ABSTRAK

Keamanan siber menjadi perhatian utama seiring dengan meningkatnya ancaman dari URL berbahaya yang dapat digunakan untuk menyebarkan malware, melakukan serangan phishing, dan aktivitas siber berbahaya lainnya. Ancaman ini dapat menyebabkan kerugian finansial serta pencurian data sensitif. Dalam upaya meningkatkan akurasi deteksi URL berbahaya, penelitian ini menerapkan algoritma Extreme Gradient Boosting (XGBoost) dengan memanfaatkan analisis fitur leksikal. Dengan pendekatan ini, sistem deteksi dapat mengidentifikasi pola yang mencurigakan dalam struktur URL tanpa ketergantungan pada daftar hitam (blacklist), sehingga lebih adaptif terhadap ancaman baru.

Metodologi penelitian ini dimulai dengan pengumpulan dataset yang terdiri dari 633.010 URL, terbagi secara seimbang antara URL benign dan malicious. Ekstraksi fitur leksikal dilakukan untuk mendapatkan karakteristik URL yang berkontribusi dalam proses klasifikasi. Sebanyak 20 fitur leksikal digunakan, di antaranya secure_http, url_length, domain_entropy, special_char_count, tld_reputation, path_length, digit_letter_ratio, subdomain_dot_count, query_string_length, protocol_count, dan suspicious_path_words. Model XGBoost digunakan sebagai algoritma utama dalam pengklasifikasian URL, dengan parameter yang dioptimalkan menggunakan GridSearchCV. Evaluasi model dilakukan dengan 8-fold cross-validation menggunakan metrik seperti accuracy, precision, recall, dan F1-score.

Hasil penelitian menunjukkan bahwa penggunaan fitur leksikal dalam model XGBoost secara signifikan meningkatkan akurasi deteksi URL berbahaya. Pada pengujian pertama dengan 5 fitur, akurasi yang diperoleh adalah 88%. Dengan penambahan fitur menjadi 10, akurasi meningkat menjadi 90%. Selanjutnya, pada percobaan dengan 15 fitur, akurasi mencapai 97%, dan pada percobaan dengan 20 fitur, model mencapai akurasi tertinggi sebesar 98,2%. Evaluasi tambahan melalui 8-fold cross-validation menunjukkan bahwa semakin banyak fitur relevan yang digunakan, semakin baik model dalam mengidentifikasi URL benign dan malicious. Hasil ini mengonfirmasi bahwa pendekatan berbasis fitur leksikal dengan algoritma XGBoost memiliki potensi besar dalam meningkatkan keamanan siber melalui deteksi URL berbahaya yang lebih akurat dan adaptif.

Kata Kunci: URL Berbahaya, Fitur Leksikal, XGBoost, Keamanan Siber, Deteksi Phishing.

ABSTRACT

Cybersecurity has become a crucial concern with the increasing use of the internet and the rising threats of malicious URLs. Malicious URLs are often used to distribute malware, conduct phishing attacks, and carry out other cyber threats that can lead to financial losses and the theft of sensitive data. This study aims to enhance the accuracy of malicious URL detection by implementing the Extreme Gradient Boosting (XGBoost) algorithm and leveraging lexical feature analysis.

The research process began with data collection, followed by lexical feature extraction. The extracted features were selected based on their importance in contributing to classification performance. Pre-processing was conducted by converting string data into numerical encoding. The dataset used in this study consists of 633,010 URLs, evenly distributed between benign and malicious categories. The classification process employed XGBoost with hyperparameter tuning using GridSearchCV. The model was evaluated using 8-fold cross-validation, utilizing metrics such as accuracy, precision, recall, and F1-score.

The experimental results demonstrate significant improvements in accuracy with the addition of selected lexical features. The initial model, trained with five features, achieved an accuracy of 88%. By expanding the feature set to ten, accuracy increased to 90%. Further increasing the features to fifteen resulted in an accuracy of 97%, while the final experiment, utilizing twenty features, attained the highest accuracy of 98.2%. The evaluation confirmed that incorporating relevant lexical features significantly improves the model's ability to distinguish between benign and malicious URLs. These findings highlight the effectiveness of lexical feature-based detection in cybersecurity, providing a reliable solution for identifying emerging threats without relying on external blacklists.

Keywords: Malicious URL, Lexical Features, XGBoost, Cybersecurity, Phishing Detection.