Malicious Url Detection Based on The Machine Learning

Reddymalla Praveen Reddy, Gurudu Rahul, Milkuri Mallikarjun, Puppala Shiva Kumar,
Mr.P.Subba Rao(Assistant professor),
Department of CSE,
MALLA REDDY INSTITUTE OF TECHNOLOGY AND SCIENCE, Telangana, hyderanad.

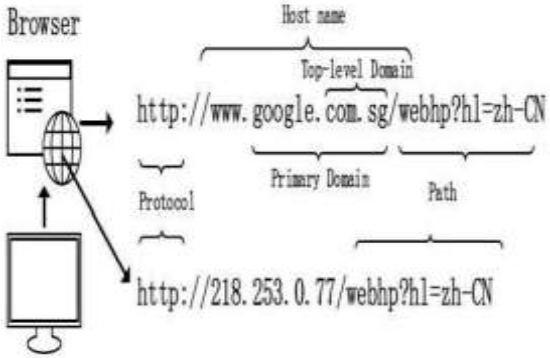
Abstract:

Both the frequency and severity of threats to network information security are on the rise at the moment. These days, hackers mostly target end-to-end systems and try to take advantage of people's weaknesses. Examples of such methods include pharming, social engineering, phishing, etc. Step one in launching such attacks is to trick people into visiting harmful URLs. Consequently, there is a lot of focus on malicious URL identification these days. Using machine learning and deep learning approaches, several scientific research have shown various strategies for detecting dangerous URLs. Based on the behaviors and features we describe for URLs, this research proposes a machine learning approach for malicious URL identification.

I. INTRODUCTION

The Internet uses the Uniform Resource Locator (URL) system to refer to online resources. Every unique resource locator (URL) follows a predetermined structure and format, as described in [1] by Sahoo et al., which consists of two parts: the protocol identifier (which specifies the protocol to use) and the resource name (which specifies the IP address or domain name where the resource is located). Attackers often attempt to alter the structure of URLs in order to trick users into sharing harmful URLs. Hyperlinks that cause harm to people are called malicious URLs. Users will be sent to resources or websites where hackers may install malware, reroute them to undesired sites, or install other phishing scams. Sharing files and messages on public networks also makes it easy for malicious URLs to hide in seemingly secure download links. Attack methods such as Drive-by Download, Phishing, Social Engineering, and Spam all make use of harmful URLs [2,3,4].

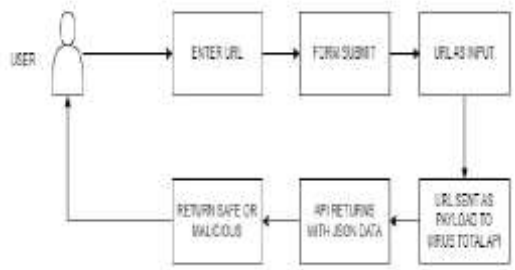
1.1 URL Structure and Format



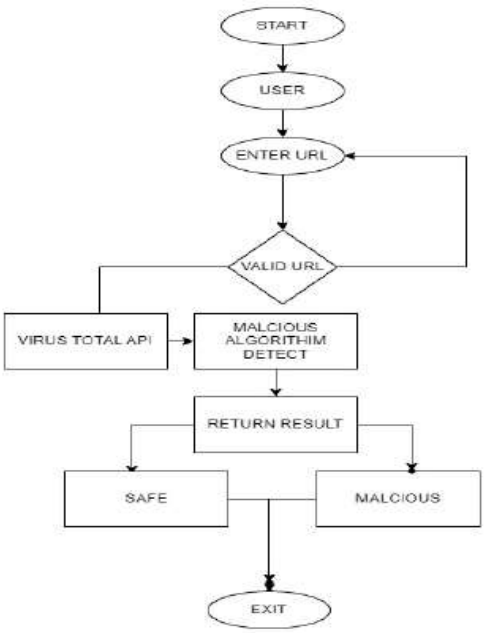
1.2 Flow Diagram for Detecting Malicious URLs

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. On the other hand, learning algorithms can categorize URLs according to their actions, but this approach can't identify new harmful URLs that don't fit the set of established indicators or criteria. This study uses machine learning methods to categorize URLs according to their properties. You may find a new approach for extracting URL attributes in the paper as well.

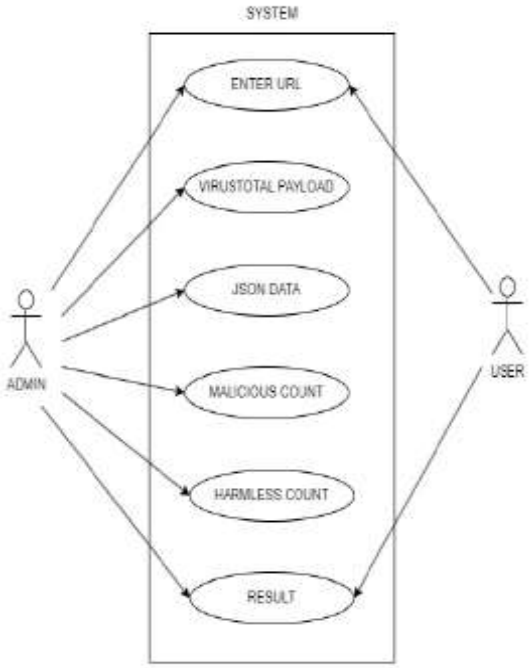
Schematic of Data Flow 1.



Data Flow Diagram -2



Data Flow Diagram -3 (User Case Diagram)



II. LITERATURE SURVEY

Title Name	Year	Publisher	Description
Phishing URL Detection: A Real-Case Scenario Through Login URLs	2022	IEEE	Phishing detection mechanism aims to improve current blacklist methods, protecting users from malicious login forms.
Malicious URL Detection using Logistic Regression	2022	IJRES	To check that if the website are malicious or not.

Section III: A SOLUTION FOR DETECTING Malicious URLS USING MACHINE LEARNING

You may make use of either static features or dynamic features. Without actually running the URL, static analysis examines a website using the data that is already accessible. Lexical characteristics from the URL string, host information, and sometimes even HTML and JavaScript content are among the aspects that are extracted. These solutions are more secure than the dynamic alternatives since they do not need execution. The basic premise is that benign and harmful URLs have differing distributions of these properties.

A. Identifying Malicious URLs using Signatures: -

For quite some time, researchers have studied and implemented methods for detecting malicious URLs using signature sets [6, 7, 8]. Research like this often makes use of databases that include

known dangerous URLs. A database query is done every time a new URL is requested. If the URL is not on a blacklist, it will not be marked as malicious; otherwise, a warning will be produced. The biggest problem with this method is that it won't be able to identify new harmful URLs that aren't already on the provided list.

Toolbox for Identifying Malicious URLs (B):-
URL Void: This application checks URLs with several engines and domain blacklists. Some instances of URL Void are MyWOT, Norton Secure Web, and Google Safe Browsing. One great thing about the Void URL tool is that it works with a wide range of browsers and can be used with a lot of other testing services. The most significant drawback of the Void URL tool is the over-reliance on a predefined collection of signatures for malicious URL identification.

IV. IMPLEMENTATION DETAILS

Training and detection are the two phases that make up the machine learning model for harmful URL detection.

Instructional phase:

It is important to gather both harmful and clean URLs in order to identify malicious ones. After that, we went on to attribute extraction after accurately labeling all the URLs, both clean and malicious. These characteristics will serve as the strongest foundation for distinguishing safe from dangerous URLs. In this work, we shall describe these characteristics in detail. Lastly, this dataset is partitioned into two parts: training data and testing data. The former is used to train machine learning algorithms, while the latter is used to test such algorithms. During the detection phase, the machine learning model will be used if its classification performance is satisfactory, indicating high classification accuracy.

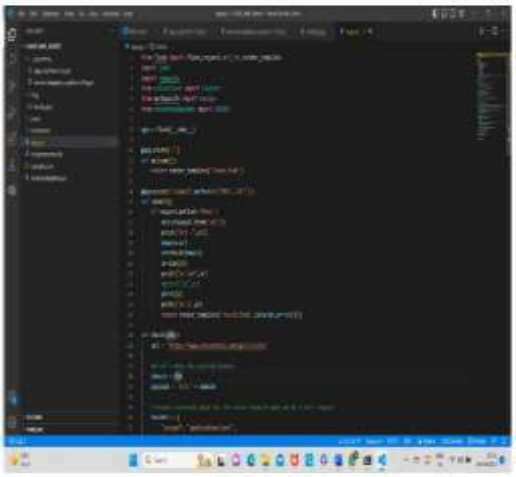
Phase of detection:

Each input URL undergoes the detecting step. Attribute extraction is the first step for the URL. After that, the classifier is fed these properties and told to determine whether the URL is safe or not. Longest possible URL, longest possible main domain, longest possible token domain, average path length, and average token length in domain are all lexical properties. Among these qualities are host-based ones, which are derived from the URLs' host attributes. Several host-based elements contribute to the URL's dangerous level; these qualities reflect the location and identity of malicious servers, as well as the degree of influence of these features.

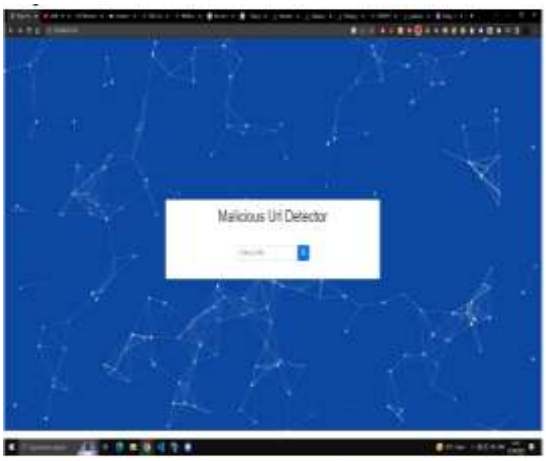
When you download a whole web page, you get these features: • Content-based Features. Due to the large amount of data that must be retrieved and the potential security risks associated with accessing that URL, these features place a high burden on the system. On the other hand, a more comprehensive understanding of a specific location should allow for the development of an improved prediction model. Most of a website's content-based functionality may be retrieved from its HTML text and PYTHON code.

V. OUTPUT

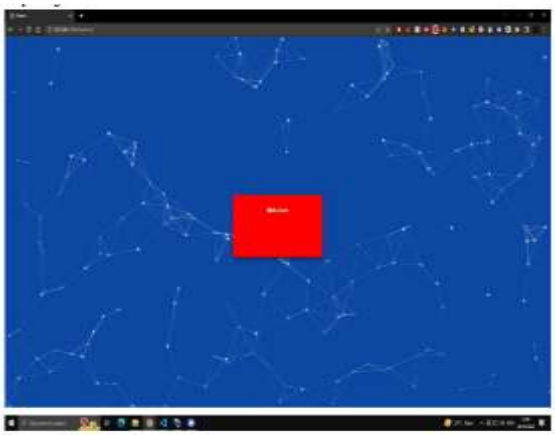
Malicious URL Detection Code



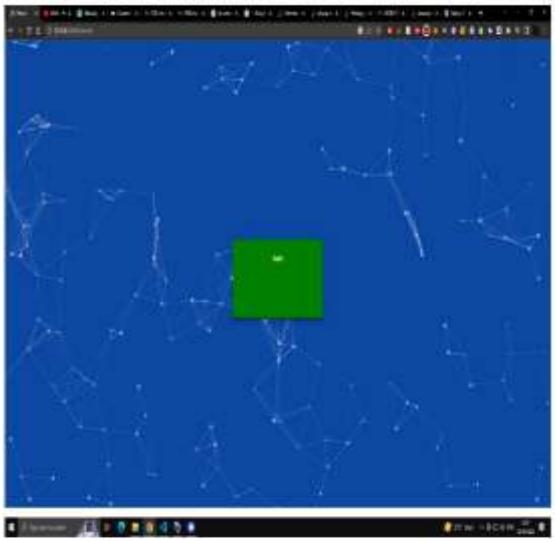
Home Page:-



Output Page:



OUTPUT SHOWING URL IS MALICIOUS



OUTPUT SHOWING URL IS HARMLESS

VI. DISCUSSION

Excellent scalability

With the ever-increasing volume of URLs, it is imperative that any system designed to identify dangerous URLs in the real world be able to train models with massive quantities of data, perhaps in the billions. Investigate more effective and scalable algorithms as a first step toward the high-scalability goal; next, construct scalable learning systems for use in distributed computing:

Powerful Adjustment
Problems that arise in practice include adversarial patterns like concept drifting, in which the distribution of malicious URLs changes over time or even in an adversarial way to evade detection, missing values, an ever-increasing number of new features, and so on. A comprehensive and effective malicious URL detection system in the real world has to be able to adapt to most situations.

Extreme Precision
When trying to identify malicious URLs, this is one of the primary objectives. We want to minimize the detection of misclassifying benign URLs as dangerous in order to increase the detection of all threats. Different degrees of detection thresholds are required to distinguish between benign and harmful ratios, since no system can achieve complete detection accuracy.

VII. CONCLUSION

Here, we displayed a comprehensive and well-structured research on the topic of harmful URL identification by machine learning. We analyzed the current system for malicious URL detection, specifically in terms of creating new feature representations and designing new learning algorithms for determining the malicious URL detection task, and we also provided an efficient design of malicious URL detection from a machine learning standpoint. As a service for real-world cyber security applications, we also outline the needs and obstacles of building malicious URL detection.

REFERENCES

[1]. Phishing URL Detection: A Real-Case Scenario Through Login URLs (April 18 IEEE) from https://ieeexplore.ieee.org/document/9759382
[2]. Malicious URL Detection using Logistic Regression (2022) from International Journal of Research in
Engineering and Science (IJRES)
[3]. Malicious URL Detection based on Machine Learning (2020) from (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 1, 2020.
[4]. Malicious URL Detection using Machine Learning: A Survey from (24 August 2019) School of Information
Systems, Singapore Management University.
[5]. Phishing Detection: A Literature Survey by Mahmoud Khonji, Youssef Iraqi, Senior Member, IEEE, and Andrew Jones from IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 15, NO. 4, FOURTH
QUARTER 2013