International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Phishing URL Classification based on Feature Engineering

*Phyo Phyo Wai, Aye Thida Win, Yimon Aung**

Associate Professor, Faculty of Information Science, University of Computer Studies(Magway), Myanmar
Associate Professor, Faculty of Computer Science, University of Computer Studies(Dawei), Myanmar
Associate Professor, Faculty of Information Science, University of Computer Studies(Dawei), Myanmar

ABSTRACT

Phishing is a form of cyberattack where attackers disguise phishing URLs as legitimate ones to steal sensitive information. The internet has become essential to daily social and financial activities. However, it exposes users to various web-based threats, such as phishing, which can cause financial damages and loss of personal data. According to the Anti-Phishing Working Group (APWG), phishing attacks reached a record high in 2023, with over 4.9 million incidents reported. This study explores the use of machine learning, specifically Support Vector Machines (SVM), for classifying URLs as phishing or legitimate. This study performs a feature engineering: the original 17 based features proposed by Banik et al. (2020)[1] and an extended feature set developed for enhanced phishing detection. Our results show that the extended feature set improves classification accuracy, offering a more effective approach to phishing URL detection.

Keywords: Phishing, Support Vector Machine, Lexical Features, Extended Feature Set.

1. Introduction

The internet is a vital role for communication, commerce, and education, but increasing online activity has led to a rise in cyber threats, especially phishing. Phishing involves tricking users into sharing sensitive information by impersonating trusted sources. Detecting phishing URLs is difficult due to attackers mimicking legitimate sites. Traditional rule-based methods struggle against evolving tactics, prompting the need for more adaptive solutions. Common approaches include blacklists, DNS filtering, user awareness training, and machine learning-based detection systems.

1. Blacklists: These are maintained by security companies like Google and Microsoft. Blacklists are updated regularly through automated crawlers and user reports, blocking access to these sites across various platforms. Popular tools like Google's Safe Browsing and Microsoft's SmartScreen filter rely on blacklists to warn users about potentially harmful websites.

2. DNS Filters: DNS filtering services, offered by companies like Cisco, Barracuda Networks, and Symantec, block access to malicious domains by redirecting users away from known phishing websites. These filters are effective in preventing users from reaching dangerous sites by blocking requests to malicious IPs or domains.

3. User Awareness Training: Educating users about phishing risks is a vital component of phishing prevention. Training programs teach individuals to recognize common phishing indicators.

4. Machine Learning Algorithms: Machine learning has emerged as a promising solution for phishing detection. Unlike static methods, machine learning algorithms can adapt to evolving threats by analyzing URL characteristics—such as domain names, length, and keyword presence—to identify patterns that suggest phishing attempts.

In this study, we investigate the use of machine learning for phishing URL detection, comparing the performance of Support Vector Machine (SVM) using two feature sets: the based 17 features from Banik et al. (2020)[1] and an extended set.

The remainder of this paper is organized as follows. Section 2 provides a review of related work in phishing URL detection. Section 3 describes the methodology employed in this study, including the feature sets and machine learning models. Section 4 presents the experimental results. Finally, Section 5 concludes the study.

2. Related Works

Phishing remains a critical issue in network and internet security, leading researchers to develop various methods to protect users from cyber-attacks. To prevent phishing, techniques such as blacklisting, whitelisting, machine learning, and deep learning have been employed to identify and block malicious URLs. This section reviews key studies that have used different approaches for detecting phishing sites.

Blacklisting and whitelisting are common methods for phishing URL detection. Blacklisting involves maintaining a database of known malicious URLs, as used by Google Safe Browsing API [11], which is regularly updated [4]. In contrast, whitelisting allows access only to trusted sites. Han et al. [5] introduced an Automated Individual White List (AIWL) using a Naïve Bayes classifier to alert users when interacting with unknown sites. Similarly, Jain and Gupta [7] developed a whitelist-based method that cross-references domain names and IPs, performing further checks on unlisted URLs using hyperlink features, and tested their approach on 1,525 URLs.

Blacklisting and whitelisting face limitations due to the short lifespan of phishing domains and the inability to handle zero-day attacks. To address this, Mohammad et al. [9] proposed an automated neural network using URL-based and external features (e.g., WHOIS, DNS, pop-ups). While this approach improves accuracy, it relies heavily on third-party services and is time-consuming to implement. Their method was tested on a small dataset of 1,400 URLs, highlighting the need for larger datasets to build more robust models.

A phishing URL classification system was proposed by Banik et al. [1], emphasizing lexical features of URLs. They extracted 17 features, which were prioritized using Chi-square feature selection.Four machine learning models—Random Forest, Decision Tree, Naive Bayes, and Support Vector Machine (SVM)—were tested, with the highest accuracy being achieved by Random Forest across three different datasets.

Another study[3] explores phishing URL detection through lexical structure analysis of URLs, utilizing classification models such as Extreme Gradient Boosting (XGBoost), SVM, and Artificial Neural Network (ANN) with datasets sourced from Phish Tank. This approach centers on static lexical features, achieving respective accuracies of 88%, 87%, and 88%, with XGBoost yielding the highest detection rate.

3. Proposed System

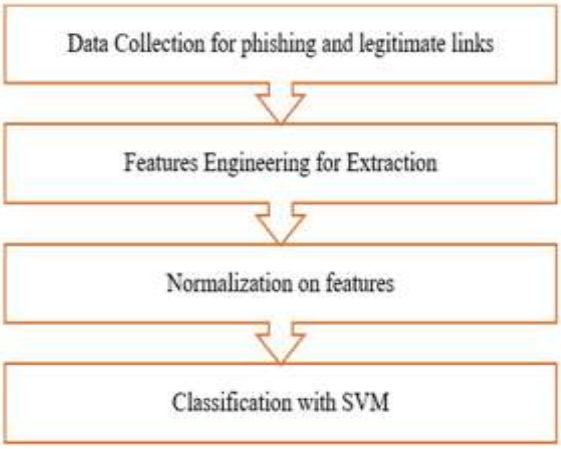
Data Collection for phishing and legitimate links
Features Engineering for Extraction
Normalization on features
Classification with SVM

Fig. 1 – Proposed System

We present a system for classifying phishing and legitimate URLs using lexical features and host based features of URLs. First, dataset of phishing and legitimate URLs are collected. Lexical features and host-based of these URLs are extracted. All values of the features extracted were then normalized by applying z-score the normalization method. Experiments have been performed using machine learning technique- SVM. Then, the performances of classification method are analyzed. Fig 1 shows the block diagram of our proposed system.

3.1 Dataset Description

The legitimate URLs Dataset and Phishing URLs Dataset utilized in this study were sourced from a dataset provided by the University of New Brunswick, which is publicly available at https://www.unb.ca/cic/datasets/. The legitimate dataset consists 35,378 of records and phishing dataset consists 10,465 of records.

3.2 Feature Extraction

Preprocessing is essential for preparing raw data for machine learning by improving model accuracy and reliability [8]. In this study, 10,465 phishing and 35,378 legitimate URLs were combined into a single dataset. Duplicate entries were removed to ensure data quality and representativeness. This process resulted in a clean and consistent dataset, forming a strong foundation for effective phishing URL detection.

Feature extraction is a crucial step in phishing URL detection, involving the selection of relevant characteristics from URLs. 17 lexical features are proposed in Banik et al. (2020) [1]. Following table 1 introduces an extended feature set designed to enhance detection accuracy. The selected features were derived and expanded based on insights from different researches [3][6][9].

In summary, feature extraction is a critical component of phishing URL detection. By combining these features, a comprehensive detection system can be built that accounts for the various techniques used by phishing attackers.

Table 1 – Extended Features Table

Feature Name	Description
Shortening Service	Detects if the URL uses a known shortening service (e.g., bit.ly, tinyurl.com). Phishing attackers often use shorteners to hide the final destination and trick users.
Top-Level Domain(TLD)	Extracts the TLD (e.g., .com,.xyz). Some TLDs are linked to phishing more often than others. Suspicious or uncommon TLDs may help identify phishing URLs.
Digit Count	Counts numeric digits in the URL. Excessive or strange digit use may indicate attempts to imitate legitimate sites or hide malicious intent.
Website Traffic	Evaluates if the website is ranked among Alexa's top 100,000. If not listed or ranked low, it may be a phishing site. Legitimate sites typically have consistent and high traffic; phishing domains are often new or unpopular.

3.3 Normalization

Normalization is a crucial preprocessing technique in machine learning, particularly when

algorithms are sensitive to the scale of input features, such as Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN). Without normalization, features with larger numeric ranges can disproportionately influence the learning process, resulting in biased outcomes and diminished model generalizability [10].

This study employed Z-Score Normalization to ensure all features contributed equally to the

classification model and to promote faster, more stable convergence during training. As Z-Score applies only to numerical features, categorical attributes were first converted using label encoding. Then, all features were standardized using the Z-Score formula:

$$z = \frac{x - \mu}{\sigma} \qquad (1)$$

where: x is the original value. μ is the mean of the feature. σ is the standard deviation of the feature.

Machine Learning Techniques

Machine learning provides a framework for developing predictive models based on patterns derived from data. This study adopts the supervised learning approach, wherein labeled data is used to train a model capable of classifying URLs as phishing or legitimate. Among various supervised algorithms, the Support Vector Machine is chosen for its robustness in high-dimensional spaces and its effectiveness in binary classification[1]. SVM constructs an optimal hyperplane that maximizes the margin between classes. The decision function is defined as:

$$f(x) = \text{sign}\ (w.\ x + b) \qquad (2)$$

However, phishing URL data often has complex, non-linear patterns that linear models struggle to separate. To handle this, the Radial Basis Function (RBF) kernel is used to map input data into a higher-dimensional space, allowing for non-linear decision boundaries[1]. The RBF kernel is defined as:

$$K\ (x\ ,\ x') = \exp(-\gamma \|x-x'\|^2) \qquad (3)$$

where γ determines the influence of individual training samples. This kernel allows SVM to effectively capture complex relationships in the data, improving classification accuracy in phishing URL detection.

4. Experimental Results

Performance evaluation is a critical phase in machine learning, intended to assess a model’s generalization capability and its effectiveness on training data. In this study, the dataset was partitioned into training and testing sets using an 80:20 split, with the 20% test set representing seen data. To provide a more detailed analysis of model performance, accuracy results for the testing phase using both feature sets are presented in Tables 2.

Table- 2 Experimental Results on different feature

FEATURE TYPE	ACCURACY	PRECISION	RECALL	F1-SCORE
BASED FEATURE (17)	98.61%	98.75%	97.30%	98.00%
EXTENDED FEATURE (21)	99.78%	99.64%	99.74%	99.69%

5. Conclusion

This study highlights the effectiveness of Support Vector Machine (SVM) in phishing URL classification, demonstrating the impact of an extended feature set on detection performance. The results indicate a substantial improvement in accuracy, underscoring the importance of comprehensive feature engineering in machine learning-based phishing detection. The incorporation of additional lexical and hosted based features enhances the model's ability to identify sophisticated phishing attempts, contributing to its reliability for real-world applications. These findings reinforce the critical role of feature extraction in optimizing classification performance and advancing phishing detection methodologies.

References

[1] B. Banik and A. Sarma, "Lexical feature-based feature selection and phishing URL classification using machine learning techniques," *Machine Learning, Image Processing, Network Security and Data Sciences*, pp. 93–105, 2020.

[2] B. Banik and A. Sarma, "Phishing URL detection system based on URL features using SVM," *International Journal of Electronics and Applied Research*, vol. 5, pp. 40–55, 2018.

[3] Cing Gel Vung and Yu Yu Win.(2023) "URL Classification Based on Lexical Features by Machine Learning (2023)". In 2023 IEEE Conference on Computer Applications (ICCA). IEEE.

[4] Gupta, B.B., Arachchilage, N.A.G., Psannis, K.E.: Defending against phishing attacks: taxonomy of methods, current issues and future directions. Telecommunication Systems 67 (2), 247–267 (2017).

[5] Han, W., Cao, Y., Bertino, E., Yong, J.: Using automated individual white-list to protect web digital identities. Expert Syst. Appl. 39, 11861–11869 (2012).

[6] Hutchinson, S., Zhang, Z., Liu, Q.: Detecting phishing websites with random forest. In: Meng, L., Zhang, Y. (eds.) MLICOM 2018. LNICSSITE, vol. 251, pp. 470–479. Springer, Cham (2018).

[7] Jain, A.K., Gupta, B.B.: A novel approach to protect against phishing attacks at client side using auto-updated white-list. EURASIP J. Inf. Secur. 2016(1), 1–11 (2016).

[8] Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). *Data preprocessing for supervised leaning*. International Journal of Computer Science, 1(2), 111–117.

[9] Mohammad, R.M., Thabtah, F., McCluskey, L.: Predicting phishing websites based on selfstructuring neural network. Neural Comput. Appl. 25(2), 443–458 (2013).

[10]Niño-Adan, I. Landa-Torres, and E. Portillo, "Influence of statistical feature normalisation methods on K-Nearest Neighbours and K-Means in the context of Industry 4.0," *Engineering Applications of Artificial Intelligence*, vol. 111, May 2022

[11] Overview Safe Browsing APIs (v4) Google Developers. https://developers.google.com/safe browsing/v4. Accessed 18 Dec 2019

[12] Sahoo, D., Liu, C., Hoi, S.C.H.: Malicious URL detection using Machine Learning: a survey. (2017)