Sri Lanka Institute of Information Technology

Malicious URL Detection using Machine Learning

Individual Assignment

ISP- (IE 3092)

IT20618872- Thisitha K.L.D

27/05/2023
Date of submission

Abstract—

Rogue websites or malicious universal resource locators (URLs) are one of the most common cybersecurity vulnerabilities. People lose billions of rupees each year by hosting unpaid content. (Spam, viruses, unethical marketing, spoofing, and so on) as well as duping unsuspecting visitors. People may be persuaded to visit these websites by email, advertisements, web searches, or links from other websites. Because of the increase in phishing, spamming, and malware incidents, a reliable system that can classify and identify harmful URLs is required. In each example, users click on the malicious URL. Classification became difficult due to the vast amount of data, changing trends and technologies, complicated linkages between attributes, a lack of training data, non-linearity, and the presence of outliers.For a variety of reasons, malicious URLs are identified in planned operations. The dataset is divided into four categories: malicious, benign, and defacement. For the proposed application, 6,51,191 URLs were used in total. To identify and categorize hazardous URLs, three machine learning algorithms were used: random forest, LightGBM, and XGBoost. URL Detection, Cybersecurity, Machine Learning, URL classification, Phishing, Benign, Defacement, and Malware are among the keywords in the index.

1. INTRODUCTION

During Covid-19, the development of internet firms such as e-commerce, social networking, and e-banking is heavily influenced. Unfortunately, sophisticated users' exploitation methods evolve in tandem with technology advancements. In these attacks, rogue websites are regularly utilized to obtain various forms of sensitive data that a hacker could use. Every URL has two unique characteristics: the identifier and the resource name. For example, the protocol identification for http.//upsssc.gov.in is http, and the resource name is upsssc.gov.in. Figure 1 depicts an example of a typical URL and identity.

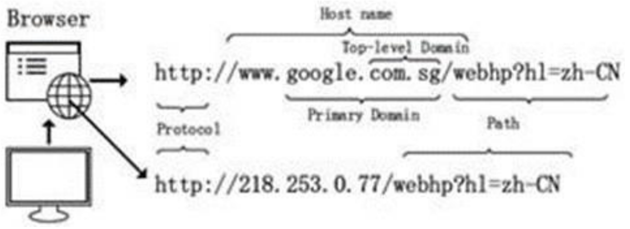

Browser
Host name
Top-level Domain
http://www.google.com.sg/webhp?hl=zh-CN
Protocol
Primary Domain
Path
http://218.253.0.77/webhp?hl=zh-CN

Fig.1. structure or format of URL

This research's proposed study considers bad URL identification and explores the assessment metrics of several Machine Learning classifiers. The data is from a public dataset on Kaggle.

This set has 450000 URLs. The best classifier is used to detect dangerous URLs from open, phishing websites. The rest of the paper is structured as follows. Section II delves into URL categorization. Section III covers the machine learning classification approaches used to address the problem.

Section IV includes dataset visualization. Section V discusses the results of the experiments. The conclusion is found in Section VI.

III. REVIEW OF LITRATURE

Rogue websites, often known as dangerous URLs, pose a significant security risk. Malicious URLs host undesirable information and dupe unsuspecting website visitors.

Every year, billions of dollars are lost as a result. Such hazards must be identified and addressed as soon as possible. Blacklists have frequently served as the primary means of detection. In contrast, blacklists are insufficient and cannot detect freshly established hazardous URLs.

Machine learning techniques are gaining popularity as a means to broaden the reach of harmful URL detectors. The dataset must be categorized, defined, and provide a formal statement of the machine learning goal of detecting harmful URLs using research on many aspects of this topic.

PROBLEM DESCRIPTION

URL has been extensively utilized and abused to exploit the user's vulnerability. This study focuses on determining if a URL is benign or malicious. It also analyzes the outcomes of other machine learning classification techniques, including Logistic Regression (LR), Stochastic Gradient Descent (SGD), Random Forest (RF), Support Vector Machine (SVM), Nave Bayes (NB), K-Nearest Neighbors (KNN), and Decision Tree (DT). To detect malicious URLs from the Open Phish website, the best performing classifier is employed. The suggested framework is divided into five stages:

- Data Collection: A tagged dataset of harmful and benign websites is acquired from the Kaggle repository.

- Data Cleaning and Extraction: Pre-processing includes further characteristic extraction, normalization, category value encoding, value standardization, and missing data management.

- Model Training: The Sklearn Python library is used to train the model on 80% of the data, employing various machine learning techniques such as Logistic Regression (LR), Stochastic Gradient Descent (SGD), Random Forest (RF), Support Vector Machine (SVM), Nave Bayes (NB), K-Nearest Neighbours (KNN), and Decision Tree (DT)

• Model Testing and Optimization: The trained model is validated against the remaining 20% of the data. Tuning hyperparameters improves accuracy, precision, and recall.

• Model Comparison: Evaluation metrics are used to compare machine learning classification approaches. URL has been widely used and misused to take advantage of the user's vulnerabilities. The purpose of this research is to determine whether a URL is benign or malicious.

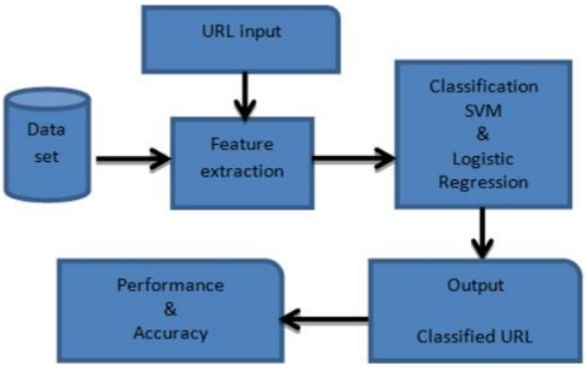
IV. PROPOSED MODELS

Data collection-A tagged dataset of hazardous, benign, defacement, and malware URLs is gathered from the Kaggle repository. There are no null or empty cells in the dataset under consideration. Data cleaning entails managing missing data, extracting new attributes, normalization, categorical value encoding, and value standardization. Model training- The Sklern Python library is used to train the model, which uses a range of machine learning methods such as random forest classifier, Light GBM classifier, and XGboost classifier.

Validation and model optimization were carried out using the remaining 20% of the data. Hyperparameter tuning improves sensitivity, F1 score, memory, and accuracy. Model comparison- Using assessment measures, the machine learning classification algorithms are compared.

V. methodology

. Traditionally, the blacklist method was used for detection. However, it was insufficiently thorough and lacked the ability to detect newly formed malicious URLs. New strategies have recently been investigated to improve the detection of rogue sites. Exploring the machine learning technique for comprehending and applying the attributes of current dangerous URLs for the detection of new malicious URLs is required for blacklisting bad sites. We usually get the following characteristics when we enter the URL: Features such as host-based, lexical, and popularity. We examine the reliability and genuineness of host sites in host-based because malicious ones are provided by less distinctive and verified hosts.In lexical one, we first distinguish two components of the URL: hostname and link path, and then tokens are searched for in the path and domain name because malicious sites contain a significant number of tokens. Next, we look at the URL length, as fraudulent URLs can have long and suspicious terms. The popularity of the sites is analyzed in popularity analysis since fraudulent URLs are less popular than genuine ones. URLs are classed as harmful (1), spoof (2), or benign (0) after the analysis. We will group spoof and malevolent together. The advantage of this strategy is that the user is not exposed to risky websites since we directly distinguish harmful URLs from benign URLs, hence boosting cybersecurity.



IV. CONCLUSION

To detect and classify harmful URLs, three machine learning algorithms are used. The random forest model yields the most accurate results. By training the random forest classifier with more balanced data, such as data containing almost equal proportions of harmful and beneficial websites. The experiment demonstrates that it is possible to detect fraudulent URLs by training a model with a database of defined features and then using the model to predict future attacks. Furthermore, the given models can be used in search engines or websites to inform visitors when they approach fake URLs.

VII. References

[1] Dhanalakshmi Ranganayakulu et al. “Detecting Malicious URLs in Email – An Implementation”, AASRI Procedia, vol. 4, pp. 125-131, 2013.

[2] Fuqiang Yu et al. “Malicious URL Detection Algorithm based on BM Pattern Matching”, International Journal of Security and Its Applications, vol. 9, pp. 33-44.

[3] K. Nirmal et al. “Phishing - the threat that still exists”, International Conference on Computing and Communications Technologies (ICCCT), pp. 139-143, 2015.

[4] F. Vanhoenshoven et al. “Detecting malicious URLs using machine learning techniques”, IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1-8, 2016.

[5] Doyen Sahoo et al. “Malicious URL Detection using Machine Learning: A Survey”, arXiv:1701.07179v3, 2019.

[6] Rakesh Verma et al. “What’s in a URL: Fast Feature Extraction and Malicious URL Detection”, Seventh ACM Conference on Data and Application Security and Privacy, pp.55-63,2017.

[7] Shantanu B et al. “Malicious URL Detection: A Comparative Study,” International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 1147-1151.

[8] A. Vikram et al. “Anomaly detection in Network Traffic Using Unsupervised Machine learning Approach,” 5th International Conference on Communication and Electronics Systems (ICCES), 2020, pp. 476-479.

[9] R. J. Franklin et al. “Anomaly Detection in Videos for Video Surveillance Applications using Neural Networks,” International Conference on Inventive Systems and Control (ICISC), 2020, pp. 632-637.

[10] Ritika H J et al. “Fraud Detection and Management for Telecommunication Systems using Artificial Intelligence (AI),” 3rd International Conference on Smart Electronics and Communication (ICOSEC), 2022, pp. 1016-1022.

[11] Niranjan DR et al. “Jenkins Pipelines: A Novel Approach to Machine Learning Operations (MLOps),” International Conference on Edge Computing and Applications (ICECAA), 2022, pp. 1292-1297.

V. AUTHOR PROFILE

Thisitha K.L.D
Information Technology BSc
(Hons)Specialization in CyberSecurity
Sri Lanka Institute of InformationTechnology
(SLIIT)