential to safeguarding users from these potential threats.

1.1.2 Improving Productivity:

A cluttered inbox filled with spam can be overwhelming and time-consuming to manage. Effective spam filtering helps users focus on important emails, reduces distractions, and enhances productivity by ensuring that only relevant messages reach their inboxes.

1.1.3 Conserving Resources:

Spam emails take up storage space and use network bandwidth, putting strain on email servers and systems. Effective spam detection and filtering help minimize this burden, optimizing resource usage and lowering operational costs for individuals and organizations.

1.1.4 Enhancing User Experience:

A spam-free inbox improves usability, enabling users to manage their emails more efficiently without the frustration of sifting through irrelevant messages. Effective spam filtering ensures a smoother and more organized email experience.

1.1.5 Regulatory Compliance:

Different regions have laws mandating organizations to protect personal data and maintain secure communication practices. Implementing robust spam detection helps organizations comply with these regulations by preventing unauthorized access, reducing the risk of data breaches, and ensuring that communication remains secure, trustworthy, and aligned with legal requirements

1.1.7 Reducing Legal Liabilities:

Effective spam detection helps organizations minimize legal risks associated with data breaches, phishing scams, and violations of anti-spam laws. Properly managing spam ensures compliance with regulations, enhances security, and reduces the likelihood of facing legal consequences due to unauthorized access or fraudulent communications.

1.2 Scope

This topic focuses on detecting spam emails using machine-learning techniques and strategies for optimizing email deliverability. It examines the nature of spam, its various types, associated risks, and its impact on both users and organizations.

Key machine learning algorithms, including convolutional neural networks (CNN), support vector machines (SVM), and naïve Bayes (NB), play a crucial role in distinguishing spam from legitimate emails. The discussion also covers essential processes such as dataset collection, preprocessing, feature extraction, and model training to improve spam detection accuracy. Additionally, evaluating and implementing machine learning models is essential for enhancing precision and reducing false positives.

Another important aspect is email optimization, which involves structuring emails effectively to prevent them from being misclassified as spam. The topic also explores emerging trends in AI-driven spam detection, advancements in cybersecurity, regulatory compliance, ethical considerations, and the continuous evolution of machine learning models to strengthen email security and efficiency.

1.3 Aims

13.1 User Education and Awareness

Training Programs: Develop educational initiatives to help users identify and manage spam emails effectively.

Public Awareness Campaigns: Conduct outreach efforts to educate users about the risks of spam and the importance of email security.

1.3.2 Global Collaboration

Information Sharing: Partner with organizations and agencies to exchange insights on spam trends and best practices.

Standardization Efforts: Work towards establishing industry-wide spam detection protocols to improve efficiency and interoperability.

Enhancing Email Security

Protection Against Malicious Content: Identify and block emails containing malware, phishing attempts, or harmful links to prevent security breaches.

1.3.3 Improving User Experience

Reducing Inbox Clutter: Minimize spam emails to ensure important messages are easily accessible, improving organization and enhancing overall email management efficiency.

Preventing Disruptions: Reduce spam-related distractions, allowing users to focus on legitimate communication without unnecessary interruptions, boosting productivity and workflow effectiveness.

2. LITERATURE SURVEY

Researchers have extensively studied spam email detection due to the growing number of unsolicited messages that threaten security and privacy. Numerous studies have investigated machine learning techniques to identify spam, strengthen email security, improve message deliverability, and enhance filtering accuracy while minimizing false positives for a better user experience.

2.1 Machine Learning Approaches in Spam Detection:

2.1.1 Naïve Bayes (NB):

One of the widely used models is the Naïve Bayes (NB) algorithm, known for its probabilistic approach and ability to process large datasets efficiently. Research by Androutsopoulos et al. (2000) highlighted its effectiveness in spam classification by evaluating word probabilities in spam and legitimate emails.

2.1.2 Support Vector Machine (SVM):

Support Vector Machine (SVM) is a supervised learning algorithm widely used for classification and regression tasks. It is especially effective for text classification, including spam email detection, due to its capability to handle high-dimensional data efficiently.

SVM operates by identifying an optimal hyperplane that separates data points into distinct categories. In spam detection, the algorithm is trained on a dataset containing emails labeled as spam or non-spam (ham). It maps these emails into a high-dimensional space and determines a hyperplane that maximizes the margin between the two classes. For cases where the data is not linearly separable, SVM applies kernel functions, such as polynomial or radial basis function (RBF) kernels, to transform the data into a higher-dimensional space where classification becomes more effective.

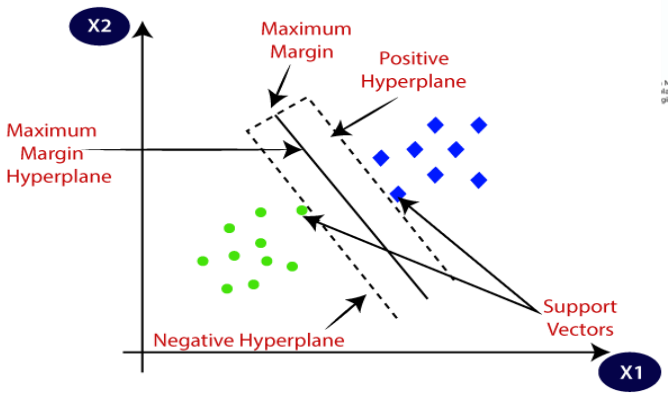

Fig. 2: Support Vector Machine

2.2 Feature Engineering and Data Preprocessing

Research highlights the significance of preprocessing email data to enhance spam detection accuracy. Zhou et al. (2016) emphasized key techniques, including:

Tokenization and stop-word removal to eliminate irrelevant words.

Stemming and lemmatization to standardize word forms.

Term Frequency-Inverse Document Frequency (TF-IDF) for effective feature extraction.

Word embeddings (Word2Vec, GloVe) to improve deep learning model performance.

2.3 Writing Optimized Emails

Studies on email deliverability indicate that proper email structuring reduces the likelihood of messages being marked as spam. SpamAssassin (2021) recommended best practices such as:

Avoid excessive capitalization and special characters to maintain professionalism.

Use clear formatting and professional language for better readability.

Minimizing promotional language to prevent triggering spam filters.

Maintaining a balanced text-to-image ratio to improve deliverability.

3. SYSTEM ARCHITECTURE AND DESIGN

A spam email detection system consists of several essential components, including gathering and processing data, extracting relevant features, training a machine learning model, and implementing a scoring mechanism to determine whether an email is spam or legitimate.

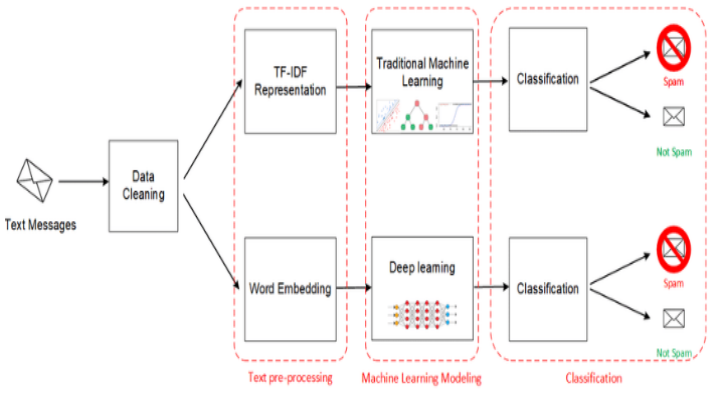

Fig. 3: SYSTEM ARCHITECTURE AND DESIGN

3.1 Hardware components

Processors (CPUs):

Efficient and high-performance processors are crucial for handling complex tasks such as analyzing emails, running spam filters, processing large datasets, and ensuring accurate, real-time spam detection for improved cybersecurity.

Memory (RAM):

Maximum RAM is useful for storing and processing email data, algorithms, and filters.

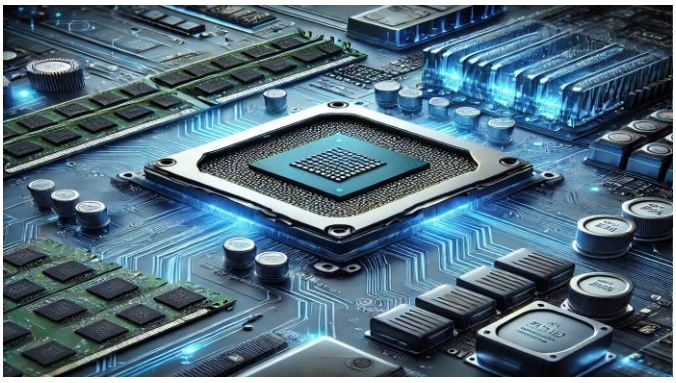

Fig. 4: Hardware

Storage (Hard Disk/SSD):

High storage capacity is required to store email in databases, spam lists, and other necessary data.

Network Infrastructure:

A strong network infrastructure, including routers and switches, is essential for managing large volumes of email traffic while ensuring efficient delivery and effective filtering.

4. PROPOSED SYSTEM AND IMPLEMENTATION

This system utilizes machine learning and natural language processing (NLP) to enhance spam detection accuracy, helping to prevent users from receiving malicious or unwanted emails. This section details the methodology and steps involved in its implementation. The proposed spam detection system follows a structured workflow to enhance accuracy and efficiency. It begins with data collection and preprocessing, where email datasets are gathered, cleaned, and standardized by removing duplicates, addressing missing values, and processing text through tokenization and stopword removal. Following this, feature extraction is conducted by analyzing email metadata, subject lines, and body content using Natural Language Processing (NLP) techniques. The extracted text data is then converted into numerical representations, such as TF-IDF or word embeddings, to facilitate effective processing.

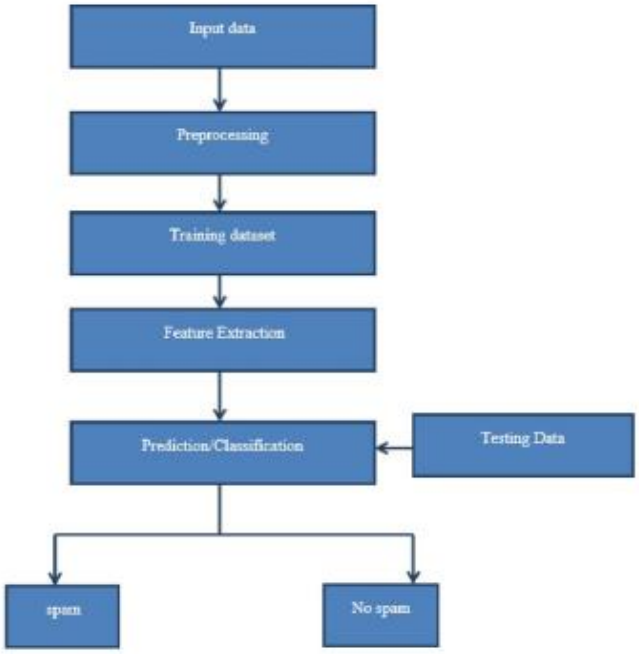

Fig. 5: Work Flow

This system architecture consists of three feature

- Detecting Spam Email With Machine Learning
- Set spam filter
- Writing Optimized Email

4.1 Detecting Spam Email With Machine Learning

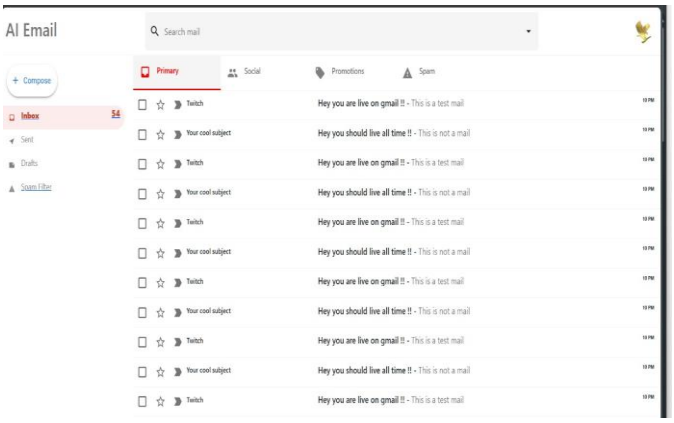

Fig. 6: Interface

This AI-powered email inbox system is designed to streamline email organization and enhance spam detection. The interface features multiple tabs, such as Primary, Social, Promotions, and Spam, allowing users to efficiently categorize and manage their emails. A navigation panel on

the left provides quick access to key email functions, including Inbox, Sent, Drafts, and the Spam Filter.

Within the Primary tab, emails are displayed in a list format, featuring checkboxes for selection, star icons for marking important messages, and sender details along with subject lines. The integrated AI system identifies potential spam and automatically categorizes such emails to maintain a clutter-free inbox. Additionally, the Spam Filter option in the left panel allows users to review flagged emails, ensuring that no important messages are mistakenly classified as spam.

4.2 Set Spam Filter

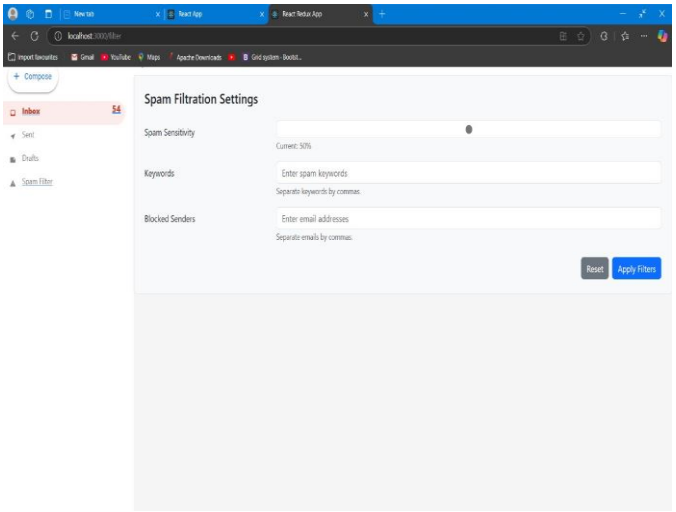

Fig. 7: Spam Filter

This interface highlights the Spam Filtration Settings feature, enabling users to personalize their spam detection preferences for a tailored email experience.

Spam Sensitivity: A slider allows users to adjust the level of spam detection, determining how strictly the system filters out unwanted emails.

Keywords: Users can specify certain words associated with spam, ensuring that emails containing these terms are automatically flagged.

Blocked Senders: This section allows users to manually add email addresses they wish to block, preventing messages from unwanted senders.

Action Buttons: The "Reset" button removes all applied settings, while the "Apply Filters" button saves the user's customized preferences.

This feature enhances email security and organization by allowing users to fine-tune their spam detection settings, ensuring important messages remain accessible while filtering out unwanted emails.

4.3 Writing Optimized Email

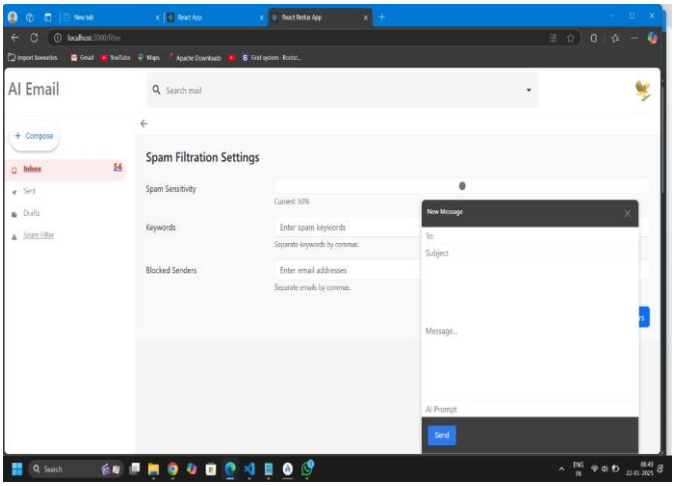

Fig. 8: Optimized email writing

This feature presents an AI-powered email composition tool designed to help users draft emails quickly and efficiently.

New Message Window: Users can enter recipient details in the "To" field, along with the subject line and message body.

AI Prompt: A dedicated field where users can provide a brief input or request, allowing the AI to generate a well-structured email accordingly.

Send Button: Enables users to send the composed email instantly with a single click.

This tool streamlines the email drafting process, ensuring professional and well-written messages with minimal effort, making communication faster and more efficient.

5 Conclusions

This research introduced a machine learning-based approach to improving spam detection and optimizing email composition. The proposed system enhances email security by effectively filtering spam through automated classification models and customizable filtering settings. By leveraging machine learning algorithms, the system identifies spam emails based on content analysis, sender reputation, and user-defined preferences. Additionally, the integration of an AI-powered email composition tool simplifies the writing process using natural language processing (NLP) techniques. This allows users to generate well-structured and contextually relevant emails with minimal effort. The research findings indicate that machine learning significantly enhances email filtering accuracy, minimizing user exposure to spam and phishing threats. Moreover, the AI-driven email composition feature boosts productivity by enabling faster and more efficient communication. Future improvements may include expanding the dataset, incorporating deep learning models

for more precise spam detection, and refining AI-generated email personalization based on user behavior and intent.

6. ACKNOWLEDGEMENT

I am delighted to present this project report on **"Detecting Spam Emails with Machine Learning and Writing Optimized Emails."** I sincerely appreciate and extend my heartfelt gratitude to everyone who contributed their valuable insights and support toward the successful completion of this report.

It is often said that every achievement is built on the contributions of others. I am deeply thankful to my guide, **Prof. Radhika Nanda**, from the **Department of Computer Engineering, Bharat College of Engineering, Badlapur, Thane**, whose unwavering support, encouragement, and expertise played a crucial role in the success of this project. Her valuable guidance, insightful suggestions, and constant motivation greatly contributed to the completion of this work.

I also take this opportunity to express my sincere gratitude to **Prof. Dr. B.M. Shinde, Principal, BCOE**, for providing the necessary resources and support. My thanks extend to **Prof. Radhika Nanda, Head of the Department of Computer Engineering**, and **Prof. Deepa Athawale, Project Coordinator, Department of Computer Engineering, BCOE**, for their assistance and encouragement throughout this project.

Finally, I am grateful to everyone who directly or indirectly contributed to the successful completion of this work. This project would not have been possible without the collective efforts and support of all those mentioned above.

6. REFERENCES

[1] Emmanuel Gbenga Dada et al. "Machine learning for email spam filtering: review, approaches and open research problems", Heliyon 5 (2019) e01802 Received 3 September 2018; Received in revised form 25 February 2019; Accepted 20 May 2019

[2] K. Varun Kumar, et al. "Machine Learning-based spam detection usingNaïve Bayes Classifier in comparison with logistic Regression for improving accuracy", Journal of Pharmaceutical Negative Results ¦ Volume 13 ¦ Special Issue 4 ¦ 2022, Page No. 548-554

[3] Nebojsa Bacanin, et al. "Application of Natural Language Processing and Machine Learning Boosted with Swarm Intelligence for Spam Email Filtering", Mathematics 2022, 10, 4173. https://doi.org/10.3390/math10224173

[4] S. S. a. N. N. Kumar, "Email Spam Detection Using Machine Learning Algorithms," in Second International Conference on Inventive Research in Computing Applications (CIRCA), 2020.

[5] İ. A. D. a. M. D. H. Karamollaoglu, "Detection of Spam E-mails with Machine Learning Methods," in Innovations in Intelligent Systems and Applications Conference (ASYU), 2018.

[6] Pooja Malhotra, et al. "Spam Email Detection using Machine Learning and Deep Learning Techniques", https://ssrn.com/abstract=4145123

[7] Neha Karadkar, et al. "Spam Mail Classification Using SVM and Genetic Algorithm", Journal of Emerging Technologies and Innovative Research (JETIR), Page No. e513-e517

[8] Naresh Vinod Wankhade, et al. "Paper on Spam Email Detection with Classification Using Machine Learning", INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY, IJIRT 156181, Page No. 1055-1059

[9] Rajesh Kumar J, et al. "Email Spam Detection using Machine Learning Techniques", International Advanced Research Journal in Science, Engineering and Technology Vol. 8, Issue 6, June 2021, DOI: 10.17148/IARJSET.2021.8632 Page No. 189-193

[10] A. Sharaff, A. Dhadse and Naresh K. Nagwani (2016), Comparative study of classification algorithms for spam email detection, in Emerging Research in Computing, communication and applications, Information, pp. 237-244, Springer, Berlin, Germany, DOI: 10.1007/978-81-322-2553-9_23.

[11] Alfandi O., Dahmani N. and Kaddoura S., "A Spam Email Detection Mechanism for English Language Text Emails Using Deep Learning Approach", IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises, France, Bayonne, pp. 193-198, DOI: 10.1109/WETICE49692.2020.00045.

7. BIOGRAPHIES

Prof. Dr. Radhika Nanda[5]
[5] HOD, Department of Computer Engineering, Bharat College of Engineering, Opp. Gajanan Maharaj Temple, Kanhor Road, Badlapur (West), Thane, Maharashtra - 421503

Siddharth More
Pursuing B.E in computer from the department of computer at Bharat College of Engineering Badlapur, thane, Maharashtra

Rajat Pandit
Pursuing B.E in computer from the department of computer at Bharat College of Engineering Badlapur, thane, Maharashtra

Rohit Jawale
Pursuing B.E in computer from the department of computer at Bharat College of Engineering Badlapur, thane, Maharashtra

Yash Salunke
Pursuing B.E in computer from the department of computer at Bharat College of Engineering Badlapur, thane, Maharashtra