Deep Learning for Spam Detection Using Neural Networks and NLP

1st Kavya Shree J
Department of Artificial Intelligence and Data Science East West Institute of Technology
kavyashree1108@gmail.com

2nd Sahana B Surasgar
Department of Artificial Intelligence and Data Science East West Institute of Technology
sahanasurasgar@gmail.com

Prof Ashwini R
Department of Artificial Intelligence and Data Science East West Institute of Technology
ashwiniramegowda@ewit.edu.in

***Abstract*--- Spam detection is a major challenge in modern digital communication systems due to the rapid growth of fraudulent and unsolicited messages. Traditional machine learning methods rely on manual feature engineering and perform poorly on multilingual and context-dependent text, while deep learning approaches are computationally expensive. This paper presents a hybrid spam detection framework that integrates machine learning and deep learning models using Natural Language Processing (NLP). The system combines Naïve Bayes and Support Vector Machine classifiers with LSTM-based neural networks to improve accuracy and robustness. A fusion strategy is used to combine predictions, reducing false classifications. The model is deployed through an interactive interface for real-time detection and supports multilingual input. Experimental results demonstrate that the proposed hybrid approach outperforms individual models in accuracy and reliability**.

I. INTRODUCTION

The rise of digital communication has led to a rapid increase in spam and phishing messages across email and messaging platforms, causing security and privacy risks [1]. Classical machine learning approaches such as Naïve Bayes and SVM are efficient but depend on handcrafted features and struggle with contextual and multilingual text [2], [4]. Deep learning models such as LSTM and BiLSTM improve accuracy by learning sequence patterns, but they are computationally expensive and harder to deploy at scale [5].

This paper proposes a hybrid spam detection framework that combines machine learning and deep learning with NLP to balance accuracy and efficiency. The system integrates TF-IDF with Naïve Bayes and SVM, alongside LSTM-based models, and applies fusion to improve robustness. The solution supports multilingual input and is deployed as a lightweight web application for real-time detection.

II. LITERATURE SURVEY

Early spam detection research primarily relied on traditional machine learning algorithms such as Naïve Bayes, Support Vector Machines, Decision Trees, and Random Forest classifiers using hand-crafted features like Bag-of-Words, TF-IDF, and n-grams. These approaches achieved reasonable accuracy on structured and short-text data such as emails and SMS. However, their performance degraded when applied to evolving spam patterns, informal text, multilingual messages, and code-mixed content due to limited semantic understanding.

With advancements in deep learning, researchers introduced neural network architectures including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Bidirectional LSTM (BiLSTM) for spam detection. These models automatically learned hierarchical features from data, reducing the dependence on manual feature engineering. LSTM-based systems demonstrated improved performance in capturing long-term dependencies and contextual relationships in text, while BiLSTM models further enhanced detection by reading text in both forward and backward directions. Despite achieving higher accuracy, deep learning models often required large training datasets, high computational power, and longer training time, which limited their practical deployment.

Recent research has explored transformer-based models such as BERT and multilingual BERT for spam detection. These models use attention mechanisms to capture long-range dependencies and contextual semantics, resulting in state-of-the-art accuracy in many NLP tasks. However, their computational cost, high memory usage, and inference latency make them unsuitable for lightweight systems and real-time applications in low-resource environments.

Several studies proposed hybrid systems that combine classical machine learning and deep learning models to balance performance and computational efficiency. These approaches leverage the simplicity of statistical models and the contextual learning ability of deep networks. Comparative studies show that hybrid frameworks tend to outperform individual models by improving classification accuracy and reducing false positives. Researchers also recommend using ensemble and fusion strategies to enhance model robustness across different datasets, message lengths, and languages.

Although many contributions have been made, gaps still exist in providing scalable, multilingual, and resource-efficient spam detection systems. Most solutions focus on English

datasets and fail to generalize across diverse languages. Few models provide real-time deployment frameworks with user interfaces. Additionally, explainability and model interpretability remain open problems, as deep models continue to act as black-box classifiers.

III. METHODOLOGY

The methodology behind the proposed Spam Detection System combines the power of deep learning–based text classification with classical machine learning techniques through an intuitive, web-based interface. The system is designed to provide real-time spam detection while ensuring accuracy, ease of use, and robustness across multilingual text. The design philosophy focuses on hybridization, which integrates statistical learning with neural networks to enhance detection capabilities.

A. System Architecture

The proposed system follows a three-layer architecture:

1) Frontend (User Interface)

The user interface is developed using Streamlit, which enables users to enter text messages and obtain classification results instantly. The interface is simple and intuitive, requiring minimal technical expertise. Users can paste any message into the input field and receive the classification output along with confidence scores in real time.

2) Backend (Processing Server)

The backend acts as a bridge between the user interface and the trained classification models. It is implemented in Python and manages text preprocessing, model inference, and decision fusion. The backend handles input validation, feature extraction, probability calculation, and final prediction. Models are dynamically loaded during execution, and the fusion algorithm combines multiple outputs into a single classification decision [3].

3) Model Layer (Classification Engine)

The core processing layer uses multiple independent learning models that analyze the input message and generate prediction scores. The following classifiers are implemented:

- Naive Bayes classifier using TF-IDF features [4]
- Support Vector Machine (SVM) classifier [2]
- LSTM-based deep learning model [5]
- BiLSTM with Attention mechanism [13]

A fusion mechanism integrates predictions from each model to improve robustness and minimize misclassification.

Together, these components form an end-to-end spam detection pipeline that processes user input and returns classification results through the frontend interface.

B. Core Algorithms

Naive Bayes (NB)

Naive Bayes is a probabilistic classifier based on Bayes' theorem. It estimates the likelihood of a message being spam by learning word frequency distributions during training [4].

Support Vector Machine (SVM)

SVM constructs an optimal separating hyperplane between spam and non-spam messages using TF-IDF features. It is effective for high-dimensional sparse datasets and offers strong generalization capabilities [2].

LSTM and BiLSTM Networks

LSTM models capture sequential dependencies within text and store long-term contextual information [11]. BiLSTM extends LSTM by processing sequences in both forward and backward directions, further improving semantic understanding [5].

Attention Mechanism

Attention assigns importance weights to specific words in a sentence, enabling the model to concentrate on spam-indicative terms [13]. This increases classification accuracy and interpretability.

C. Workflow

1. User enters a message through the interface.
2. Input is validated and cleaned using NLP preprocessing methods.
3. Tokenization is performed.
4. TF-IDF features are created for machine learning models.
5. Word sequences are generated for deep learning models.
6. Naïve Bayes and SVM generate probabilities.
7. LSTM/BiLSTM outputs are computed.
8. Fusion algorithm integrates all results.
9. Final classification is displayed to the user.

D. Key Features and Advantages

- Hybrid Classification combining ML and DL [1] -Integrates statistical and deep learning models to improve accuracy and reduce false detection rates.

- Multilingual input support [7] - Enables detection across multiple languages and code-mixed text without requiring separate models.

- Fusion-based prediction strategy [6] - Combines outputs from multiple classifiers to deliver more reliable final predictions.

- Lightweight web-based deployment - Allows real-time detection through a browser without complex installation.

• Modular and scalable design - Supports easy integration of new algorithms and datasets for future expansion.

E. Applications

The system can be applied across:

• Email filtering systems - Automatically identifies and blocks spam and phishing emails.

• Messaging applications - Detects spam messages in real time on SMS and chat platforms.

• Social media moderation tools - Filters unwanted content and fake promotional messages.

• Enterprise communication platforms - Protects organizational communication from malicious attacks.

• Academic research in NLP and cyber security [10] - Serves as a research framework for experimentation.

IV. CONCLUSION

This paper presented a hybrid spam detection framework that integrates classical machine learning techniques with deep learning models using NLP. The combination of Naïve Bayes and SVM with LSTM and BiLSTM neural networks improves accuracy and robustness compared to standalone models. The use of feature-based learning along with contextual modeling enables effective identification of spam messages across multilingual and code-mixed text.

The fusion-based decision strategy enhances classification reliability by combining probabilities from multiple models. The Streamlit-based deployment framework allows real-time interaction and user-friendly operation, making the system suitable for academic and prototype-level applications. Experimental observations show reduced false positives and improved detection performance across diverse datasets.

Future enhancements include transformer-based architectures, multilingual fine-tuning, and real-time integration with communication platforms. Further research may also explore explainable AI methods to improve interpretability and trust in automated spam detection systems.

The proposed framework also demonstrates that hybrid learning offers an efficient alternative to computationally intensive transformer models, making it suitable for low-resource environments. The modular design enables easy upgrades of components such as feature extraction techniques and classification models, allowing the system to adapt to evolving spam patterns.

Additionally, the fusion strategy improves generalization across different languages and message formats, resulting in stable performance on real-world data. The system also provides a foundation for extending hybrid learning approaches to applications such as fake news detection and content moderation.

V. REFERENCES

[1] G. Svadasu and M. Adimoolam, “Spam detection in social media using artificial neural networks and SVM,” *IEEE International Conference on Communication and Signal Processing (ICCSP)*, 2022.
[2] B. Sonare, S. Dharmadhikari, and A. Patil, “E-mail spam detection using machine learning,” *IEEE International Conference on Emerging Techniques in Computational Intelligence (INCET)*, 2023.
[3] A. Jain, R. Mehta, and P. Sharma, “Gmail spam detection using natural language processing,” *IEEE International Conference on Data Science and Analytics*, 2023.
[4] Y. Arora, N. Saini, and A. Sharma, “SMS spam detection using enhanced Naïve Bayes classifier,” *IEEE International Conference on Advances in Computing, Communication and Informatics (ICACCI)*, 2022.
[5] A. A. E. Damanik, S. Simanjuntak, and R. Sihotang, “E-mail spam detection using LSTM network,” *IEEE International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2024.
[6] M. Ketcham, S. Wattanapongsakorn, and N. Kaewkamnerd, “Spam text detection using machine learning models,” *IEEE International Conference on Artificial Intelligence and IoT*, 2023.
[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *Proceedings of NAACL 2019*, 2019.
[8] A. Mozafari, A. El-Sayed, and M. Abdel-Basset, “Spam detection using deep learning techniques: A survey,” *IEEE Access*, vol. 9, pp. 1–18, 2021.
[9] K. Shukla and A. Dwivedi, “A hybrid model for spam classification using machine learning and deep learning,” *IEEE International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2021.
[10] S. Wang, S. Hong, and L. Zhu, “Survey on spam detection using deep learning models,” *IEEE Access*, vol. 10, pp. 1–15, 2022.
[11] S. Kaur and A. Madaan, “Multilingual spam detection using deep learning models,” *IEEE International Conference on Advances in Computing, Communication and Automation (ICACCA)*, 2022.
[12] R. Singh and P. Sharma, “Hybrid spam detection system using NLP and ensemble learning,” *IEEE International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 2023.
[13] P. Patil and R. S. Bhat, “Email spam classification using deep learning and ensemble learning,” *IEEE International Conference on Computing, Analytics and Security Trends (CAST)*, 2022.
[14] S. Kumar and N. K. Soni, “Multilingual text classification using deep neural networks,” *IEEE International Conference on Computational Intelligence and Data Science (ICCIDS)*, 2021.
[15] A. Mishra and S. Gupta, “A deep learning approach for spam detection in social media,” *IEEE International Conference on Network Intelligence and Digital Content (INDICON)*, 2023.