Building a phishing website classification system based on machine learning

Pham Quoc Dat

Abstract: Research and development project on a solution for detecting phishing websites using machine learning techniques, replacing the traditional blacklist matching method. The author extracts feature groups from URLs, such as lexical features and statistical features, to train models like SVM and Random Forest. The process includes data collection, preprocessing using the Pandas library, and feature normalization to optimize classification performance. Experimental results through Ablation Study show that URL structural features play a key role, helping the model achieve high performance with an impressive F1-score.

Keywords: Fake website (Phishing), Machine Learning, URL feature extraction

References

[1] Cục An toàn thông tin – Bộ Thông tin và Truyền thông, "Báo cáo tổng quan an toàn thông tin mạng Việt Nam," 2024. [Online]. Available:

[2] NCSC, "Cảnh báo và thống kê tình hình lừa đảo trực tuyến tại Việt Nam," 2024. [Online]. Available:

[3] VNCERT/CC, "Hướng dẫn phòng chống tấn công phishing," 2024. [Online]. Available:

[4] Khoa CNTT - ĐH Bách Khoa Hà Nội, Giáo trình Học máy (Machine Learning). Hà Nội, 2023.

[5] VietAI, "Tổng quan về các thuật toán học máy và ứng dụng," 2023. [Online]. Available:

[6] N. T. Bình et al., "Ứng dụng học máy trong phát hiện website phishing," Tạp chí Khoa học & Công nghệ Thông tin Việt Nam, 2023.