ISSN: 2321-2152

IJMECE

International Journal of modern electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

ISSN2321-2152
www.ijmece .com
Vol 12, Issue 2, 2024

MALICIOUS URL DETECTION USING MACHINE LEARNING

[1]SMT.S ARUNA, [2]YANDRAPU MONIKA, [3]SASI KUMARI

[1](Assistant Professor), MCA, Swarnandhra College

[23]MCA, scholar, Swarnandhra College

ABSTRACT

There is a current, exponential growth in both the frequency and severity of threats to the security of network information. Attacking end-to-end technology and taking advantage of human weaknesses are the main tactics used by hackers nowadays. Examples of such methods include pharming, social engineering, phishing, etc. Misleading users with harmful URLs is a stage in carrying out these types of attacks. Because of this, detecting bad URLs is a hot topic right now. Using machine learning and deep learning approaches, several scientific research have shown various strategies for detecting dangerous URLs. Our suggested URL behaviors and properties form the basis of our malicious URL detection system that utilizes machine learning methods. This method is presented in this work. The capacity to identify harmful URLs based on aberrant behaviors is further enhanced by using bigdata technologies. To summarize, the suggested detection method is built on a big data technology, a machine learning algorithm, and a new set of attributes and behaviors for URLs. If implemented, the suggested URL properties and behavior may greatly enhance the capability to identify harmful URLs, according to the testing

ISSN2321-2152
www.ijmece .com
Vol 12, Issue 2, 2024

findings. It follows that the recommended approach might be seen as a user-friendly and efficient solution for detecting dangerous URLs.

1.INTRODUCTION

It is common practice to refer to online resources by their Uniform Resource Locator (URL). According to Sahoo et al. [1], there are two primary parts to a URL: the protocol identifier (which tells you which protocol to use) and the resource name (which gives you the IP address or domain name of the resource). There is a clear pattern to the structure and syntax of each URL. Attackers often attempt to alter the structure of URLs in order to trick people into sharing their malicious URLs. The term "malicious URL" describes links that cause harm to people. Attackers may insert malicious code onto users' systems using these URLs, or they can lead users to undesirable websites, harmful phishing sites, or malware downloads. Hidden malicious URLs in seemingly secure download links may also propagate rapidly via shared networks' file and message sharing capabilities. Spam, Drive-by Download, and Phishing and Social Designing are a couple of examples of assault strategies that utilization malevolent URLs [2, 3, 4].

The most popular attack tactic in 2019 was the distributing malicious URL approach, according to information given in [5]. In particular, this data shows that the three most common methods of URL propagation—malicious URLs, botnet URLs, and phishing URLs—increasing the frequency and severity of assaults.

The data showing a steady rise in the spread of harmful URLs over the last several years makes it quite evident that research and implementation of strategies to identify and stop these URLs are urgently required. At the moment, there are two major schools of thought when it comes to the issue of malicious URL detection: one that relies on signals or rules, and the other that uses behavior analysis approaches [1, 2]. Malicious URLs may be swiftly and correctly detected using the approach that relies on a set of markers or criteria. Nevertheless, this approach cannot identify newly dangerous URLs that do not belong to the collection of predetermined indicators or regulations. Machine learning and deep learning algorithms are used to categorize URLs according to their actions in the behavior

analysis approach for identifying malicious URLs. In this research, we use ML algorithms to categorize URLs according to their properties. The article also features an innovative approach to extracting URL attributes.

Our study use machine learning techniques to categorize URLs according to their characteristics and actions. This article introduces new features derived from URLs' static and dynamic characteristics. The key contribution of the study is the set of recently recommended attributes. The entire strategy for identifying hazardous URLs incorporates AI procedures. The two supervised machine learning techniques that are used are Random Forest (RF) and Support Vector Machine (SVM).

The following is the paper's structure. In the second section, we take a look at some of the more recent articles on malicious URL detection. Section III presents the suggested technique for detecting dangerous URLs using machine learning. Also detailed here are the latest additions to the URL detection method. Section IV presents the experimental data and discusses them. In Section V, the paper comes to a close.

2.LITERATURE SURVEY

A comprehensive literature survey on malicious URL detection using machine learning reveals a growing body of research dedicated to combatting cyber threats. Various studies have explored a range of techniques and methodologies aimed at effectively identifying malicious URLs amidst the vast expanse of the internet. Common approaches include feature extraction methods such as lexical analysis, content-based features, and behavior-based features. Machine learning algorithms such as Support Vector Machines (SVM), Random Forests, and Deep Learning models have been widely employed for classification tasks. Researchers have evaluated the performance of detection systems using metrics like accuracy, precision, recall, F1-score, and ROC-AUC. Despite significant advancements, challenges persist, including the need for large and diverse datasets, scalability concerns, and adaptability to emerging threats. Nonetheless, recent literature highlights promising developments, such as the integration of ensemble methods, deep learning architectures, and the utilization of novel features for enhanced detection accuracy. Addressing these challenges and building upon existing research will be crucial for the

continued advancement of malicious URL detection systems in the ever-evolving landscape of cybersecurity.

3. EXISTING SYSTEM

A. Malicious URL Detection Based on Signatures

There has been much research and implementation of signature sets for the purpose of malicious URL identification for quite some time [6, 7, 8]. In the majority of these investigations, lists of known harmful URLs are frequently utilized. A data set question is run at whatever point another URL is visited.In the absence of a blacklist, URLs will be deemed safe; in the presence of one, a warning will be sent. It will be exceptionally difficult to distinguish new hurtful URLs that are not on the given rundown; This technique's biggest flaw is this.

B. Detection of Malicious URLs using Machine Learning

In order to identify malicious URLs, one may use one of three machine learning algorithms: regulated learning, unaided learning, or semi-directed learning. The recognition calculations are worked around the ways of behaving of URLs. [1] examines a number of malicious URL systems that use machine learning techniques. Decision Trees, Ensembles, Support Vector Machines, Logistic Regression, Online Learning, etc. are instances of AI calculations. The RF and SVM calculations are utilized in this article. The exploratory outcomes will feature the exactness of these two calculations with differed boundary configurations. There are two primary categories into which URL behaviors and attributes fall: static and dynamic. Lexical, Content, Host, and Popularity-based approaches for assessing and extracting static behavior of URLs were given in research [9, 10, 11]. In these investigations, SVM and Online Learning algorithms were used as machine learning tools. In [12, 13], we see an example of malicious URL detection that makes use of the dynamic activities of URLs. Both static and dynamic behaviors are used to extract URL information in this article. Groups of attributes, such as character and semantic, are studied; Website abnormalities and host-based abnormalities; Connected set.

Disadvantages:

- The system is not implemented Machine Learning Algorithm Selection.

ISSN2321-2152
www.ijmece .com
Vol 12, Issue 2, 2024

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

3.1 PROPOSED SYSTEM

• To categorize URLs according to their characteristics and actions, the suggested method employs machine learning techniques. This article introduces new features derived from URLs' static and dynamic characteristics.

• The study mostly contributes to those newly suggested features. The whole method for detecting dangerous URLs includes machine learning techniques. The two supervised machine learning techniques that are used are Random Forest (RF) and Support Vector Machine (SVM).

Advantages:

To identify malicious URLs, the suggested algorithms are well-suited to make advantage of the newly-selected attributes.

Although they are not the primary emphasis of the proposed study, SVM and RF are chosen to demonstrate the overall detection system's strong performance. The use of additional algorithms like Naïve Bayes, Decision trees, k-nearest neighbors, neural networks, etc., is highly suggested for readers.

4. OUTPUT SCREENS

Home page:

Remote user login:

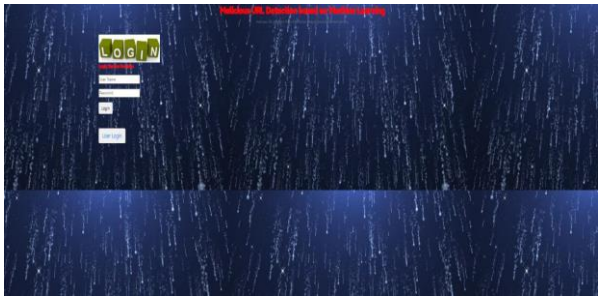
Registration:

Service providerlogin:

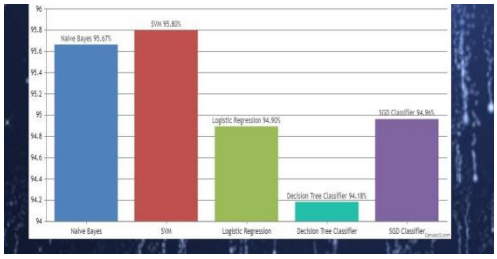
ISSN2321-2152

www.ijmece .com

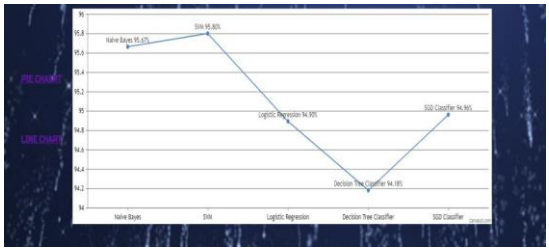
Vol 12, Issue 2, 2024



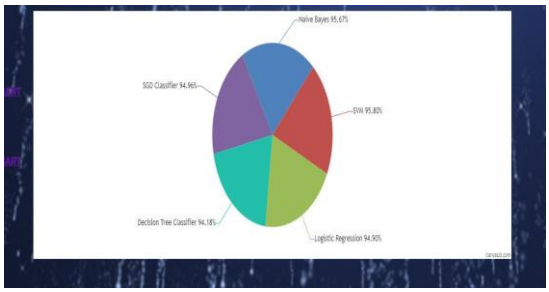
Bar chart:



Line chart:



Pie chart:



Accuracy:

5. CONCLUSION

This study suggests a smart method for efficiently identifying phishing emails. It compares and contrasts SVM, Random Forests, and Naive Bayes. When it comes to email phishing, finding the most intelligent classification model is the main objective. To assess how well each classifier performed, separate experiments were run on each of the three benchmarking testing levels. In the future, we want to evaluate SVM's efficacy using other benchmarking datasets. Also included is a comparison of SVM's performance with other kernels, including sigmoid and Gaussian kernels.

6.REFERENCES

• Sahoo, C. Liu, S.C.H. Hoi, “Malicious URL Detection using Machine Learning: A Survey”. CoRR, abs/1701.07179, 2017.

• M. Khonji, Y. Iraqi, and A. Jones, “Phishing detection: a literature survey,” IEEE

Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013.

• M.Cova, C.Kruegel, and G.Vigna, "Detection and analysis of drivebydownload attacks and malicious javascript code," in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 281–290.

• R.Heartfield and G.Loukas, "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, 2015.

• 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].

• 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.

• 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.

• 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 on Machine Learning. ACM, 2009, pp. 681–688.

• 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.