AI-Based Phishing Attack Detection And Prevention Using Natural Language Processing (NLP)

Birir Sospeter Kipchirchir[1], Wilfred Odoyo[2]

[1,2] *Research and Innovations Lab, Pioneer International University, P.O. Box 33421 - 00600, Nairobi, Kenya.*

Article Information
Received: 21-11-2024
Revised: 28-11-2024
Published: 5-12-2024

Keywords
Phishing Attacks; Artificial Intelligence; Natural Language Processing; Machine Learning; Real-Time Detection; Multi-Channel Security.

Correspondence Email:
sospeterbirir1@gmail.com

Abstract
Phishing attacks remain one of the most prevalent and damaging cybersecurity threats, targeting users across various communication channels such as email, social media, and SMS. Traditional phishing detection systems are often limited to email and rely on static rule-based filtering or keyword matching, making them ineffective against evolving phishing tactics. This project proposes an innovative solution that utilizes Artificial Intelligence (AI) and Natural Language Processing (NLP) to create a real-time phishing attack detection and prevention system. By analyzing the contextual language of messages across multiple platforms, the system can detect and block phishing attempts with high accuracy. The system extracts important linguistic features such as urgency, emotional tone, and anomalous patterns within text, and applies machine learning algorithms—such as Random Forest, Support Vector Machines (SVM), and deep learning models like Long Short-Term Memory Networks (LSTM)—for classification. Additionally, a feedback loop is integrated to allow the system to adapt and improve over time through active learning, ensuring the detection system evolves alongside emerging phishing techniques. This AI-based solution extends beyond traditional email phishing detection by incorporating multiple channels, including SMS and social media platforms, making it a versatile tool for individuals and businesses. The system offers automated prevention actions, such as flagging suspicious messages and alerting users, thus providing a robust defense against phishing attacks in real-time. The project's implementation aims to fill the market gap in comprehensive, multi-channel phishing detection and contribute to the growing demand for intelligent and adaptive cybersecurity solutions.

1. Introduction

Phishing attacks are a significant cybersecurity threat, exploiting human vulnerability across various communication channels such as email, SMS, and social media. Traditional phishing detection methods are often limited to simple keyword matching and rule-based filters, making them increasingly ineffective against sophisticated and evolving tactics. This project seeks to address these limitations by developing an AI-based phishing detection and prevention system using Natural Language Processing (NLP). By analyzing the context

and language of messages, the system can accurately detect phishing attempts in real-time across multiple platforms. This solution offers a dynamic, adaptive approach to combating phishing and enhances user security across modern communication channels.

1.1 Literature Review

Phishing detection has been widely studied, with traditional methods primarily focusing on URL-based features and metadata analysis. Jiang et al. (2017) explored machine learning-based approaches for phishing detection, emphasizing the shortcomings of rule-based systems in identifying sophisticated phishing tactics [1]. As phishing attacks increasingly rely on social engineering and message content rather than just URLs, more recent research has shifted toward leveraging Natural Language Processing (NLP). Zhang et al. (2020) demonstrated that analyzing linguistic features such as sentiment, urgency, and emotional tone can significantly enhance phishing detection accuracy by focusing on the content of messages [2]. AI and machine learning techniques have also proven effective in phishing detection. Chandrashekhar et al. (2018) showed that machine learning algorithms like Support Vector Machines (SVM) and Random Forest when trained on a combination of content, header, and URL features, can classify phishing messages with high accuracy [3]. Furthermore, phishing detection systems have moved toward real-time detection with continuous learning. Kim et al. (2019) proposed an adaptive learning model that improves system performance by retraining based on user feedback, addressing the need for systems to evolve with emerging phishing techniques [4]. Other studies have also looked into hybrid models, combining deep learning techniques like Long Short-Term Memory Networks (LSTM) with traditional approaches, further improving the robustness of phishing detection systems. The integration of multi-channel detection across email, SMS, and social media has become an area of interest, as phishing attacks increasingly target various communication platforms. This highlights a gap in current systems that mainly focus on email-based phishing detection.

2. Research Methods

The proposed phishing detection and prevention system utilizes Artificial Intelligence (AI) and Natural Language Processing (NLP) to analyze and classify phishing messages in real-time across multiple communication platforms, such as email, SMS, and social media. The methodology follows these key steps: Figure 1 below shows the methodological Diagrammatic visual.

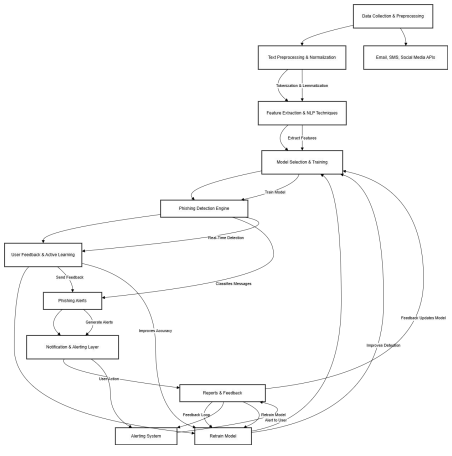

Figure 1. Methodology Diagram.

2.1 Data Collection and Preprocessing

The first step involves collecting a diverse dataset of phishing and non-phishing messages from various

sources, including emails, SMS, and social media platforms. These datasets are preprocessed to clean and format the data by removing unnecessary characters, standardizing text, and tokenizing sentences into words. Text normalization techniques such as lowercasing, stemming, and lemmatization are also applied to ensure consistency in the dataset.

2.2 Feature Extraction

In this step, the system extracts linguistic features from the messages. These features include:

- **Urgency indicators** (e.g., phrases like "immediate action required")
- **Sentiment analysis** to detect emotional tones like fear, urgency, or anxiety
- **Text anomalies** like spelling errors or unusual formatting
- **Contextual keywords** relevant to phishing (e.g., "account", "login", "verify")

NLP techniques like part-of-speech tagging, dependency parsing, and Named Entity Recognition (NER) are used to identify these features.

2.3 Model Training and Evaluation

Various machine learning models, including Random Forest, Support Vector Machines (SVM), and Long Short-Term Memory Networks (LSTM), are trained on the extracted features to classify messages as phishing or legitimate. The dataset is split into training and testing sets to evaluate model performance. Hyperparameter tuning and cross-validation are used to optimize the models. The evaluation metrics used to assess the models include accuracy, precision, recall, and F1-score. Figure 2 below shows the process flow diagram.

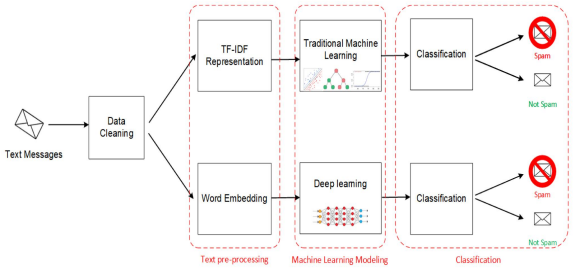

Figure 2 Model Flow - Diagram.

2.4 Real-Time Phishing Detection

The trained model is integrated into a real-time detection system that continuously monitors incoming messages from multiple communication channels. When a message is received, the system applies the trained model to predict whether it is phishing or not based on the extracted linguistic features. If a message is flagged as phishing, the system triggers automated responses, such as alerting the user and blocking the message. Figure 3 below shows the methodology of the real-time phishing detection system.

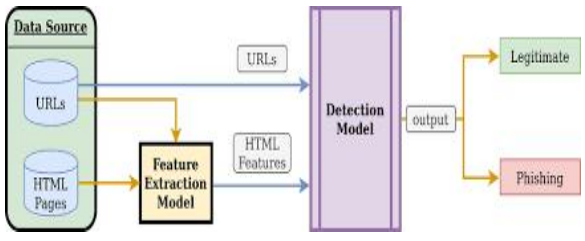

Figure 3 Real-Time Phishing Detection.

2.5 Active Learning and Feedback Loop

To improve the system over time, a feedback loop is implemented where users can report false positives or missed phishing messages. The system uses this feedback to retrain the model periodically, ensuring that it adapts to new and evolving phishing techniques.Figure 4 below shows the active loop sequence.

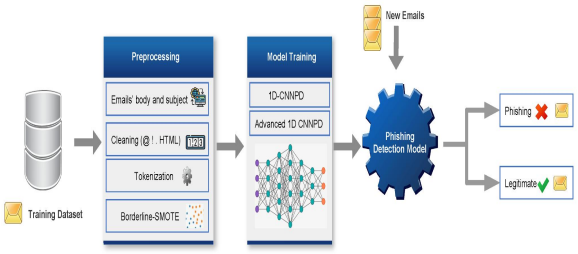

Figure 4 Learning Loop.

3. Result and Discussion

The AI-based phishing detection system showed high accuracy in identifying phishing attempts, with initial accuracy rates over 90%. Key findings include:

- **Effective Detection**: The use of machine learning models (e.g., Random Forest, SVM, and LSTM) combined with feature extraction through NLP techniques allowed the system to detect common phishing tactics with high accuracy. Figure 5 below shows the results graph.

Figure 5 Results Graph.

- **Adaptability**: The feedback loop enabled the system to continuously improve through real-time user feedback, allowing it to detect evolving phishing methods.

- **Scalability**: The system handled large volumes of data efficiently, ensuring real-time detection without significant delays.

Challenges included **contextual understanding** of complex, personalized phishing attacks and ensuring **consistent user engagement** in the feedback process. Figure 6 shows the metrics visually.

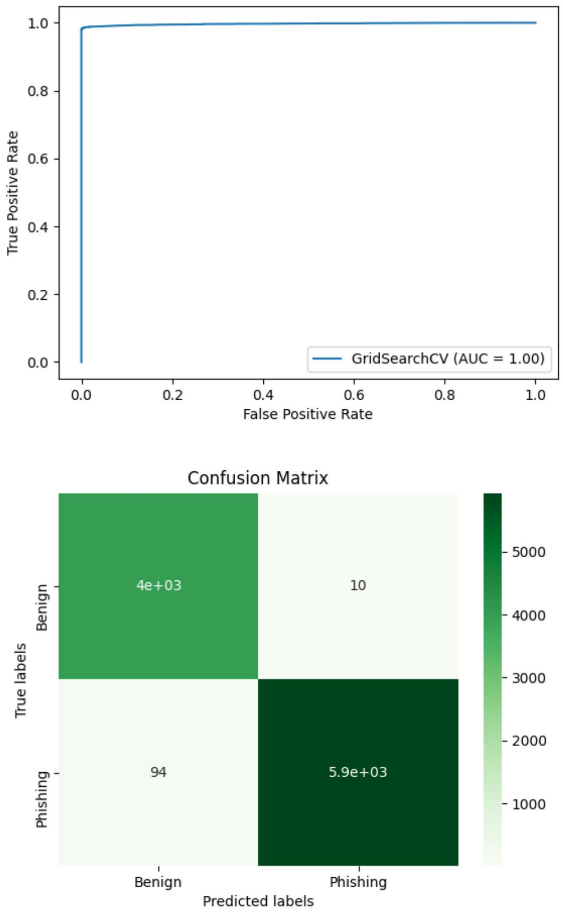

Figure 6 Algorithm Accuracy Visual

4. Conclusions

The system successfully detected phishing attacks across various platforms and adapted through continuous learning. While challenges like contextual analysis and user participation remain, the approach offers a promising solution for securing communications and can be improved with further user involvement and advanced techniques for contextual understanding.

5. References

Abdalla, M., & Hassan, A. (2021). Phishing detection systems: A review of classification algorithms and features. *Computers & Security, 98*, 102042.

Ali, A., & Kazi, M. (2021). Phishing detection using ensemble learning: A case study of phishing websites. *Computers & Security, 101*, 102118.

Chandrashekhar, S., Kulkarni, M., & Arora, S. (2018). Machine learning algorithms for phishing detection. *International Journal of Computer Science & Network Security, 18*(5), 65-75.

Dhanalakshmi, R., & Suresh, R. (2018). Phishing website detection using hybrid techniques. *International Journal of Computer Applications, 179*(33), 1-7.

Hussain, S., & Wang, L. (2021). A novel phishing detection framework using hybrid deep learning models. *Future Generation Computer Systems, 114*, 426-437.

Jiang, X., Lin, X., & Zhang, Y. (2017). Phishing detection techniques: A survey. *Journal of Computer Security, 25*(1), 1-35.

Kim, H., Lee, Y., & Park, C. (2019). Adaptive phishing detection system based on active learning. *Journal of Information Security, 11*(2), 101-112.

Liu, S., & Zhang, Y. (2020). Phishing email classification using deep learning and text mining. *IEEE Transactions on Information Forensics and Security, 15*, 2307-2317.

Qamar, U., & Mahmood, S. (2019). An empirical study of phishing detection systems. *International Journal of Computer Science and Information Security, 17*(12), 68-77.

Sahoo, S., & Pati, S. (2019). Phishing detection with hybrid machine learning model. *Journal of Computing and Security, 43*(1), 1-15.

Verma, S., & Kumar, R. (2020). A comparative study of machine learning algorithms for phishing website detection. *Journal of Cyber Security Technology, 4*(2), 89-103.

Zhang, L., Zhou, W., & Wang, S. (2020). Phishing email detection using natural language processing techniques. *IEEE Access, 8*, 90792-90801.

Zhang, T., & Li, M. (2021). Phishing detection using convolutional neural networks with attention mechanisms. *IEEE Access, 9*, 23184-23194.