IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 7, July 2023

DOI: 10.17148/IARJSET.2023.107112

Machine Learning-Based Dangerous URL Identification

Chinmaya Shrinivasa Bhat Joshi[1], Suma N R[2]

Department of MCA, Bangalore Institute of Technology, Bengaluru[1]

Assistant professor, Department of MCA, Bangalore Institute of Technology, Bengaluru[2]

Abstract: Currently, the risk of network information is high. Insecurity is rising in both quantity and severity. Hackers' most popular tactics nowadays are to target end-to-end technology and exploit human weaknesses. Approaches include social engineering, phishing, and pharming. The use of malicious Uniform Resource Locators (URLs) to deceive users is one stage in carrying out these assaults. As a result, malicious URL detection is gaining popularity. Several scientific studies employing machine learning and deep learning approaches have been published, demonstrating a range of methods for identifying malicious URLs. In this paper, we present a machine learning-based malicious URL detection technique based on our hypothesized URL behaviors and attributes. In addition, Bigdata technology is being utilized to improve the identification of harmful URLs based on anomalous behavior. In summary, the proposed detection system consists of a novel set of URL attributes and behaviors, a machine learning algorithm, and bigdata technologies. The testing results indicate that the proposed URL characteristics and behavior can significantly improve the ability to detect malicious URLs. The proposed approach should be regarded as an optimized and user-friendly solution for dangerous URL detection.

Keywords: URL; malicious URL detection; feature extraction; feature selection; machine learning

I. INTRODUCTION

To refer to Internet resources, the Uniform Resource Locator (URL) is employed. In their paper, Sahoo et al. explored the URL's features and two essential components: protocol identifier, which defines which protocol to use, and resource name, which specifies the IP address or domain name where the resource is located. As can be seen, each URL has its own structure and format. Attackers usually attempt to change one or more URL structure components in order to mislead users and distribute their malicious URL. Malicious URLs are links that are harmful to users. These URLs will take visitors to resources or pages where attackers can execute code on users' computers, redirect users to undesired, dangerous, or malicious websites, or redirect users to malicious websites. or the installation of malware. Malicious URLs can also be disguised as secure download links and spread quickly. via file and message sharing across shared networks. Drive-by Download, Phishing and Social Engineering, and Spam are some of the attack methods that take use of malicious URLs. According to statistics released in, assaults employing the spreading malicious URL method rank first among the top ten most prevalent attack strategies in 2019. The three primary URL spreading tactics, which are dangerous URLs, botnet URLs, and phishing URLs, are rising in both the quantity of assaults and the danger level, according to this data. The statistics on the rise in the number of malicious URL deployments throughout the years indicates that there is a problem. The challenge of identifying rogue URLs has a solution.

There are two key developments at the moment: hazardous URL detection based on signals or sets of rules, and dangerous URL detection based on machine learning.The approach of identifying malicious URLs based on a set of markers or criteria may quickly and correctly discover hazardous URLs. However, this method is incapable of identifying new harmful URLs that do not meet any of the stated indicators or standards. To identify URLs based on their behaviors in the web, machine learning or deep learning techniques are utilized. In this research, machine learning methods are utilized to classify URLs based on their attributes. In addition, the magazine includes a new

II. LITERATURE SURVEY

A. Signature-Based Malicious URL Detection Long ago, investigations on malicious [1] URL detection using signature sets were investigated and implemented. The bulk of these studies often employ lists of known dangerous URLs. When a new URL is accessed, a database query is performed. IEEE International Journal of Machine Learning, Volume 11, Issue 1, July 5, 2022. If a URL is blacklisted, it is deemed dangerous and a warning is sent; otherwise, URLs are considered safe. The main downside of this strategy is that identifying new malicious URLs that are not on the given list

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 7, July 2023

DOI: 10.17148/IARJSET.2023.107112

will be incredibly tough. [2] Machine Learning for Malicious URL Detection To detect problematic URLs, three types of machine learning techniques may be used: supervised learning, unsupervised learning, and semi-supervised learning. In addition, the detection algorithms are based on URL behavior. [1] investigated a variety of harmful URL systems based on machine learning approaches. Examples include SVM, Logistic Regression, Nave Bayes, Decision Trees, Ensembles, Online Learning, and other machine learning approaches.

In this article, the two methods, RF and SVM, are applied. The experimental results will demonstrate the accuracy of these two methods with various parameter combinations. URL behaviors and attributes are classified as static or dynamic. In their research, the authors offered Lexical, Content, Host, and Popularity-based techniques for analyzing and extracting static behavior of URLs. As machine learning algorithms, Online Learning algorithms and SVM were used in these investigations. Use dynamic URL operations to identify malicious URLs. This study extracts URL attributes based on both static and dynamic behaviors. Among the attribute groups being studied are character and semantic groups. The abnormal website group and the host-based group Associated organization [3] Software for Detecting Malicious URLs URL omitted: URL Void is a URL verification application that makes use of a number of engines and domain blacklists. Google Safe Browsing, Norton Safe Web, and My WOT are some URL Void instances. The Void URL tool has the benefit of being cross-browser compatible and providing a wide range of extra testing services. The biggest issue with the Void URL tool is that the malicious URL detection technique relies heavily on a preset set of signatures. Among the attribute groups being studied are character and semantic groups. The abnormal website group and the host-based group Associated organization [3] Software for Detecting Malicious URLs URL omitted: URL Void is a URL verification application that makes use of a number of engines and domain blacklists. Google Safe Browsing, Norton Safe Web, and My WOT are some URL Void instances. The Void URL tool has the benefit of being cross-browser compatible and providing a wide range of extra testing services. The biggest issue with the Void URL tool is that the malicious URL detection technique relies heavily on a preset set of signatures. • an instrument for detecting holes. This allows people to check URLs or webmasters to set up daily checks by [5]. Primary attribute groups for malicious URL detection are extracted and selected. URL length, primary domain length, maximum token domain length, path average length, and domain average token length are lexical properties. Host-based characteristics: These variables are generated from the host properties of URLs. These features represent the location and identity of malicious servers, as well as the degree of impact of several host-based criteria that contribute to the URL's hazardous level.

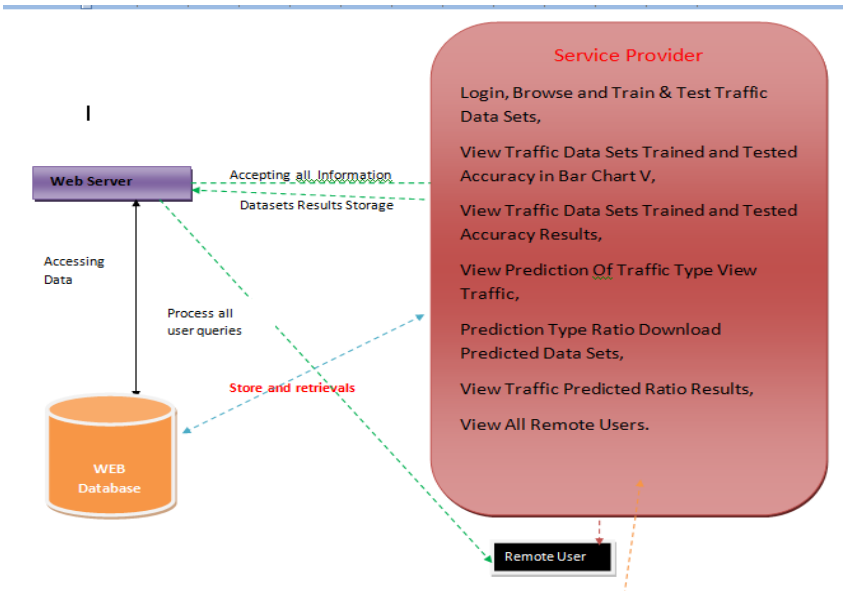

Fig [1] System Architecture

III. EXISTING SYSTEM

Signature-Based Malicious URL Detection Long ago, investigations on malicious URL detection using signature sets were investigated and implemented. The bulk of these studies often employ lists of known dangerous URLs. When a new URL is visited, a database query is executed. If a URL is blacklisted, it is deemed dangerous and a warning is sent; otherwise, URLs are considered safe. The main downside of this strategy is that identifying new malicious URLs that are

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 7, July 2023

DOI: 10.17148/IARJSET.2023.107112

not on the given list will be incredibly tough. To detect problematic URLs, three types of machine learning techniques may be used: supervised learning, unsupervised learning, and semi-supervised learning. And the detecting algorithms are based on URL behaviors. [1] investigated a variety of harmful URL systems based on machine learning approaches. Examples include SVM, Logistic Regression, Nave Bayes, Decision Trees, Ensembles, Online Learning, and other machine learning approaches. This study employs the two methods, RF and SVM. The experimental results will demonstrate the accuracy of these two methods under various parameter setups. There are two types of URL behaviors and qualities: static and dynamic. In their research, the authors presented ways for analyzing and extracting static behavior.

Based on both static and dynamic performance. Character and semantic groups, as well as the Abnormal group in websites and the Host-based group; Correlated group, are investigated.
Disadvantages: Machine Learning Algorithm Selection is missing from the system.
The system does not support URL Attribute Extraction and Selection.

IV. PROPOSED SYSTEM

In the suggested system, machine learning algorithms are used to classify URLs based on their features and behaviors. The novel qualities in the literature are retrieved from static and dynamic URL behaviors. The novel proposed characteristics are the research's main contribution. Machine learning methods are used in the malicious URL detection system. The supervised machine learning methods Support vector machine (SVM) and Random forest (RF) are used. Advantages The algorithms provided are well-suited to maximizing the value of our newly selected malicious URL detection characteristics. Although SVM and RF are not our major emphasis, they are provided as examples to highlight the general good performance of the detection system in the proposed research. Readers are welcome to implement other algorithms such as Nave Bayes, Decision trees, k-nearest neighbors, neural networks, and so on.

V. IMPLEMENTATION

MODULES:
Service Provider
View and Authorize Users
Remote User

[1] Service Provider
The Service Provider must login to this module using a valid user name and password. He can perform some actions after successfully logging in, such as Login, Search for URL Datasets and Train & Test Data Sets. View Urls Datasets Accuracy Trained and Tested in a Bar Chart, View Trained and Tested Accuracy Results for Urls Datasets, View Urls Type Prediction, View Urls Type Ratio, Download Predicted Data Sets, View Urls Type Ratio Results, View Every Remote User

[2] View and Authorize Users
The admin can view a list of all registered users in this module. The admin can examine the user's details such as user name, email, and address, and the admin can authorise the users.

[3]Remote User
There are a n number of users in this module. Before doing any operations, the user must first register. When a user registers, their information is saved in the database. After successfully registering, he must login using his authorised user name and password. Once logged in, the user can perform the following actions: REGISTER AND LOGIN, PREDICT URLS TYPE, and VIEW YOUR PROFILE.

VI. CONCLUSION

This study describes an artificial intelligence-based approach for identifying malicious URLs. The empirical data in Tables V and VI show that the proposed extracted attributes are effective. We do not employ unique attributes in our research, nor do we aim to build enormous datasets to improve the system's accuracy, as many other conventional publications do. In this scenario, a mix of easy-to-calculate qualities and big data processing methods determines the system's processing speed and accuracy. This study's findings can be used and incorporated in information security technologies and systems. The findings of this study have been utilized to develop a free online tool [20] for identifying fake URLs.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 7, July 2023

DOI: 10.17148/IARJSET.2023.107112

REFERENCES

[1] D. Sahoo, C. Liu, and S.C.H. Hoi, "Malicious URL Detection Using Machine Learning," A Survey of Learning". CoRR, abs/1701.07179, 2017.

[2] "Phishing detection: a literature review," M. Khonji, Y. Iraqi, and A. Jones. IEEE Communications Surveys & Tutorials, vol. 15, no. 4, 2013, pp. 2091-2121.

[3] M. Cova, C. Kruegel, and G. Vigna, "Detection and analysis of drivebydownload attacks and malicious javascript code," Proceedings of the 19th international World Wide Web Conference. ACM, 2010, pp. 281- 290.

[4] R. Heartfield and G. Loukas (2015), "A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks," ACM Computing Surveys (CSUR), vol. 48, no. 3, p. 37.

[5] Symantec's Internet Security Threat Report (ISTR) 2019. https://www.symantec.com/content/dam/symantec/docs/reports/istr-24-2019-en.pdf [Last updated 10/2019].

[6] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," Sixth Conference on Email and Anti-Spam (CEAS), 2009.

[7] Identifying malicious web pages using static heuristic and telecommunications networks and application conference in the year of [2008] in IEEE Conference by Seifert

[8] in the year of [2010] sinha proposed that shades of grey on the efficacy of reputation based blacklist In the IEEE Conference.