RADemics

Natural Language Processing-Based AI for Real-Time Phishing and Social Engineering Attack Detection in Email and Messaging Systems

Shruthi N, Suresh Kadarkarai, S. Nanthini
SJCE, JSS STU, KARPAGA VINAYAGA COLLEGE OF ENGINEERING AND TECHNOLOGY, SIMATS.

Natural Language Processing-Based AI for Real-Time Phishing and Social Engineering Attack Detection in Email and Messaging Systems

[1]Shruthi N, Assistant Professor, Computer Science and Engineering, SJCE, JSS STU, MYSURU, nshruthi@jssstuniv.in

[2]Suresh Kadarkarai, Assistant Professor, Electrical and Electronics Engineering, Karpaga Vinayaga College of Engineering and Technology, Chinna Kolampakkam, Mail id: srshk549@gmail.com

[3]S. Nanthini, Professor, Saveetha School of Engineering, SIMATS, Chennai, nanthini27j88@gmail.com,

Abstract

The increasing sophistication of phishing and social engineering attacks poses a significant threat to email and messaging security. Traditional rule-based and heuristic-driven approaches are often ineffective against evolving attack methodologies that exploit linguistic deception and psychological manipulation. Natural Language Processing (NLP)-based Artificial Intelligence (AI) has emerged as a powerful solution for real-time detection of phishing attempts by analyzing textual patterns, semantic structures, and contextual cues. However, deploying NLP-driven phishing detection in high-throughput email environments presents critical challenges, including scalability, concept drift, adversarial evasion, and multilingual attack vectors. This book chapter explores advanced NLP techniques, such as deep learning-based transformers, contextual embeddings, and hybrid AI models, for enhancing phishing and social engineering attack detection. it examines adaptive learning strategies to address concept drift, adversarial resilience mechanisms to counter evasive phishing tactics, and real-time processing architectures to ensure high-speed threat identification. Privacy-preserving AI frameworks and explainable NLP models are also discussed to enhance transparency and regulatory compliance in cybersecurity applications. By integrating state-of-the-art AI methodologies with real-time security monitoring, this chapter provides a comprehensive roadmap for developing robust, scalable, and intelligent phishing detection systems capable of mitigating emerging cyber threats in dynamic email and messaging ecosystems.

Keywords: Phishing Detection, Social Engineering, Natural Language Processing, Deep Learning, Adversarial Resilience, Concept Drift.

Introduction

Phishing and social engineering attacks have become one of the most pervasive cyber threats, targeting individuals and organizations through deceptive email and messaging tactics. These

attacks exploit human psychology, trust mechanisms, and communication vulnerabilities to manipulate users into divulging sensitive information, such as login credentials, financial details, and corporate secrets. Traditional security mechanisms, such as rule-based email filtering and signature-based detection, have proven inadequate in identifying advanced phishing attempts that employ sophisticated language patterns, social engineering strategies, and obfuscation techniques. The increasing reliance on digital communication channels further amplifies the risks, as cybercriminals continue to refine their attack methodologies to evade detection. The emergence of artificial intelligence (AI)-driven security solutions, particularly Natural Language Processing (NLP), has provided a promising avenue for enhancing phishing detection by analyzing textual characteristics, contextual relationships, and intent-driven deception patterns within messages.

Natural Language Processing (NLP)-based AI models leverage deep learning techniques, semantic understanding, and contextual embeddings to automatically detect fraudulent emails and messages. Unlike traditional heuristic-based methods, NLP models analyze linguistic structures, identify anomalous patterns, and detect subtle manipulations within the text that are indicative of phishing attempts. Transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) have demonstrated significant advancements in understanding text meaning, enabling precise classification of phishing emails. However, implementing NLP-driven phishing detection in large-scale, real-time email environments presents several challenges, including processing latency, computational resource constraints, and the need for continuous model adaptation. Addressing these challenges is crucial for ensuring the effectiveness of AI-driven security frameworks in detecting and mitigating phishing threats in modern communication systems.

A major challenge in NLP-based phishing detection is the phenomenon of concept drift, where phishing techniques evolve dynamically, rendering static detection models ineffective. Attackers frequently modify their language patterns, exploit emerging social engineering trends, and craft messages tailored to specific targets to increase the likelihood of deception. As a result, phishing detection models trained on historical data often struggle to generalize to new attack vectors. To mitigate the impact of concept drift, continuous learning techniques such as online learning, incremental model updates, and active learning strategies must be integrated into NLP-based security frameworks. These approaches enable AI models to adapt to evolving phishing tactics by incorporating real-time threat intelligence and user feedback, ensuring sustained detection accuracy in dynamic cyber environments.