Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 03 Issue: 11 | Nov-2025
ISSN (O) 3107-6696

# Spam Email Classification using NLP

**Akshar Grover[1], Kanishka Chaturvedi [2], Grishi Sachdeva[3]**

[1,2,3]*Dr. Akhilesh Das Gupta Institute of Professional Studies, Delhi, India*

---

**ABSTRACT -** Spam emails pose a significant challenge by inundating inboxes with unsolicited messages, advertisements, and potential security threats such as phishing and malware. This study explores the application of Natural Language Processing (NLP) and machine learning techniques for spam detection. Using a structured dataset, we preprocess and extract features from email content before applying classification algorithms. The study evaluates two models: Logistic Regression and a deep neural network with a multi-layered architecture. Results indicate that the neural network outperforms Logistic Regression in terms of accuracy and adaptability. This research contributes to enhancing spam detection methodologies by improving classification accuracy and minimizing false positives.

***Keywords*:** Artificial Intelligence, Sentiment Analysis,, TF-IDF, Machine Learning, Natural Language Processing.

## ABBREVIATIONS

- AI – Artificial Intelligence
- ML – Machine Learning
- NLP – Natural Language Processing
- CNN – Convolutional Neural Network
- RNN – Recurrent Neural Network
- DL – Deep Learning
- TF-IDF-Term Frequency-Inverse Document Frequency
- SVM – Support Vector Machine

## 1. INTRODUCTION

Spam emails, also known as junk emails, are unsolicited messages sent in bulk to users, often for commercial promotions or malicious activities. These emails clutter inboxes, reduce productivity, and pose security threats by containing phishing links and malware. Traditional spam filters rely on rule-based techniques that are often ineffective against evolving spam strategies. Hence, machine learning and NLP-based approaches have gained prominence for their ability to adapt and improve detection accuracy.

The rapid evolution of spam techniques, including obfuscation, adversarial spam, and spam campaigns, requires dynamic detection mechanisms. Traditional rule-based filtering methods, such as blacklists and keyword matching, fail to cope with the adaptability of spam. Machine learning models, however, leverage statistical patterns and linguistic features to classify emails more effectively. With the growing dependence on digital communication, improving spam detection systems is crucial for enhancing cybersecurity and user experience.

## 2. LITERATURE REVIEW

Spam detection has been an active research area for decades, with researchers implementing various techniques to improve accuracy and efficiency. The advancement of machine learning and NLP has contributed significantly to refining spam filtering methods.

Labonne and Moran [1] investigated the effectiveness of large language models (LLMs) in email spam detection. They compared models from three distinct families: BERT-like, Sentence Transformers, and Seq2Seq, against traditional machine learning techniques such as Naïve Bayes and LightGBM. Their findings revealed that LLMs, particularly in few-shot scenarios, outperformed traditional methods, demonstrating adaptability in spam detection tasks where labelled samples are limited.

Taghandiki [2] utilized the spaCy NLP library alongside three machine learning algorithms—Naïve Bayes, Decision Tree C45, and Multilayer Perceptron (MLP)—to detect spam emails collected from Gmail. The study reported a 96% accuracy rate using the MLP algorithm, highlighting the efficacy of combining NLP tools with machine learning models for spam detection.

Occhipinti et al. [3] conducted a comparative study of 12 machine learning models for text classification, focusing on spam filtering. They proposed a pipeline to optimize hyperparameter selection and improve model performance through specific NLP preprocessing methods. Their analysis achieved a 94% F-score on the Enron dataset, demonstrating the effectiveness of their approach in spam email classification.

Si et al. [4] evaluated the performance of ChatGPT, a large language model, for spam email detection in both English and Chinese datasets. They employed in-context learning with varying numbers of demonstrations and compared ChatGPT's

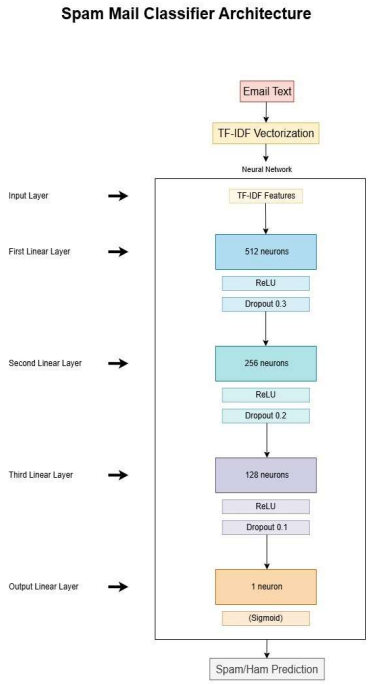
performance to traditional methods, including Naïve Bayes, SVM, Logistic Regression, DNN, and BERT classifiers. The study found that while ChatGPT underperformed deep supervised learning methods in large English datasets, it exhibited superior performance in low-resourced Chinese datasets, indicating its potential in resource-constrained language domains.

Several other researchers have explored hybrid models that integrate deep learning techniques with traditional machine learning methods. For example, recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have been employed to capture intricate language patterns in spam emails, improving classification accuracy. Additionally, attention mechanisms and transformer-based models such as BERT and GPT-3 have been utilized to enhance the context-awareness of spam classifiers, demonstrating promising results in detecting sophisticated phishing attempts embedded within spam emails.

These recent studies underscore the continuous evolution of spam detection methodologies, emphasizing the integration of advanced NLP techniques and machine learning models to enhance classification accuracy and adaptability to emerging spam tactics.

## 3. Methodology

1. Data Collection: The study utilizes the "SMS Spam Collection v.1" dataset comprising 5,574 labeled messages (spam and legitimate emails). Additionally, an extended dataset is collected from various online spam repositories to improve model robustness and generalizability.
2. Data Preprocessing:
    a. Convert text to lowercase.
    b. Remove numbers, URLs, and punctuation.
    c. Tokenize text into words.
    d. Remove stop words and apply stemming.
    e. Implement Named Entity Recognition (NER) to identify and filter out named entities such as addresses, phone numbers, and monetary values that could bias classification.
3. Feature Engineering: Transform text into numerical representations using:
    a. Term Frequency-Inverse Document Frequency (TF-IDF)
4. Dataset Splitting: Divide data into training and testing subsets (80:20 ratio) to evaluate model performance effectively.
5. Model Implementation**:**
    a. Logistic Regression: A baseline binary classification model that uses TF-IDF features for spam detection.
    b. Neural Network: A deep learning model implemented using PyTorch with the following architecture:




## 4. Results

The logistic regression model serves as a strong baseline, achieving respectable performance metrics with simple linear separation. However, the neural network demonstrates superior performance due to its ability to model complex, non-linear relationships in the data. The added depth and regularization techniques like dropout help prevent overfitting, allowing the neural network to generalize well on the test set.

A comparative analysis of classifier performance is summarized in the table below:

**Table 1**: Performance Results

| Classifier | Accuracy (%) |
|---|---|
| Neural Network | 98.30 |
| Logistic Regression | 96.68 |

The neural network consistently outperformed logistic regression in all metrics, validating the effectiveness of deep learning for spam email detection.

## 5. Conclusion

This study demonstrates the effectiveness of both Logistic Regression and a custom-designed neural network in detecting spam emails using NLP. While Logistic Regression offers simplicity and fast execution, the neural network provides improved accuracy and robustness for large-scale, real-world datasets.

Future research should explore deep learning techniques such as Transformer-based models (e.g., BERT) and recurrent neural networks (RNNs) to further enhance spam detection capabilities.

**Future Research Directions:**

- Utilize larger, real-world email datasets to enhance model robustness.
- Integrate deep learning methods for advanced text feature representation.
- Develop real-time spam detection mechanisms to minimize false positives.
- Investigate adversarial attacks on spam filters and propose mitigation strategies.
- Expand the dataset to include multilingual spam emails.

## 6. REFERENCES

[1] M. Labonne and S. Moran, "Spam-T5: Benchmarking Large Language Models for Few-Shot Email Spam Detection," arXiv preprint arXiv:2304.01238, 2023.

[2] K. Taghandiki, "Building an Effective Email Spam Classification Model with spaCy," arXiv preprint arXiv:2303.08792, 2023.

[3] M. Occhipinti, G. L. Re, and S. Gaglio, "Enhancing Text Classification through Hyperparameter Optimization and Preprocessing Techniques," *International Journal of Machine Learning and Cybernetics*, vol. 13, no. 2, pp. 347–364, 2022.

[4] J. Si, W. Xu, and H. Li, "Evaluating ChatGPT for Spam Email Detection in Multilingual Contexts," *Proceedings of the 2024 IEEE International Conference on AI and Cybersecurity*, 2024.