Journal of Science and Technology
ISSN: 2456-5660 Volume 8, Issue 07 (July -2023)
www.jst.org.in *DOI:https://doi.org/10.46243/jst.2023.v8.i07.pp39 - 44*

ENHANCED MALICIOUS URL DETECTION SYSTEM WITH MACHINE LEARNING ALGORITHMS

Mrs. Bessy [1], Archana Sharma [2], CH. Rasmitha [3], CH. Sindhu [4], D. Satvika [5]
[1] Associate Professor, Department of CSE, Malla Reddy Engineering College for Women, Hyderabad, Telangana, India.
[2,3,4,5] UG Scholar, Department of CSE, Malla Reddy Engineering College for Women, Hyderabad, Telangana, India.
bessybijo@gmail.com
sharchana2001@gmail.com, rasmithachinthalapally@gmail.com,
sindhuchirra23@gmail.com, sathwikadasari01@gmail.com.

To Cite this Article

Mrs. Bessy , Archana Sharma , CH. Rasmitha , CH. Sindhu , D. Satvika, " MALICIOUS URL DETECTION SYSTEM WITHMACHINE LEARNING ALGORITHMS" ***Journal of Science and Technology, Vol. 08, Issue 07,-July 2023, pp39-44***

Article Info
Received: 26-06-2023 Revised: 28-06-2023 Accepted: 10-07-2023 Published: 18-07-2023

ABSTRACT

Currently, the risk of network information insecurity is increasing rapidly in number and level of danger. The methods mostly used by hackers today is to attack end-to end technology and exploit human vulnerabilities. These techniques include social engineering, phishing, pharming, etc. One of the steps in conducting these attacks is to deceive users with malicious Uniform Resource Locators (URLs). As a results, Malicious URL detection is of great interest nowadays. There have been several scientific studies showing several methods to detect malicious URLs based on machine learning and deep learning techniques. In this paper, we propose a malicious URL detection method using machine learning techniques based on our proposed URL behaviors and attributes. Moreover, bigdata technology is also exploited to improve the capability of detection malicious URLs based on abnormal behaviors. In short, the proposed detection system consists of a new set of URLs features and behaviors, a machine learning algorithm, and a big data technology. The experimental results show that the proposed URL attributes and behavior can help improve the ability to detect malicious URL significantly. This is suggested that the proposed system may be considered as anoptimized and friendly used solution for malicious URL detection.

INTRODUCTION

Uniform Resource Locator (URL) is used to refer to resources on the Internet. In [1], Sahoo et al. presented about the characteristics and two basic components of the URL as: protocol identifier, which indicates what protocol to use, and resource name, which

specifies the IP address or the domain name where the resource is located. Each URL has a specific structure and format. Attackers often try to change one or more components of the URL's structure to deceive users for spreading their malicious URL. Malicious URLs are known as links that adversely affect users. These URLs will redirect users to resources or pages on which attackers can execute codes on users' computers, redirect users to unwanted sites, malicious website, or another phishing site, or malware download. Malicious URLs can also be hidden in download links that are deemed safe and can spread quickly through file and message sharing in shared networks. Some attack techniques that use malicious URLs include [2, 3, 4]: Drive-by Download, Phishing and Social Engineering, and Spam.

According to statistics presented in [5], in 2019, the attacks using spreading malicious URL technique are ranked first among the 10 most common attack techniques. Especially, according to this statistic, the three main URL spreading techniques, which are malicious URLs, botnet URLs, and phishing URLs, increase in number of attacks as well as danger level.

From the statistics of the increase in the number of malicious URL distributions over consecutive years, there is a need to study and apply techniques or methods to detect and prevent these malicious URLs.

Regarding the problem of detecting malicious URLs, there are two main trends at present as malicious URL detection based on signs or sets of rules, and malicious URL detection based on behavior analysis techniques [1, 2]. The method of detecting malicious URLs based on a set of markers or rules can quickly and accurately detect malicious URLs. However, this method is not capable of detecting new malicious URLs that are not in the set of predefined signs or rules. The method of detecting malicious URLs based on behavior analysis techniques adopt machine learning or deep learning algorithms to classify URLs based on their behaviors. In this paper, machine learning algorithms are utilized to classify URLs based on their attributes. The paper also includes a new URL attribute extraction method.

In our research, machine learning algorithms are used to classify URLs based on the features and behaviors of URLs. The features are extracted from static and dynamic behaviors of URLs and are new to the literature. Those newly proposed features are the main contribution of the research. Machine learning algorithms are a part of the whole malicious URL detection system. Two supervised machine learning algorithms are used, Support vector machine (SVM) and Random forest (RF).

EXISTING SYSTEM

A. Signature based Malicious URLDetection

Studies on malicious URL detection using the signature sets had been investigated and applied long time ago [6, 7, 8]. Most of these studies often use lists of known malicious URLs. Whenever a new URL is accessed, a database query is executed. If the URL is blacklisted, it is considered as malicious, and then, a warning will be generated; otherwise, URLs will be considered as safe. The main disadvantage of this approach is that it will be very difficult to detect new malicious URLs that are not in the given list.

B. Machine Learning based Malicious URL Detection

There are three types of machine learning algorithms that can be applied on malicious URL detection methods, including supervised learning, unsupervised learning, and semi supervised learning. And the detection methods are based on URL behaviors. In [1], several malicious URL systems based on machine learning algorithms have been investigated. Those machine learning algorithms include SVM, Logistic Regression, Nave Bayes, Decision Trees, Ensembles, Online Learning, etc. In this paper, the two algorithms, RF and SVM, are used. The accuracy of these two algorithms with different parameters setups will be presented in the experimental results.

The behaviors and characteristics of URLs can be divided into two main groups, static and dynamic. In their studies [9, 10, 11] authors presented methods of analyzing and extracting static behavior of URLs, including Lexical, Content, Host, and Popularity-based. The machine learning algorithms used in these studies are Online Learning algorithms and SVM. Malicious URL detection using dynamic actions of URLs is presented in [12, 13]. In this paper, URL attributes are extracted based on both static and dynamic behaviors. Some attribute groups are investigated, including Character and semantic groups;Abnormal group in websites and Host-based group; Correlated group.

Disadvantages

- The system is not implementedMachine Learning Algorithm Selection.
- The system is not implemented URL Attribute Extraction and Selection.

PROPOSED SYSTEM

- In the proposed system, machine learning algorithms are used to classify URLs based on the features and behaviors of URLs. The features are extracted from static and dynamicbehaviors of URLs and are new to the literature.
- Those newly proposed features are the main contribution of the research.

Machine learning algorithms are a part of

the whole malicious URL detection system. Two supervised machine learning algorithms are used, Support vector machine (SVM) and Random forest (RF).

Advantages

- The proposed algorithms are suitable to utilized the usefulness of our new features selected for malicious URL detection.
- In the proposed work, SVM and RF are selected as an example to illustrate the good performance of the whole detection system, and are not our focus. Readers are encouraged to implement some other algorithms such as Naïve Bayes, Decision trees, k-nearest neighbors, neural networks, etc.

IMPLEMENTATION

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as

Login, Browse URLs Datasets and Train & Test Data Sets, View URLs Datasets Trained and Tested Accuracy in Bar Chart, View URLs Datasets Trained and Tested Accuracy Results, View Prediction Of URLs Type, View URLs Type Ratio, Download Predicted Data Sets, View URLs Type Ratio Results, View All Remote Users

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address, and admin authorize the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN,PREDICT URLs TYPE,VIEWYOUR PROFILE.

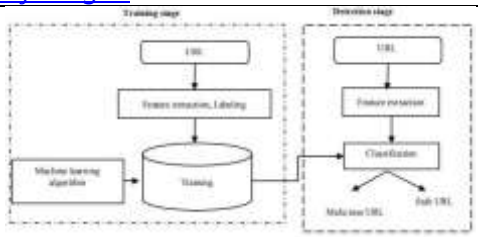

Fig.1. System architecture.

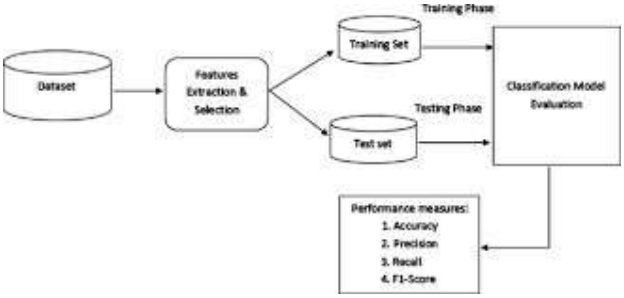

Fig.2. Output results.

CONCLUSION

In this paper, a method for malicious URL detection using machine learning is presented. The empirical results in Tables V and VI have shown the effectiveness of the proposed extracted attributes. In this study, we do not use special attributes, nor do we seek to create huge datasets to improve the accuracy of the system as many other traditional publications. Here, the combination between easy-to- calculate attributes and big data processing technologies to ensure the balance of the two factors is the processing time and accuracy of the system. The results of this research can be applied and implemented in information security technologies in information security systems. The results of this article have been used to build a free tool [20] to detect malicious URLs on web browsers.

REFERENCES

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

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

[3] M. Cova, C. Kruegel, and G. Vigna, "Detection and analysis of drivebydownload attacks and maliciousjavascript code," in Proceedings of the

19th international conference on Worldwide web. ACM, 2010, pp. 281– 290.

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

[5] Internet Security Threat Report (ISTR) 2019–Symantec.

https://www.symantec.com/content/dam

/symantec/docs/reports/istr-24- 2019- en.pdf [Last accessed 10/2019].

[6] S. Sheng, B. Wardman, G. Warner, L.

F. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in Proceedings of Sixth Conference on Email and Anti-Spam (CEAS), 2009.

[7] C. Seifert, I. Welch, and P. Komisarczuk, "Identification of malicious web pages with static heuristics," in Telecommunication Networks and Applications Conference, 2008. ATNAC 2008. Australasian. IEEE,2008, pp. 91–96.

[8] S. Sinha, M. Bailey, and F. Jahanian, "Shades of grey: On the effectiveness of reputation-based "blacklists"," in Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on. IEEE, 2008, pp. 57–64.

J. Ma, L. K. Saul, S. Savage, and G.
M. Voelker, "Identifying suspicious urls:
an application of large-scale online learning," in Proceedings of the 26th Annual International Conference onMachine Learning. ACM, 2009, pp. 681–
688.

B. Eshete, A. Villafiorita, and K. Weldemariam, "Binspect: Holistic analysis and detection of malicious web pages," in Security and Privacy in Communication Networks. Springer, 2013, pp. 149–166.

S. Purkait, "Phishing counter measures and their effectiveness– literature review," Information Management & Computer Security, vol. 20, no. 5, pp. 382–420, 2012.