DOI: doi.org/10.47363/JAICC/ICADCCS2026/2026(5)12

International Conference on AI, Data Science, Cybersecurity, Cloud Architectures, and Software Engineering

Conference Proceedings April 23, 2026 - (Virtual)

DecepNet: A Hybrid NLP Framework for Robust Phishing Detection Under Adversarial Conditions

Gargi Choudhury and Shivam Choudhury

Independent Researchers, India

Abstract

Phishing attacks continue to evolve with increasing sophistication, particularly with the emergence of AI-generated content that closely mimics legitimate communication. Traditional phishing detection systems, which rely on static text classification models, often struggle to generalize across diverse datasets and fail under subtle adversarial modifications. In this work, we propose DecepNet, a hybrid phishing detection framework that integrates natural language processing with structural and behavioral feature analysis to enhance detection performance and robustness.

The proposed system combines TF-IDF-based representations and transformer-based semantic modeling with a learned deception function designed to capture persuasive intent through linguistic, contextual, and behavioral signals such as urgency, authority cues, and threat-reward framing. The model is trained on a multi-source dataset constructed from publicly available corpora, including Apache SpamAssassin and Enron Spam Dataset, along with additional phishing samples derived from URL-based sources and synthetically generated adversarial data.

To evaluate robustness, we introduce an adversarial testing setup involving paraphrasing, tone modification, and obfuscation techniques. Experimental results demonstrate that conventional machine learning models experience noticeable performance degradation under such conditions, while the proposed framework maintains improved stability and generalization across datasets. Comparative analysis with baseline models, including Logistic Regression, Random Forest, and transformer-based classifiers, highlights the effectiveness of integrating deception-aware features.

This study emphasizes the importance of combining semantic understanding with behavioral modeling in phishing detection and provides a practical, adaptable framework for addressing evolving social engineering threats in real-world environments.

Keywords: Phishing Detection, Natural Language Processing (NLP), Adversarial Machine Learning, Deception Modeling, Transformer Models, Cybersecurity, Email Security, Hybrid Machine Learning