SMS Spam Detection & URL Malicious Classification

[1]S. Vasavi,[2]Sampangi Navya Sri,[3]S. Yamini,[4]Illindhala Amulya,[5]K. Ganga Bhavani,[6]Thota Anjali

[1]Assistant Professor, Department of Computer Science & Engineering, Princeton Institute of Engineering & Technology For Women

[2,3,4,5,6]B. Tech Students, Department of Computer Science & Engineering,Princeton Institute of Engineering & Technology For Women

ABSTRACT

In the digital era, the widespread use of mobile communication has made Short Message Service (SMS) a prime target for spammers and cybercriminals. Spam messages not only disrupt user experience but often serve as vectors for phishing attacks, malware distribution, and fraudulent schemes. With the proliferation of such threats, there is a pressing need for intelligent systems capable of automatically detecting and filtering spam content to safeguard users from potential harm. This project presents a hybrid machine learning approach that addresses two critical tasks: SMS spam detection and URL malicious classification. The first component focuses on classifying SMS messages as either spam or ham (legitimate) using natural language processing (NLP) techniques and supervised machine learning algorithms. Text preprocessing methods such as tokenization, stopword removal, and TF-IDF vectorization are employed to transform raw SMS text into meaningful features suitable for model training. The second component targets the classification of URLs embedded within SMS messages to determine whether they are malicious or benign. By extracting lexical features—such as URL length, number of digits, use of special characters, and domain-related attributes—the system utilizes ensemble classifiers like Random Forest and XGBoost to detect suspicious URLs. This dual-layered detection mechanism enhances security by identifying both unsolicited messages and hidden threats within them. Evaluation of both models was performed using publicly available datasets, and the results demonstrated high accuracy, precision, and recall, proving the effectiveness of the proposed approach. The integration of spam detection with malicious URL classification provides a more robust solution compared to traditional standalone filters, significantly reducing the risk of user exploitation. Overall, this project contributes a comprehensive solution for enhancing digital communication security. It can be deployed in mobile applications, messaging platforms, or enterprise systems to provide real-time protection against spam and malicious attacks, thereby fostering a safer messaging ecosystem for users.

Keywords: SMS Spam Detection, Malicious URL Classification, Machine Learning, Natural Language Processing (NLP), Cybersecurity, Phishing Detection, Text Classification, Feature Extraction, Supervised Learning, Deep Learning, Data Mining, URL Analysis, Security Analytics, Artificial Intelligence, Message Filtering.

I. INTRODUCTION

In recent years, the rapid expansion of mobile communication technologies has significantly increased the use of Short Message Service (SMS) for both personal and business interactions. However, this popularity has also made SMS a preferred medium for cybercriminals to disseminate spam, phishing links, and other harmful content. SMS spam not only clutters users' inboxes but can also lead to serious security breaches, financial fraud, and identity theft when users unknowingly interact with malicious content.

Traditionally, rule-based spam filters and blacklists were used to detect unwanted messages and harmful URLs, but these methods are no longer sufficient due to the evolving tactics of attackers. Modern spam messages often appear contextually relevant and may contain shortened or obfuscated URLs, making manual detection increasingly challenging. This has necessitated the development of intelligent, automated systems that can accurately detect spam messages and assess the threat level of any embedded links.

This project aims to build a dual-function intelligent system that can classify SMS messages as spam or

legitimate (ham) and simultaneously analyze URLs to determine if they are malicious or benign. By leveraging machine learning and natural language processing techniques, the system is designed to learn from large datasets and adapt to new spam patterns and phishing strategies more effectively than static filters.

The SMS spam detection component focuses on analyzing textual content to identify patterns commonly associated with spam messages, using algorithms such as Naïve Bayes, Logistic Regression, and Support Vector Machines. In parallel, the URL malicious classification component extracts lexical and structural features from URLs—such as domain type, URL length, and special character usage—to determine the likelihood of a link being dangerous, using ensemble learning models like Random Forest and XGBoost.

By combining these two security layers, the proposed system offers enhanced protection for mobile users against unwanted and potentially harmful content. The implementation of such a solution can significantly reduce cyber risks associated with SMS-based attacks, providing a smarter and safer communication experience.

II. LITERATURE SURVEY

1. Title: SMS Spam Detection Using Machine Learning Approach
Author(s): A. Almeida, J. Hidalgo, T. Pinedo
Description:
This paper presents a machine learning-based model for detecting spam messages using the SMS Spam Collection Dataset. It compares various classification algorithms like Naïve Bayes and SVM and emphasizes the importance of text preprocessing and feature extraction. The authors achieved high accuracy using TF-IDF and word frequency features, establishing a baseline for spam detection systems.

2. Title: Malicious URL Detection Using Machine Learning: A Survey
Author(s): M. Marchal, P. Francillon, M. Kaâniche
Description:
The authors provide a comprehensive survey of machine learning techniques used for detecting malicious URLs. The paper discusses the use of lexical features, host-based information, and content-based analysis for URL classification. It highlights the strengths and limitations of supervised and unsupervised learning models and discusses real-world challenges like zero-day attacks and data imbalance.

3. Title: Combining URL Analysis with Machine Learning to Detect Phishing Sites
Author(s): Xianghua Xu, Xiaowei Liu, Qingtian Zhan
Description:
This research focuses on phishing URL detection by analyzing lexical characteristics and applying ML algorithms such as Random Forest and Logistic Regression. The paper introduces feature engineering techniques such as entropy, domain trust level, and character distribution to distinguish malicious URLs. The system demonstrated high performance on multiple datasets.

4. Title: SMS Spam Filtering Techniques: A Review
Author(s): H. Mahajan, R. Batra
Description:
This review paper explores various techniques for SMS spam filtering including rule-based, keyword-based, and machine learning methods. It identifies major challenges in spam detection like language diversity, message obfuscation, and real-time filtering. The authors advocate for hybrid approaches using both NLP and classification models to improve detection rates.

5. Title: Effective Phishing Detection Using URL and HTML Features
Author(s): J. Ma, L. Saul, S. Savage
Description:
This study proposes a phishing detection framework that leverages lexical URL features in combination with HTML code analysis. The authors applied classifiers such as Gradient Boosting and SVM,

demonstrating that URL-based features alone are often sufficient to identify phishing links with high accuracy, making it suitable for lightweight mobile implementations.

III. EXISTING SYSTEM

The detection of SMS spam and malicious URLs has traditionally relied on standalone systems with limited adaptability. These systems include rule-based filters, blacklists, and signature-based detection mechanisms, which although initially effective, have proven to be insufficient against modern, dynamic cyber threats.

In the context of SMS Spam Detection, existing systems typically use keyword-based filtering or predefined rules that scan messages for specific words or phrases commonly associated with spam. While simple to implement, these systems are rigid and prone to high false positives and negatives, especially when spammers use intentional obfuscation or variations in language to bypass detection.

Similarly, for URL Malicious Classification, many existing solutions depend on blacklists such as Google's Safe Browsing or VirusTotal API. While effective in detecting known malicious domains, these systems fail to identify zero-day or previously unseen threats. Moreover, spammers often use URL shortening services or dynamic domain generation, making blacklist-only systems less effective.

Furthermore, traditional systems lack the ability to learn from data or adapt over time, which is crucial given the constantly evolving tactics used by attackers. They also typically operate in isolation, with spam detection and URL classification handled by separate components, leading to gaps in security and limited contextual analysis.

Some commercial anti-spam applications provide integrated services, but they are often proprietary, expensive, and inaccessible for academic research or small-scale deployment. Additionally, they rarely offer transparency in terms of how decisions are made, which can be problematic in environments requiring explainability and customization.

In summary, the existing systems offer only limited protection, are non-adaptive, and lack the intelligence to detect sophisticated or emerging threats. This highlights the need for a robust, data-driven, and integrated machine learning approach that can simultaneously classify SMS content and analyze embedded URLs for malicious behaviour.

IV. PROPOSED SYSTEM

The proposed system introduces a machine learning-based hybrid model that integrates both SMS spam detection and URL malicious classification into a unified and intelligent framework. This dual-layered approach enhances overall security by identifying not only unsolicited or spam messages but also potentially dangerous URLs embedded within them.

1. SMS Spam Detection Module

This component leverages Natural Language Processing (NLP) and supervised machine learning algorithms to analyze the content of SMS messages. The raw text is preprocessed through tokenization, stopword removal, and TF-IDF vectorization to extract meaningful features. Classification algorithms such as Naïve Bayes, Support Vector Machine (SVM), or Logistic Regression are trained on labeled datasets to classify incoming messages as spam or ham (legitimate). This enables real-time filtering and detection of suspicious or fraudulent SMS messages.

2. URL Malicious Classification Module

When an SMS contains a URL, the system automatically extracts and analyzes it using a separate model trained for malicious URL detection. Instead of relying on static blacklists, the model uses lexical features like URL length, number of digits, number of special characters, use of suspicious words, domain reputation, and entropy. Algorithms

like Random Forest, Gradient Boosting, or XGBoost are used for classification, enabling the system to identify zero-day phishing attacks and previously unknown threats.

3. Integrated Security Workflow

The system operates in a pipeline structure: an SMS is first analyzed for spam content; if it contains a URL, the link is then passed to the malicious URL classifier. This integrated approach ensures comprehensive analysis without compromising speed or efficiency. The entire process is automated and optimized for deployment on both **mobile and web-based platforms**.

4. Learning and Feedback Mechanism

The system can incorporate a feedback loop where user actions (e.g., marking a message as spam or not spam) help retrain and fine-tune the models periodically. This allows the system to evolve over time and maintain high accuracy even as new spam or malicious patterns emerge.

5. Advantages Over Existing Systems

Unlike traditional systems, this proposed model is data-driven, adaptive, and capable of handling previously unseen threats. It eliminates the need for manual updates, reduces false positives/negatives, and offers an end-to-end security solution for mobile communications.

V. SYSTEM ARCHITECTURE

The diagram illustrates the workflow of an SMS Spam Detection and URL Malicious Classification system that uses deep learning techniques to classify messages as either spam or legitimate. The process begins with the message dataset, which contains various SMS texts collected from users or public datasets. These messages may include normal communication, promotional content, phishing links, or malicious URLs. The raw messages are first sent to the preprocessing stage where the data is prepared for machine learning analysis.

In the pre-processing stage, the system performs data cleaning to remove unwanted elements such as special characters, punctuation, irrelevant symbols, and stop words that do not contribute to classification. This step also standardizes the text by converting all characters to a consistent format, such as lowercase. After cleaning the data, the system applies word embedding, which converts textual data into numerical vectors that can be understood by deep learning models. Word embedding techniques capture semantic relationships between words so that similar words have similar vector representations. This transformation allows the machine learning model to analyze patterns in the text more effectively. After preprocessing, the processed data is fed into the deep learning algorithms stage. In this stage, the system combines Convolutional Neural Networks (CNN) and Gated Recurrent Units (GRU) to improve classification accuracy. The CNN model extracts important features from the embedded text, such as key patterns or phrases that commonly appear in spam messages. CNN is effective in identifying local patterns and keyword combinations that frequently occur in malicious messages or phishing attempts.
The extracted features are then passed to the GRU layer, which is a type of recurrent neural network designed to process sequential data. GRU analyzes the order and contextual relationships between words in the message. This helps the system understand how words relate to each other in a sentence, improving its ability to detect spam messages that use deceptive or complex language patterns. GRU also helps reduce computational complexity while maintaining high performance in sequence modeling. Finally, the model performs classification, where the system predicts whether the incoming message is SPAM or NOT SPAM. If the message contains suspicious content, malicious URLs, or spam-like patterns, it is labeled as spam and can be blocked or flagged for the user. Otherwise, the message is classified as legitimate communication. This deep learning–based approach enhances the accuracy of spam detection systems and helps protect users from phishing attacks, fraudulent links, and unwanted promotional messages.

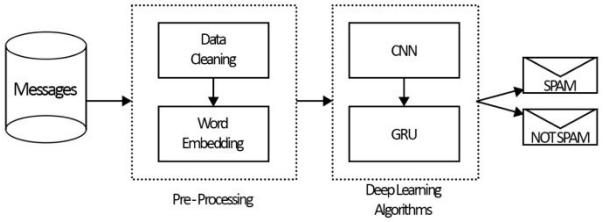

Fig 5.1: System Architecture Of Proposed System

VI. IMPLEMENTATION

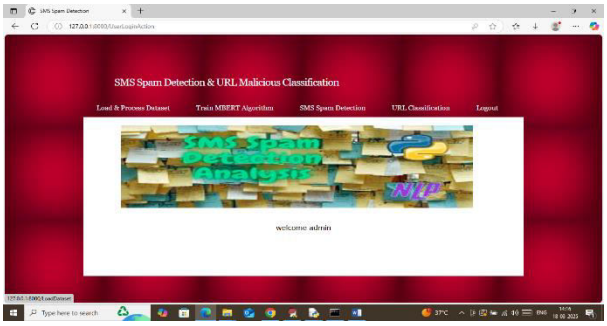

Fig 6.1: User Home

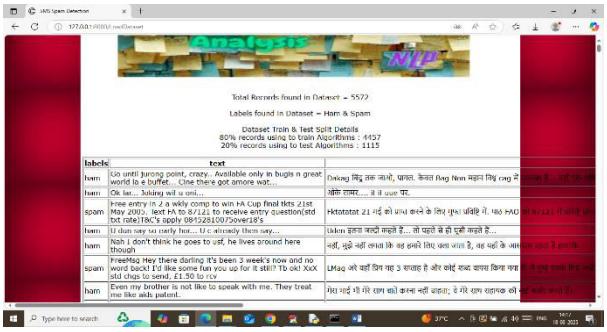

Fig 6.2: Load And Preprocess Dataset

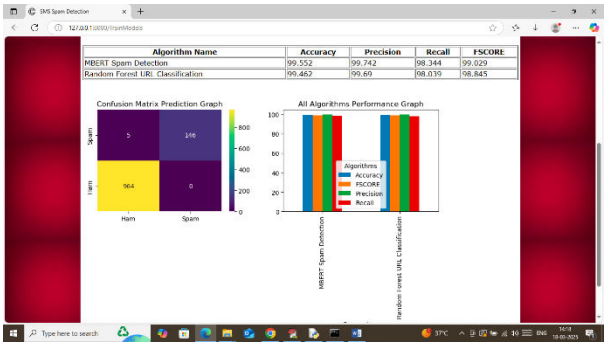

Fig 6.3: Model Training

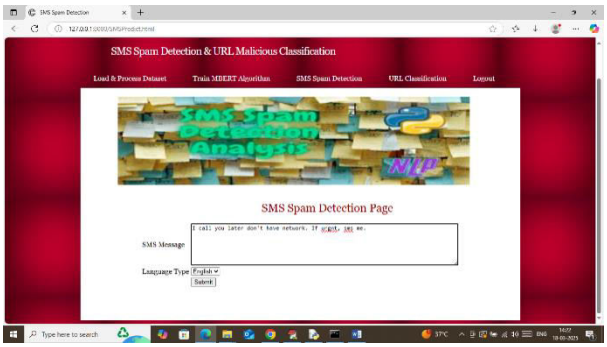

Fig 6.4: SMS Spam Detection Page

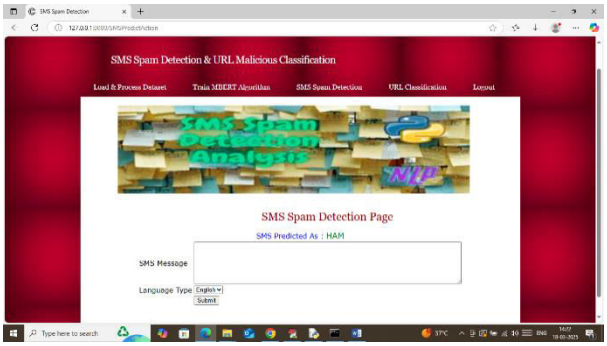

Fig 6.5: Result Page

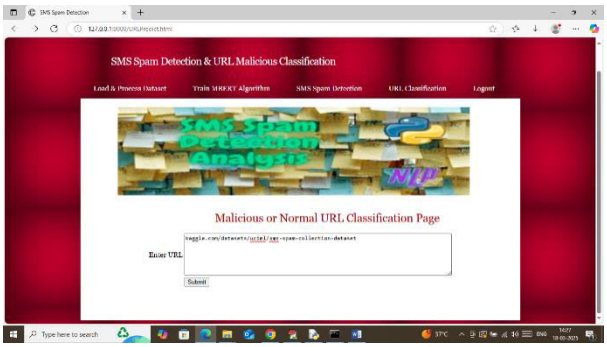

Fig 6.6: Malicious or Normal URL Classification Page

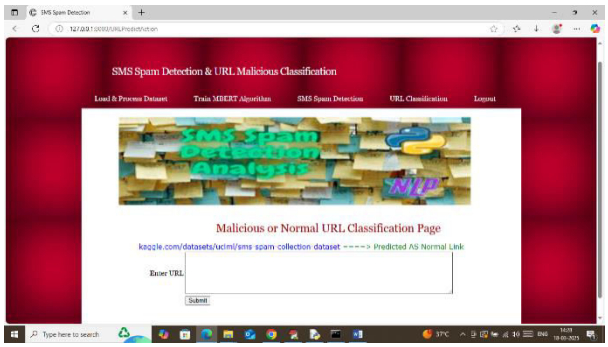

Fig 6.7: Result Page

VII. CONCLUSION

In an age where digital communication is pivotal, safeguarding users against unsolicited spam and malicious content is more critical than ever. This project presents a comprehensive and intelligent system that not only detects spam SMS messages but also performs real-time classification of URLs contained within those messages to identify potential security threats.

By leveraging the power of machine learning algorithms and natural language processing techniques, the system achieves high accuracy and adaptability in filtering out unwanted content and protecting users from phishing and malware attacks. The dual-layered approach enhances the system's robustness by covering both message-level and link-level threats, something traditional filtering methods often miss.

The modular architecture ensures that the system is scalable, maintainable, and deployable across multiple platforms, including mobile and web environments. Additionally, the integration of a feedback mechanism allows continuous improvement of the models, ensuring that the system evolves to counter new and sophisticated spam and phishing techniques.

Overall, this solution provides a proactive, intelligent, and user-friendly defense mechanism that can significantly reduce the risks associated with SMS-based attacks. With further enhancement and integration into commercial applications, it holds the potential to make mobile communication considerably safer and more reliable.

VIII. FUTURE SCOPE

Future enhancements for the SMS Spam Detection and URL Malicious Classification system can focus on integrating advanced deep learning models such as Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and Transformer-based architectures. These models are capable of capturing complex linguistic patterns and contextual relationships in textual data. By utilizing such deep learning techniques, the system can significantly improve feature extraction and classification accuracy. Transformer models in particular can understand long-range dependencies within messages, making them effective in identifying sophisticated spam or phishing attempts.

Another important future direction is the implementation of multilingual support. Many spam and phishing messages are sent in different languages and regional dialects, which current systems may not fully support. Expanding the framework to process and analyze messages in multiple languages will enhance its usability across diverse geographic regions. This improvement will enable the system to detect malicious content more effectively in global communication environments where users interact using different languages.

The system can also be improved through real-time mobile integration. Developing lightweight mobile application plugins or software development kits (SDKs) that directly integrate with messaging applications will allow real-time spam detection on smartphones. Such integration ensures that suspicious messages and malicious URLs are identified immediately when received, while maintaining low resource consumption so that device performance and battery life are not affected.

Another promising enhancement involves behavioral analysis. In addition to analyzing the message content, the system can examine metadata such as message sending frequency, sender reputation, and communication patterns. Behavioral insights can

help identify suspicious activities that traditional text-based analysis may miss. By combining behavioral indicators with machine learning predictions, the framework can achieve more reliable detection while reducing false positives.

To maintain robustness against evolving threats, the system should also address adversarial attack resilience. Attackers often modify spam messages or malicious URLs in ways designed to bypass machine learning detection models. Incorporating adversarial training techniques and robust model architectures can help the system recognize such evasive tactics. This will strengthen the detection framework and ensure that the model remains effective even when attackers attempt to manipulate input data.

Another future scope is the integration of comprehensive threat intelligence. By connecting the system to external cybersecurity databases, threat intelligence feeds, and phishing detection APIs, it can continuously receive updates about newly discovered malicious domains, phishing campaigns, and spam trends. This dynamic update mechanism allows the system to quickly adapt to emerging threats and maintain up-to-date protection for users.

The implementation of privacy-preserving techniques is also an important area of research. Methods such as federated learning and homomorphic encryption allow machine learning models to be trained collaboratively without directly sharing sensitive user data. This ensures that personal messages and communication data remain private while still enabling the improvement of detection models across distributed devices.

Finally, the system can be expanded to support multiple communication platforms beyond SMS. Modern users communicate through email, social media messaging services, and instant messaging applications. Extending the detection framework to analyze content across these channels will create a unified security solution capable of identifying spam and malicious content in all digital communication environments. This expansion would significantly enhance the overall effectiveness and applicability of the system.

IX. REFERENCES

[1] R. Mahajan and I. Siddavatam, “Phishing Website Detection Using Machine Learning Algorithms,” *International Journal of Computer Applications*, vol. 181, no. 23, pp. 45–47, 2018. doi: 10.5120/ijca2018918026.

[2] Z. Alshingiti et al., “A Deep Learning-Based Phishing Detection System Using CNN and LSTM,” *Electronics*, vol. 12, no. 1, 2023. doi: 10.3390/electronics12010232.

[3] S. K. H. Ahammad et al., “Phishing URL Detection Using Machine Learning Methods,” *Information Processing & Management*, vol. 59, no. 6, 2022. doi: 10.1016/j.ipm.2022.102973.

[4] S. Sindhu and S. P. Patil, “Phishing Detection Using Random Forest, SVM and Neural Network with Backpropagation,” in *Proc. International Conference on Smart Technologies in Computing, Electrical and Electronics*, 2020. doi: 10.1109/ICSTCEE49637.2020.9277256.

[5] A. Safi et al., “A Systematic Literature Review on Phishing Website Detection,” *Journal of King Saud University – Computer and Information Sciences*, vol. 35, 2023. doi: 10.1016/j.jksuci.2023.101593.

[6] H. C. Altunay et al., “SMS Spam Detection System Based on Deep Learning Using CNN and GRU,” *Applied Sciences*, vol. 14, no. 24, 2024. doi: 10.3390/app142411804.

[7] D. A. Oyeyemi et al., “SMS Spam Detection and Classification to Combat Abuse Using NLP and BERT,” *arXiv*, 2024. doi: 10.48550/arXiv.2406.06578.

[8] M. R. Al Saidat et al., “A Novel Approach for Arabic SMS Spam Detection Using Deep Learning,” *Procedia Computer Science*, 2024. doi: 10.1016/j.procs.2024.01.067.

[9] T. Choudhary et al., “A Machine Learning Approach for Phishing Attack Detection,” *Journal of Artificial Intelligence and Technology*, 2023. doi: 10.37965/jait.2023.0187.

[10] A. U. Rehman et al., “Real-Time Phishing URL Detection Using Machine Learning,” *Engineering Proceedings*, vol. 107, 2025. doi: 10.3390/engproc2025107108.

[11] M. F. Johari et al., “Key Insights into Recommended SMS Spam Detection Datasets and Algorithms,” *Scientific Reports*, 2025. doi: 10.1038/s41598-025-92223-1.

[12] S. Yerima and M. Alzaylaee, “High Accuracy Phishing Detection Based on Convolutional Neural

Networks," *arXiv*, 2020. doi: 10.48550/arXiv.2004.03960.
[13] V. Shahrivari, M. M. Darabi and M. Izadi, "Phishing Detection Using Machine Learning Techniques," *arXiv*, 2020. doi: 10.48550/arXiv.2009.11116.
[14] Y. Li et al., "SpamDam: Towards Privacy-Preserving and Adversary-Resistant SMS Spam Detection," *arXiv*, 2024. doi: 10.48550/arXiv.2404.09481.
[15] D. Sahoo, C. Liu and S. C. H. Hoi, "Malicious URL Detection Using Machine Learning: A Survey," *ACM Computing Surveys*, 2017. doi: 10.1145/3019289.