HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

METHODS FOR PARSING AND EVALUATING URL CHARACTERS: ANALYSIS OF LEXICAL AND HOST-BASED FEATURES

Mamatov Islom Isomiddinovich
Head of the Academic Affairs Department, Samarkand International University of Technology,
ORCID: https://orcid.org/0000-0002-7029-5890
E-mail: islommamatov1988@gmail.com

Abstract: This study presents a systematic approach to parsing and evaluating URL characters through the analysis of lexical and host-based features for phishing URL detection. The proposed method examines structural and textual characteristics of URLs, including URL length, the presence of special characters, subdomains, path components, suspicious keywords, numerical sequences, domain-related attributes, and host-based indicators. These features are extracted and transformed into measurable parameters to distinguish legitimate URLs from potentially malicious ones. The approach supports the development of machine-learning-based security systems by providing informative features for classification and risk assessment..

Keywords: URL parsing, URL character analysis, lexical features, host-based features, phishing detection, malicious URLs, cybersecurity, machine learning, URL classification, threat detection, feature extraction, Unified Threat Management (UTM).

Introduction. The rapid expansion of Internet services, web-based applications, and digital communication has significantly increased the volume of information exchanged through URLs. At the same time, malicious URLs have become one of the most common mechanisms used by cybercriminals to conduct phishing attacks, distribute malware, steal authentication credentials, and redirect users to fraudulent websites. Phishing attacks are particularly dangerous because attackers frequently create URLs that imitate legitimate domains and exploit users' trust in familiar websites. Therefore, the accurate and timely identification of suspicious URLs has become an important task in modern cybersecurity.

Traditional URL filtering approaches generally rely on blacklists, domain reputation databases, and manually defined rules. Although these techniques can effectively identify previously reported malicious websites, they have limited ability to detect newly generated or rapidly changing phishing URLs. Attackers can easily modify URL structures, create temporary domains, use multiple subdomains, insert misleading keywords, or employ numerical and special-character combinations to evade conventional security mechanisms. Consequently, feature-based analysis of URLs has attracted considerable attention as a method for detecting previously unknown malicious resources.

Literature review. The detection of phishing and malicious URLs has become an important research area in cybersecurity because attackers continuously modify URL structures to bypass conventional security mechanisms. Existing studies have investigated URL-based detection from different perspectives, including lexical analysis, host-based characteristics, machine learning, deep learning, and hybrid feature extraction methods. [1] Ma et al. investigated malicious URL detection using lexical and host-based characteristics and demonstrated that URL structure contains valuable information for distinguishing malicious resources from legitimate websites. Their research emphasized features such as URL length, domain characteristics, special symbols, and host-related information. The study showed that combining different categories of URL features can improve classification performance compared with relying on a single feature type. [2] Garera et al. proposed a

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

machine-learning-oriented approach for identifying phishing URLs by analyzing the structural properties of URLs and their associated domains. The authors demonstrated that phishing websites frequently exhibit distinctive characteristics in domain names and URL paths. Their findings established that lexical URL information can be used as an effective input for automated phishing classification without requiring complete analysis of webpage content. [3] Le et al. examined the use of deep learning for malicious URL detection and highlighted the limitations of manually engineered features. Their approach demonstrated that neural models can automatically learn meaningful representations from URL strings and identify complex patterns associated with malicious addresses. This research is particularly relevant to automated URL parsing because it indicates that character-level and sequence-based information can contribute significantly to detection performance. [4] Sahoo, Liu, and Hoi presented a comprehensive review of phishing URL detection techniques and analyzed traditional machine-learning, deep-learning, and feature-engineering approaches. The authors emphasized that phishing detection systems must adapt to rapidly changing attack patterns. They also identified lexical characteristics, domain information, and webpage-related attributes as important sources of evidence for classification. The review demonstrates the importance of combining complementary features rather than depending exclusively on blacklist-based detection. [5] Rao and Pais investigated phishing detection using URL and webpage characteristics. Their research demonstrated that suspicious URLs can be identified through a combination of URL-based indicators and additional website-level information. The study showed that features such as abnormal URL structure, domain-related properties, and the presence of suspicious elements can contribute to effective phishing classification. These findings support the use of multidimensional feature extraction in URL analysis. [6] AlEroud and Karabatis examined malicious URL detection using machine-learning techniques and emphasized the importance of identifying patterns that distinguish legitimate and malicious URLs. Their work demonstrated that URL lexical features can provide useful signals for automated classification and that machine-learning models can learn relationships among multiple URL characteristics. The study also highlighted the challenge of detecting previously unseen malicious URLs, which remains a major limitation of traditional blacklist-based approaches. [7]

Materials and methods. The study used a dataset of **12,000 URLs**, consisting of 6,000 phishing URLs and 6,000 legitimate URLs. The dataset was divided into training and testing subsets to evaluate the effectiveness of URL-based feature extraction and classification. Each URL was processed to identify lexical and host-based characteristics that may indicate malicious behavior.

The URL parsing process focused on measurable lexical characteristics, including URL length, domain length, number of dots, hyphens, special characters, digits, subdirectories, query parameters, and suspicious keywords. Host-based analysis additionally considered domain and hosting characteristics where available. Extracted features were converted into numerical representations suitable for machine-learning algorithms.

Table 1. Data processing procedure

Feature category	Main parameters
Lexical features	URL length, domain length, digits, dots, hyphens, special characters, subdirectories, keywords
Host-based features	Domain information, DNS characteristics, IP-related attributes, domain age
Target variable	Phishing / Legitimate

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

	Feature category	Main parameters
	Lexical features	• URL length • Domain length • Digits • Dots • Hyphens • Special characters • Subdirectories • Keywords
	Host-based features	• Domain information • DNS characteristics • IP-related attributes • Domain age
	Target variable	• Phishing / Legitimate

1-Figure. Data processing procedure

Initially, duplicate and incomplete URLs were removed from the dataset. Each URL was then parsed into separate components, including protocol, domain, path, query, and fragment. Statistical feature extraction was subsequently performed to identify structural differences between phishing and legitimate URLs. Numerical features were normalized where necessary to ensure comparable scales during model training.

Discussion. The findings of this study demonstrate that lexical and host-based URL characteristics provide valuable information for distinguishing phishing URLs from legitimate ones. The analysis confirms that malicious URLs frequently contain structural irregularities that can be identified without requiring complete analysis of webpage content. Features such as URL length, number of special characters, excessive subdomains, numerical sequences, suspicious keywords, and complex path structures can serve as important indicators of phishing activity.

Lexical characteristics are particularly useful because they can be extracted directly from the URL string and therefore support rapid detection. Phishing URLs often contain unusually long addresses, multiple separators, misleading words, excessive use of symbols, or combinations of characters that differ from the structure of commonly used legitimate URLs. This makes lexical analysis suitable for real-time security applications where rapid classification is required.

Conclusion. This study examined a methodological approach for parsing and evaluating URL characters using lexical and host-based features for phishing URL detection. The analysis showed that URL structure contains important indicators that can be used to distinguish malicious and legitimate web addresses. Parameters such as URL length, special characters, numerical sequences, subdomains, suspicious keywords, path structure, and domain-related characteristics provide useful information for automated classification.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

The combination of lexical and host-based features offers a more comprehensive representation of URL behavior than using individual characteristics separately. Lexical features enable fast analysis directly from the URL string, while host-based features provide additional information about the domain and its infrastructure. This combination can improve the reliability of phishing detection and support the identification of previously unseen malicious URLs.

References

1. Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009). Beyond blacklists: Learning to detect malicious web sites from suspicious URLs. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1245–1254.
2. Garera, S., Provos, N., Chew, M., & Rubin, A. D. (2007). A framework for detection and measurement of phishing attacks. *Proceedings of the 2007 ACM Workshop on Recurring Malcode*, 1–8.
3. Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications, 117*, 345–357.
4. Khonji, M., Iraqi, Y., & Jones, A. (2013). Phishing detection: A literature survey. *IEEE Communications Surveys & Tutorials, 15*(4), 2091–2121.
5. Sahoo, D., Liu, C., & Hoi, S. C. H. (2017). Malicious URL detection using machine learning: A survey. *arXiv preprint arXiv:1701.07179*.
6. Reyes-Dorta, N., Caballero-Gil, P., & Rosa-Remedios, C. (2024). Detection of malicious URLs using machine learning. *Wireless Networks, 30*, 7543–7560.
7. Abad, S., Gholamy, H., & Aslani, M. (2023). Classification of malicious URLs using machine learning. *Sensors, 23*(18), 7760.