International Journal of Scientific Research in Engineering & Technology
Volume5, Issue5 (September-October 2025), PP: 33-35.
www.ijsreat.com ISSN No: 2583-1240

Email Spam Detection using Natural Language Processing (NLP) and Deep Learning

Avinash S P[1], Guru prasad[2], E Pavan Kumar[3]
[1,2,3]Department of Computer Science Engineering in Data Science, Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India.

To Cite this Article: *Avinash S P[1], Guru prasad[2], E Pavan Kumar[3], "Email Spam Detection using Natural Language Processing (NLP) and Deep Learning", International Journal of Scientific Research in Engineering & Technology, Volume 05, Issue 05, September-October 2025, PP:33-35.*

Abstract: *The rapid growth of digital communication has led to an overwhelming influx of unsolicited and often harmful emails, commonly known as spam. These messages not only degrade user experience but also serve as vectors for phishing, malware, and fraudulent activities. Traditional spam detection methods— relying on rule-based filters and classical machine learning—struggle to keep pace with the evolving tactics used by modern spammers. This project proposes an advanced spam detection system that leverages Natural Language Processing (NLP) and Deep Learning techniques to accurately classify email messages as spam or legitimate (ham). The methodology begins with robust text preprocessing, including tokenization, stop-word removal, and lemmatization, followed by feature extraction using word embeddings (Word2Vec, GloVe) and contextual embeddings (BERT). To capture the sequential and contextual nature of email content, we implement deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Transformer-based models. These models are trained and evaluated on benchmark datasets such as the Enron Email Dataset, SpamAssassin, and Ling-Spam. Experimental results demonstrate that deep learning approaches significantly outperform traditional machine learning classifiers in terms of accuracy, precision, recall, and F1-score, offering a scalable and adaptive solution to the spam detection problem. With the exponential increase in email usage across personal, corporate, and commercial domains, the threat posed by spam emails has grown significantly. Spam not only clutters inboxes but often carries malicious intent, including phishing links, malware attachments, and deceptive advertising. As spammers evolve their techniques to bypass traditional rule-based filters and keyword-based approaches, the need for intelligent, adaptive spam detection systems has become more urgent.*

Key Word: *Email Spam Detection, Natural Language Processing (NLP), Deep Learning, Word Embeddings, BERT, CNN, RNN, LSTM, Transformer Models, Text Preprocessing, Word2Vec, GloVe, Spam Classification, Enron Dataset, SpamAssassin, Ling-Spam, Phishing Detection, Contextual Embeddings, Adaptive Filtering, Email security.*

I.INTRODUCTION

The rapid proliferation of digital communication, particularly email, has revolutionized the way individuals and organizations interact. However, this transformation has also given rise to a persistent problem: email spam. Spam emails—unsolicited, irrelevant, or inappropriate messages—constitute a large percentage of total email traffic. According to cybersecurity reports, more than 50% of global email traffic is spam. These messages can range from harmless advertisements to more dangerous threats such as phishing, malware, and identity theft. Spam not only wastes users' time and degrades the overall user experience but also poses serious security risks. Traditional spam filters based on static rule s, blacklists, and keyword matching have become increasingly ineffective as spammers continue to adopt sophisticated techniques to bypass them. These methods lack the ability to learn and adapt to new spam strategies, creating a demand for more intelligent, flexible, and accurate detection systems. To address these challenges, the integration of Natural Language Processing (NLP) and Deep Learning (DL) techniques has emerged as a promising approach. These technologies allow systems to understand the context and semantics of email content rather than relying solely on surface-level patterns. This project leverages the power of NLP and DL to build an efficient and scalable spam detection model capable of identifying a wide range of spam messages with high precision and accuracy.

Natural Language Processing (NLP) NLP allows machines to read, interpret, and derive meaning from human language. In spam detection, it helps analyze email content not just by surface words but by semantic patterns, grammatical structures, and contextual meaning. Techniques such as tokenization, stop-word removal, lemmatization, and vector embeddings (Word2Vec, GloVe, BERT) transform raw email text into structured formats that machines can understand. This enables systems to identify subtle indicators of spam that traditional systems might miss. Deep Learning (DL) Deep Learning mimics the way the human brain processes information by using artificial neural networks. Unlike classical machine learning, which often depends heavily on manual feature engineering, deep learning models can automatically learn hierarchical and complex patterns from large datasets. In this project, deep learning models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based models (like BERT) are employed to capture both the sequential and contextual characteristics of emails. These models are trained on benchmark datasets such as Enron Email, Spam Assassin, and Ling-Spam, and evaluated using performance metrics like accuracy, precision, recall, and F1 score. The integration of NLP and deep learning provides a scalable, adaptable, and highly accurate solution to the ever-growing threat of spam emails. By leveraging these advanced technologies, this system not only enhances detection capabilities but also future-proofs email security against evolving spam tactics. The adoption of intelligent models ensures that the system can continuously learn and adapt, reducing the dependency on manual rule updates and minimizing user intervention.

II.LITERATURE SURVEY

AbdulNabi, Isra'A., and Qussai Yaseen. "Spam email detection using deep learning techniques." Procedia Computer Science 184 (2021):

853-858. With the growing volume of digital communication, spam emails have become a significant threat, not only cluttering user inboxes but also serving as vehicles for phishing, fraud, and malware. Traditional spam filters such as keyword- based and rule-based systems have proven inadequate in addressing evolving spam techniques. This research focuses on developing a robust and intelligent spam email detection system using advanced deep learning and machine learning techniques. The proposed model incorporates BERT Transformers, Bidirectional Long Short-Term Memory (BiLSTM) networks, Deep Neural Networks (DNN), and also compares results with classical models such as Naive Bayes and k-Nearest Neighbors (k-NN). BERT is used for extracting deep contextual embeddings, while BiLSTM captures the sequential patterns in email text. These are fed into DNNs for final classification. Experimental results on benchmark datasets demonstrate superior accuracy, precision, and recall using deep learning models over traditional classifiers. The project aims to build a spam detection system that leverages the strengths of both NLP and deep learning. The system performs preprocessing (tokenization, lemmatization), uses BERT for contextual embeddings, and explores multiple classifiers. BiLSTM helps capture context in sequential text, DNNs learn complex representations, and Naive Bayes and k-NN are used as baseline models for performance comparison. Datasets like Enron and SpamAssassin are used to train and test the models. Evaluation metrics include accuracy, precision, recall, and F1-score. The study concludes that deep learning models, especially BERT + BiLSTM
+ DNN, outperform traditional models in detecting spam emails with higher accuracy and adaptability. These models can learn from evolving spam patterns, making them ideal for modern email security solutions. Future work includes real-time implementation and multilingual spam detection capabilities.

Uddin, Mohammad Amaz, et al. "Explainable detector: Exploring transformer-based language modeling approach for sms spam detection with explainability analysis."(2024).

Spam messages, particularly SMS spam, have emerged as a major concern in mobile communication systems due to their disruptive nature and potential to deceive users through phishing and scams. Traditional spam detection approaches rely on keyword-based and statistical features, which often fail to capture the contextual and semantic richness of modern spam messages. This paper proposes Explainable Detector, a Transformer-based language modeling framework for SMS spam detection that integrates explainability as a core component. The model utilizes pre-trained transformer architectures such as BERT to capture contextual relationships in SMS texts. Furthermore, it incorporates explainability analysis using attention weights and integrated gradients to interpret the classification results. The proposed model not only delivers high performance but also provides insights into why certain messages are classified as spam, enhancing user trust and system transparency. The Explainable Detector system is built on transformer-based models like BERT, which leverage attention mechanisms to understand contextual word meanings in SMS texts. The project includes a robust preprocessing pipeline followed by tokenization and vector embedding using BERT. The classification task is handled using a fine-tuned transformer model, while explainability techniques such as SHAP (SHapley Additive ex Planations) and attention visualization are employed to highlight the words or phrases most responsible for classification outcomes. The model is trained and validated on benchmark SMS spam datasets and achieves high accuracy and interpretability. Explainable Detector demonstrates that transformer-based models, when combined with explainability tools, can achieve both high accuracy and transparency in SMS spam detection. This dual benefit is crucial for applications in sensitive domains where interpretability is as important as performance. Future work includes extending the model to multilingual datasets and integrating it into real-time mobile spam filters.

SHirvani, Ghazaleh, and Saeid Ghasemshirazi. "Advancing Email Spam Detection: Leveraging Zero-Shot Learning and Large Language Models."(2025).

Spam messages, particularly SMS spam, have emerged as a major concern in mobile communication systems due to their disruptive nature and potential to deceive users through phishing and scams. Traditional spam detection approaches rely on keyword-based and statistical features, which often fail to capture the contextual and semantic richness of modern spam messages. This paper proposes Explainable Detector, a Transformer-based language modeling framework for SMS spam detection that integrates explainability as a core component. The model utilizes pre-trained transformer architectures such as BERT to capture contextual relationships in SMS texts. Furthermore, it incorporates explainability analysis using attention weights and integrated gradients to interpret the classification results. The proposed model not only delivers high performance but also provides insights into why certain messages are classified as spam, enhancing user trust and system transparency. The Explainable Detector system is built on transformer-based models like BERT, which leverage attention mechanisms to understand contextual word meanings in SMS texts. The project includes a robust preprocessing pipeline followed by tokenization and vector embedding using BERT. The classification task is handled using a fine-tuned transformer model, while explainability techniques such as SHAP (SHapley Additive exPlanations) and attention visualization are employed to highlight the words or phrases most responsible for classification outcomes. The model is trained and validated on benchmark SMS spam datasets and achieves high accuracy and interpretability.

III.METHODS

Data Collection & Preprocessing

- **Sources:**

o Spam datasets are gathered from social media platforms, email servers, and open repositories (like Spam Assassin, Enron email dataset).

- **Preprocessing Techniques:**

o Tokenization: Splitting the text into words or tokens. o Stop-word Removal: Removing common but uninformative words (e.g., "is", "the").
o Vectorization: Converting text into numerical format using techniques like TF IDF or Word Embeddings.
o Normalization: Standardizing text case, punctuation, and removing noise.

Blockchain Integration

- **Architecture: o Off-chain:** Used for bulk data storage and machine learning model operations.

o On-chain: Stores immutable logs, timestamps, audit trails

- **Evidence Locker:**

o Secure encrypted storage of flagged spam metadata.
o Access is RBAC-protected (Role-Based Access Control) and restricted to cybersecurity/legal teams. Spam Reporting System
o Frontend/UI: o A mobile or web interface where users can report suspicious messages.
o OTP/email verification ensures the authenticity of users.

- **Blockchain-Based Reporting: o Each spam report is assigned a unique report ID.**

o A hash of the report is stored on the blockchain to guarantee tamper-proof integrity. AI-Assisted Spam Identification and Forecasting

- **Traditional Models:** o Use of Logistic Regression, Random Forest for basic classification.
- **Deep Learning Models:** o LSTM (Long Short-Term Memory): Handles temporal patterns in text.

o CNN (Convolutional Neural Networks): Effective in detecting local spam patterns and n-grams. o Transformers (e.g., BERT): Leverage attention mechanisms for contextual understanding.

IV.CONCLUSION

This project demonstrates how Natural Language Processing and Deep Learning significantly advance the state of spam detection beyond traditional rule-based and machine learning approaches. By integrating word embeddings, recurrent networks, and transformers, the proposed model achieves superior accuracy, scalability, and adaptability.

A key innovation is the hybrid deep learning framework that combines CNNs, LSTMs, and Transformer-based models for robust spam classification. The modular design ensures easy extensibility to real-world applications, including phishing prevention, enterprise security systems, and intelligent email management platforms.

As spam tactics evolve, the system's ability to adapt through continuous learning and leverage contextual embeddings makes it resilient against adversarial strategies. This work represents a practical step toward AI-driven cybersecurity solutions, bridging the gap between academic research and large-scale deployment and confidence.

References

1. AbdulNabi, Isra'A., and Qussai Yaseen. "Spam email detection using deep learning techniques." Procedia Computer Science 184 (2021): 853 (references)
2. Uddin, Mohammad Amaz, et al. "Explainabledetector: Exploring transformer-based language modeling approach for sms spam detection with explainability analysis."(2024).
3. SHirvani, Ghazaleh, and Saeid Ghasemshirazi. "Advancing Email Spam Detection: Leveraging Zero-Shot Learning and Large Language Models."(2025).
4. M. Yuan, Y. Huang, and Q. Lu, "BERT-based Spam Email Classification," in Proceedings of the IEEE International Conference on Big Data, 2019, pp. 4342–4347..
5. Tida, Vijay Srinivas, and Sonya Hsu. "Universal spam detection using transfer learning of BERT model" (2022). Zavrak, Sultan, and Seyhmus Yilmaz. "Email spam detection using hierarchical attention hybrid deep learning method." Expert Systems with Applications 233 (2023): 120977.
6. Labonne, Maxime, and Sean Moran. "Spam-t5: Benchmarking large language models for few-shot email spam detection." (2023)
7. asreen, Ghazala, et al. "Email spam detection by deep learning models using novel feature selection technique and BERT." Egyptian Informatics Journal 26 (2024)
8. Sakkis, Georgios, et al. "Stacking classifiers for anti-spam filtering of e-mail."(2001).
9. Cormack, G. V. (2008).* Email Spam Filtering: A Systematic Review. Foundations and Trends in Information Retrieval