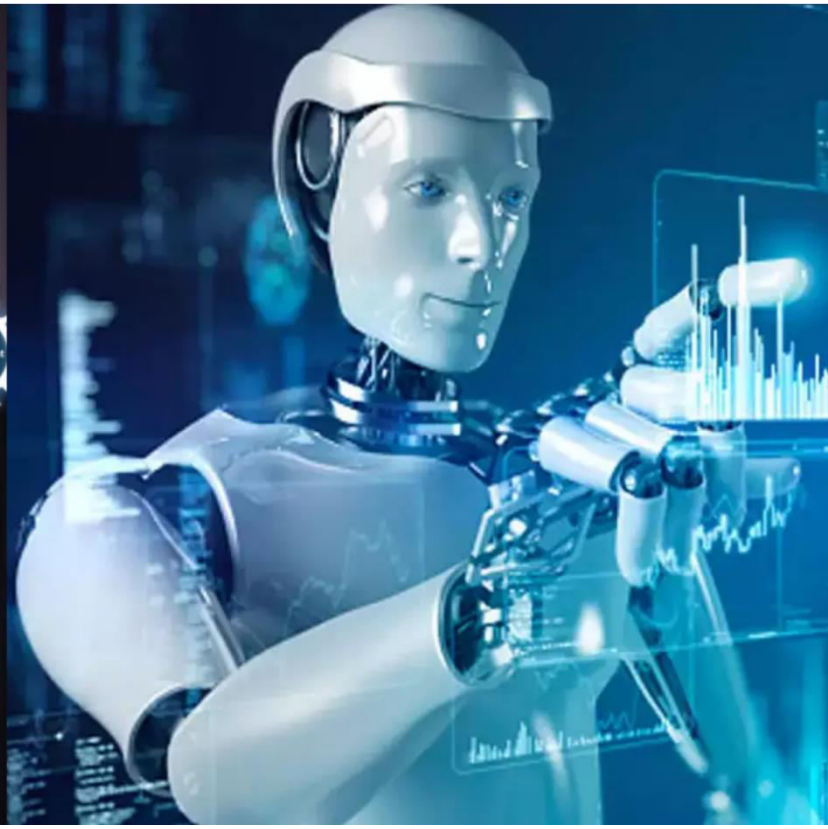
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 3, March 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503157|

Email Spam Detection System using NLP and Machine Learning

Mitali Gaikwad[1], Rajlaxmi Gaikwad[2], Sai Jagtap[3], Pratiksha Kandgule[4]

Mentor, Department of Computer Engineering, AISSMS Polytechnic, Pune, Maharashtra, India. [1]

Students, Diploma in Computer Engineering, AISSMS Polytechnic, Pune, Maharashtra, India.[2,3,4]

ABSTRACT: The rapid growth of email communication has led to a significant increase in spam messages, which negatively affect user productivity and pose security risks like phishing and fraud. Manually filtering spam emails is inefficient due to the high volume of daily messages. This project presents an Email Spam Detection System that uses Natural Language Processing (NLP) and the Naïve Bayes classifier to automatically sort messages as spam or not spam.

The dataset for this project is sourced from Kaggle and contains labeled spam and legitimate email messages. We apply NLP techniques such as text preprocessing, tokenization, stop-word removal, and TF-IDF feature extraction to convert email text into numerical form. The Naïve Bayes classifier is trained on this processed data to effectively identify spam patterns.

Besides spam detection, the system includes an Inbox for legitimate emails, a Trash Bin for spam or deleted emails, and a History column to track email activity and classification results. Experimental evaluation shows that the proposed system achieves reliable accuracy and efficient spam detection. This system offers a simple, scalable, and user-friendly solution for secure email management.

KEYWORDS: Email Spam Detection, NLP, Machine Learning, Inbox Management, Spam Classification.

I. INTRODUCTION

Email is a common method for exchanging information in personal, academic, and professional settings. However, the misuse of email services has resulted in a huge rise in spam messages, including promotional emails, fake offers, and phishing attempts. These spam emails clutter inboxes and pose serious security threats.

To tackle these challenges, we need automated spam detection systems. This project focuses on detecting whether an email is spam or not spam using NLP techniques and the Naïve Bayes machine learning algorithm. We chose the Naïve Bayes classifier for its simplicity, efficiency, and strong performance in text classification.

The proposed system not only classifies emails but also includes email management features like an inbox, trash bin, and history column, making it suitable for real-world email applications.

II. PROBLEM STATEMENT

The growing number of spam emails makes it hard for users to manage their inboxes effectively. Spam messages waste time, use storage, and may contain harmful links or fraudulent information. Traditional rule-based spam filters are inadequate because of changing spam patterns and language variations.

Manually identifying spam emails is impractical and unreliable. Therefore, we need an intelligent email spam detection system that can automatically classify messages as spam or not spam while offering features such as inbox organization, trash management, and activity history tracking. Such a system enhances email security, usability, and trust in digital communication.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503157|

III. SCOPE AND OBJECTIVE

Scope:
The goal of this project is to create an email spam detection system that uses NLP and ML techniques to accurately classify messages. The system features an inbox for legitimate emails, a trash bin for deleted or spam emails, and a history column to track email-related activities.

Objectives:
- Detect whether an email is spam or not spam.
- Preprocess and analyze email content using NLP techniques.
- Implement machine learning models for message classification.
- Provide an organized inbox for legitimate emails.
- Maintain a trash bin for removed emails.
- Store user actions and email statuses in a history column.

IV. LITERATURE REVIEW

SR NO.	AUTHOR(S) AND YEAR	Title	Methodology used	Key Findings/Contribution
1	Md. Tanvir Chowdhury et al. (2024)	NLP-based Classifier for Identifying Spam Emails and SMS	NLP, TF-IDF, Naïve Bayes, RF, SVM, DNN	Achieved high accuracy (up to 98%) using Bernoulli and Multinomial Naïve Bayes for spam detection.
2	Nathin V & Elampirai Gopika (2025)	Spam Detection Using NLP for Social Media and Email	NLP + Random Forest	Improved spam filtering accuracy by combining NLP with Random Forest for emails and social media data.
3	Pritesh A. Patil & Prayag Bhosale (2022)	Literature Survey on Spam Email Detection	Survey of ML & DL algorithms	Compared ML models like Naïve Bayes, SVM, RF, NN and highlighted research gaps in spam filtering.
4	Rajasree R.S. et al. (2022)	Email Spam Detection Using NLP	NLP, ML classifiers	Identified effective ML algorithms for detecting spam with high precision and accuracy.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503157|

V. METHODOLOGY

A. Implementation:

1) Data Acquisition

Download the email spam dataset from Kaggle and import it into the system.

2) Data Preprocessing (NLP)

- Convert text to lowercase
- Remove punctuation and special characters
- Tokenization
- Stop-word removal
- Lemmatization

3) Feature Extraction

Term Frequency–Inverse Document Frequency (TF-IDF)

4) Model Development

Train the Naïve Bayes classifier using extracted TF-IDF features because of its probabilistic nature and efficiency in text classification.

5) Training and Testing

Split the dataset into training and testing sets with an 80:20 ratio to evaluate model performance.

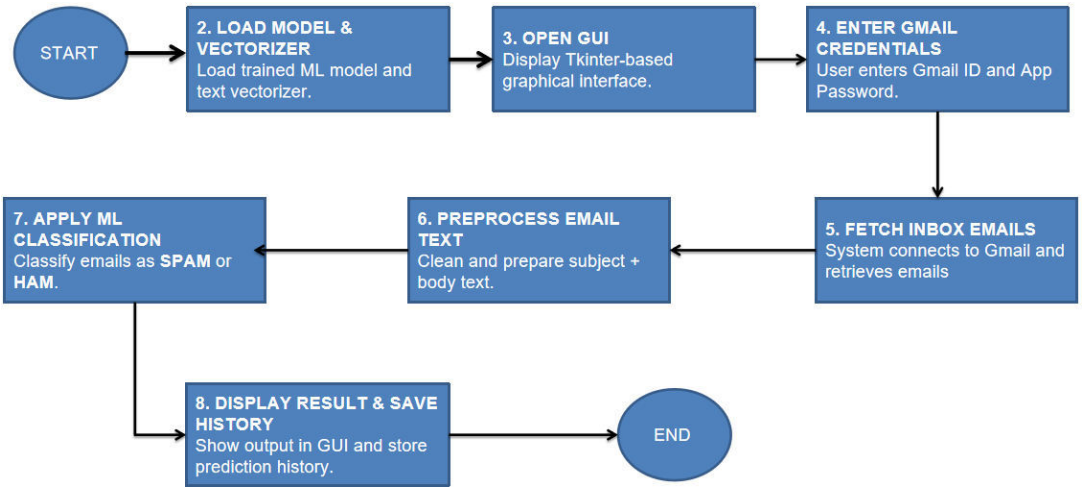


Fig 1. System Architecture

B. Technologies Used:

1) Front-end

Python Tkinter

2) Backend (Application Logic)

Python 3

3) Machine Learning Layer (ML Model Used)

Multinomial Naive Bayes
Feature Extraction- TF-IDF Vectorizer
ML Libraries- scikit-learn, numpy

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503157|

4) Data Processing (NLP Pipeline)
- Text Preprocessing Steps
- Lowercasing text
- Removing punctuation and symbols
- Tokenization
- Removing stopwords
- Feature extraction using TF-IDF

5) Development Tools & Technologies
Programming Language- Python 3
IDE / Editor- VS Code
Libraries & Frameworks- Tkinter – GUI
scikit-learn – Machine Learning
NLTK / re – Text processing
IMAP – Email fetching
Pickle – Model serialization
Operating System- Windows

6) System Architecture- Architecture Type- Standalone Desktop Application
Layered Architecture:
System Architecture Flow
User Interface (Tkinter GUI)
↓
Backend Logic (Python)
↓
Email Fetching (IMAP)
↓
Text Processing (NLP)
↓
Feature Extraction (TF-IDF)
↓
ML Model (Naive Bayes)
↓
Prediction Output (SPAM / HAM)
↓
GUI Display + History Storage

VI. CONCLUSION

The proposed email spam detection system effectively identifies spam and not spam messages using NLP and machine learning techniques. By integrating inbox management, trash bin functionality, and history tracking, the system provides a complete and user-friendly email management solution. This approach improves email security, reduces user effort, and boosts overall communication efficiency.

VII. FUTURE SCOPE

- Integration of deep learning models such as LSTM and CNN
- Real-time email filtering
- Detection of phishing and malicious links
- Multi-language email spam detection
- Cloud-based deployment for scalability

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503157|

REFERENCES

[1] M. T. Chowdhury et al., "NLP-based classifier for identifying spam emails and SMS," International Journal of Computer Applications, 2024.
[2] V. Nathin and G. Elampirai, "Spam detection using NLP for social media and email," International Journal of Advanced Research in Computer Science, 2025.
[3] P. A. Patil and P. Bhosale, "A literature survey on spam email detection," International Journal of Engineering Research & Technology (IJERT), 2022.
[4] R. S. Rajasree et al., "Email spam detection using natural language processing," International Journal of Scientific Research in Computer Science, 2022.
[5] A. Kumar et al., "Hybrid SVM with NLP and probabilistic neural network for phishing email detection," Journal of Information Security, 2020.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details