ABSTRACT

Identifying phishing emails poses a significant challenge in the realm of cybersecurity, as malicious actors continually modify their tactics to take advantage of vulnerabilities in communication systems. This study investigates the efficacy of the Bidirectional Encoder Representations from Transformers (BERT) model in identifying phishing emails, with particular attention to the impact of dataset size and diversity. Two experimental scenarios were conducted: In Scenario 1, the effectiveness of BERT was evaluated using various unique phishing email datasets. In contrast, Scenario 2 applied BERT to a larger, combined dataset that included 203,176 emails. The results of Scenario 1 demonstrate that BERT outperforms conventional machine learning models, including SVM, RF, ET, XGB, and ADB, across various datasets. BERT achieved an accuracy of 99.64% on the Ling dataset, 99.43% on the Enron dataset, and 99.82% on the TREC-07 dataset. The AUC-ROC analysis for Scenario 1 reveals exceptional outcomes, with BERT achieving an AUC of 99.88% or greater across all datasets. In Scenario 2, a larger and more diverse dataset allowed BERT to achieve an accuracy of 99.35%, precision of 99.45%, recall of 99.04%, and an F-score of 99.24%, as well as an AUC-ROC of 99.97%. This analysis demonstrates that BERT consistently outperforms other models in distinguishing between phishing and legitimate emails, irrespective of the dataset size. The findings contribute to the enhancement of more efficient detection systems and hold considerable importance for bolstering cybersecurity strategies against phishing attacks in real-world scenarios.

Index Terms

Email Phishing Detection, Cyber Attack, BERT, Tranformers, Deep Learning, Machine Learning