Phishing Email Detection Using Artificial Intelligence: An Integrated Approach

Abstract

Phishing email attacks continue to pose a significant threat to cybersecurity, leading to financial losses and privacy breaches. This paper presents an integrated review of current AI-driven approaches for phishing email detection, exploring machine learning (ML) and natural language processing (NLP) techniques. The study reviews various models, including ensemble methods, large language models (LLMs), and deep learning techniques, to understand their effectiveness in tackling phishing emails, especially those powered by generative AI.

Keywords: Phishing Detection, AI, Machine Learning, Natural Language Processing, Cybersecurity

1. Introduction

Phishing attacks exploit human vulnerabilities to steal sensitive information. Traditionally, phishing emails are detected using heuristic-based approaches, but these methods are struggling to cope with the sophistication of AI-powered phishing campaigns. Recent advances in AI, particularly through machine learning models and generative AI, present new challenges and opportunities in phishing detection.

2. Literature Review

1) Traditional Phishing Detection Methods: Relied on blacklists, URL pattern analysis, and heuristics, which are less effective against modern phishing tactics. 2) Machine Learning and Deep Learning Approaches: Models like Naive Bayes, SVM, Random Forest, and XGBoost achieve high accuracy. CNNs automate feature extraction for better performance. 3) Ensemble Methods: Techniques like AdaBoost, Gradient Boosting, and Stacking improve accuracy and reduce false positives. 4) Large Language Models (LLMs): GPT-4, Gemini, and Claude detect phishing indicators using linguistic analysis. 5) Generative AI in Phishing Attacks: Allows creation of personalized phishing emails that evade traditional filters, requiring advanced AI models for detection.

3. Methodology

This study evaluates AI-driven phishing detection methods using datasets of phishing and legitimate emails. We assess Machine Learning models (Random Forest, XGBoost, SVM), Ensemble Learning (AdaBoost, Stacking), Large Language Models (GPT-4, Gemini, Claude), and NLP-based semantic analysis for phishing indicators.

4. Results and Discussion

Results indicate that fine-tuned LLMs outperform traditional models in phishing detection. Combining ensemble learning with LLMs provides robust solutions, achieving high accuracy in detecting sophisticated phishing attacks. AI-driven approaches also excel in real-time detection and adapt to evolving threats.

5. Challenges and Future Directions

Challenges include data privacy concerns, continuous model updates, and adapting to evolving attack patterns. Future research should focus on enhancing adaptability and integrating Explainable AI (XAI) to improve transparency and user trust.

6. Conclusion

AI-based approaches, especially those combining ensemble methods and LLMs, are highly effective in phishing detection. However, with phishing tactics constantly evolving, continuous advancements are essential to maintain effective cybersecurity defenses.