Machine Learning

Miss. Narjis Subuhi Syed[*], Dr. Ranjit keole[2]

Student, Department of Computer Science &Engineering, HVPM COET, Amravati, India[1]
Head of Department, Department of Information Technology & Engineering, HVPM COET, Amravati, India[2]

ABSTRACT

A malicious URL, often known as a malicious website, is a widespread and dangerous cyber security concern. Malicious URLs host unsolicited content (spam, phishing, drive-by downloads, and so on) and mislead unwary users into becoming scam victims (monetary loss, theft of private information, and malware installation), resulting in billions of dollars in damages each year. It is critical to detect and respond to such dangers quickly. Detection has traditionally been done mostly through the use of blacklists. Blacklists, on the other hand, aren't exhaustive and can't detect newly created harmful URLs. Machine learning approaches have been studied with growing interest in recent years to improve the generality of malicious URL detectors. This article attempts to provide a complete overview and structural understanding of machine learning-based malicious URL detection approaches. We offer a formal description of Malicious URL Detection as a machine learning task, as well as categorise and review the contributions of literature research that address various aspects of the topic (feature representation, algorithm design, etc.). This article also serves as a timely and comprehensive survey for a variety of audiences, including not only machine learning researchers and engineers in academia, but also professionals and practitioners in the cyber security industry, to help them understand the state of the art and facilitate their own research and practical applications. We also explore practical system design concerns, open research challenges, and crucial future initiatives.

.

I. INTRODUCTION

The value of the World Wide Web is growing all the time. Unfortunately, technological advancement has coincided with the development of highly sophisticated tactics for attacking and defrauding people. Rogue websites that sell counterfeit goods, financial fraud that tricks users into disclosing critical information that leads to money or identity theft, and even virus installation in the user's system are examples of such attacks. Explicit hacking efforts, drive-by downloads, social engineering, phishing, watering holes, man-in-the-middle, SQL injections, device loss/theft, denial of service, distributed denial of service, and many other techniques are used to carry out such assaults. It's difficult to create strong systems to identify cyber-security breaches because of the variety of attacks, possibly novel attack types, and the numerous contexts in which such attacks can arise. Given the exponential expansion of new security threats, quick changes in new IT technologies, and a substantial scarcity of security personnel, the limits of traditional security management tools are becoming increasingly critical. The majority of these attacks are carried out through circulating compromised URLs.

What is URL?
A Uniform Resource Locator (URL), colloquially termed a web address, is a reference to a web resource that specifies its location on a computer network and a mechanism for retrieving it. A URL is a specific type of Uniform Resource Identifier, although many people use the two terms interchangeably.

what is malicious URL?

Malicious URL is a URL created with malicious purposes, among them, to download any type of malware to the affected computer, which can be contained in spam or phishing messages, or even improve its position in search engines using Blackhat SEO techniques.Within the multitude of cyber threats out there, malicious websites play a critical role in today's attacks and scams. Malicious URLs can be delivered to users via email, text message, pop-ups or shady advertisements. The end result can often be downloaded malware, spyware, ransom ware, compromised accounts, and all the headaches those threats entail. It should be evident that being aware of what a Malicious URL is, and how it can do damage, is key to your email security.

Malicious URLs are a big part of most of the cyber security threats we see today. They are a tool that cyber criminals use to:

- Launch phishing campaigns meant to steal your personal information,
- Get you to install malware, viruses or Trojans, whether by downloading a file (without knowing it's malicious) or as a drive-by-download that is prompted by something as simple as a mouse-over or other trick,
- Launch a spam campaign against you that can involve phishing, malicious advertising, scams or other cyber-assisted fraud.

II. RELATED WORK

Many researchers have proposed different methods for classification and detection of malicious Web pages and detection of different Webpage attacks.

In [1], Naga et al have described how a machine can able to judge the URLs based upon the given feature set. Specifically, they described the feature sets and an approach for classifying the given the feature set for malicious URL detection as the traditional methods

To counter these limitations, Naga et al proposed a novel approach using sophisticated machine learning techniques that could be used as a common platform by the Internet users in order to detect the malicious URLs. Various feature sets for the URL detection have also been proposed that can be used with Support Vector Machines (SVM). The feature set used in [1] is composed of the 18 features, such as token count, average path token, largest path, largest token, etc. They also propose a generic framework that can be used at the network edge. That would safeguard the naive users of the network against the cyber-attacks.Although, using this method of detecting malicious URLs based on various features did not give much accuracy and obtaining features with a high collection time maybe infeasible.The comparison has been made on the various machine learning techniques. The detailed view of the results of various machine learning techniques has been elaborated in [2]. Machine Learning approaches use a set of URLs as training data, and based on the statistical properties, learn a prediction function to classify a URL as malicious. This gives them the ability to generalize to new URLs unlike blacklisting methods.

In [2], Doyen et al conducted a comprehensive and systematic survey on Malicious URL Detection using machine learning techniques. In this survey, they categorized most, if not all, the existing contributions for malicious URL detection in literature, and also identified the requirements and challenges for developing Malicious URL detection as a service for real-world cyber-security applications.

Patil et al performed an extensive literature survey of existing techniques and approaches for malicious Web pages detection. [3] Presents a brief overview of various forms of Web pages attacks. Patil et al introduced different Web pages and URLs features used for the effective detection of the malicious Web pages and also online learning algorithms as a promising approach for the large scale and efficient detection of malicious Web pages.

In [4], considering limitations of previous work for malicious URLs detection based on key features like URL features, URL source features, domain name features and short URLs features, Patil et al proposed a methodology to detect malicious URLs and identify attack types. 117 various types of discriminative features like URL features, domain name features, URL source features and short URLs features were used. Significant results were obtained by using proposed novel features.

Although, there was still need to investigate more features of short URLs for the effective detection and attack type identification, because it is the most growing trend today on the micro blogging sites like Twitter, Facebook etc. The first task is gathering data. Some websites offer malicious links while browsing. The next task is finding out clear URLs. We can use datasets which are already available so there is no need to crawl for non-malicious URLs. The feature extraction is used which extracts suspicious keywords, host features, URL features by using extraction techniques like Blacklist, Lexical, WHOIS, HTML, etc. The host features consists of Page Rank, Age of Domain, etc. and then score is calculated for the feature set using

III. PROBLEM STATEMENT

In existing there are various types of approaches all determination done on the various kind of processes in which some part of web pages consider originating root considers various type of data downloading protocols check but the actual checking will fails lot of times so that it is necessary to define a framework which will efficient in determination of malicious urls.

IV ANALYSIS

This chapter addresses the concept or issues related with the previous existing systems and their limitations over their implementations. It also focuses on how proposed system will overcome the drawbacks.

The use of World Wide Web is increasing continuously. Day by day the World Wide Web becomes a victim of Web attacks like spamming, phishing and malware. When the innocent user unknowingly visits the URL, it becomes the victim of the attacks. While visiting a torrent page, you click on a link, and then 2-3 browser windows will pop-up in the background. In other cases, you'll get pop-ups that ask you to download a new software or browser extension. These sites run on only two things: traffic and ad clicks. To maximize both, they will use shady software and ad networks in order to extract as many clicks as possible from you, the end user.The verification of URLs is very essential in order to ensure that user should be prevented from visiting malicious websites. For detecting these malicious URLs, various methods were proposed. However, these methods are sometimes time consuming and if not, they do not provide better accuracy. From what stated in the introduction, Blacklisting methods thus have severe limitations, and it appears almost trivial to bypass them, especially because blacklists are useless for making predictions on new URLs. Every time a new URL is introduced, it needs to be looked up in the existing database of malicious URLs. The new URL will be matched and tested with every previously known malicious URL in the black list. The update has to be made in black list whenever system comes across a new malicious URL. The technique is repetitive, time-consuming, and computationally intensive with ever increasing new URLs.

Heuristic Classification is an improvement to the blacklisting method. Here the signatures are matched and tested in order to find the correlation between the new URL and signature of existing malicious URL. Even though both Blacklisting and Heuristic Classification can effectively classify the malign and benign URLs, however, they cannot cope up with the evolving attack techniques. [1]Recent statistics imply that there is 20 - 25% growth in the attacks yearly and the threats that are coming from the newly created URLs are on the rise. One serious limitation of these techniques is that they are inefficient to classify the newly generated URLs.

proposed model.

4.1 Feasibility study

A feasibility study is an analysis that considers all of the relevant aspects for a project, including economic, technical, legal, and scheduling issues, in order to determine the chance of the project being completed successfully. Before investing a lot of time and money into a project, project managers conduct feasibility studies to determine the benefits and drawbacks of doing so.

4.2 Requirement Analysis

(Software tools, OS, Languages, Libraries etc.)

S/W Requirement

1. Python IDLE 3.0
2. MySQL

H/W Requirement

Following are the hardware requirements for the Project

- Processor : Intel Core i3(6th Generation)
- RAM : 4 GB or More
- Hard Disk : 500 GB and above

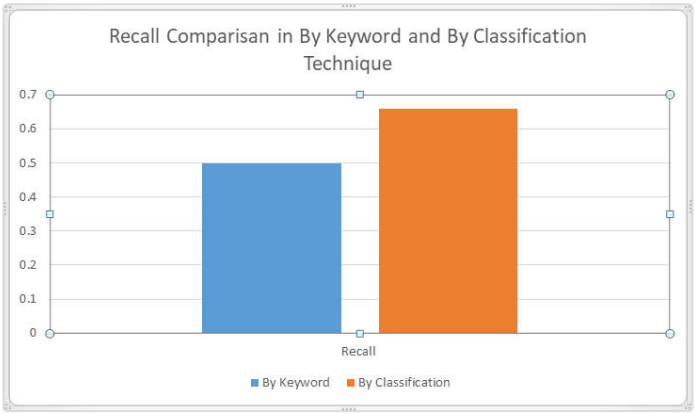
4.3 Performance Measures

Malicious ration Determination Using Keyword Comparison

URL/Category	Violent	Offensive	Sexual	Harmful	R(MRK) Ratio	Is Malicious
https://www.dictionary.com/browse/kill	3	1	0	0	1.0	Yes
https://www.merriam-webster.com/dictionary/hate	10	1	0	0	2.75	Yes
https://www.epa.gov/pesticide-worker-safety/paraquat-dichloride-one-sip-can-kill	0	0	2	3	1.25	No
https://www.google.co.in/url?/sex	4	0	2	0	1.5	No
https://en.wikipedia.org/wiki/Acid_attack	0	0	0	1	0.25	no
https://en.wikipedia.org/wiki/Murder_mystery_game	2	5	0	5	3.0	Yes

The table shows the word count and then the ratio of the searched urls.

Total Numbers of true positive T(p)=3



The above figure shows the recall comparison in By Keywords and By Classification Technique. .

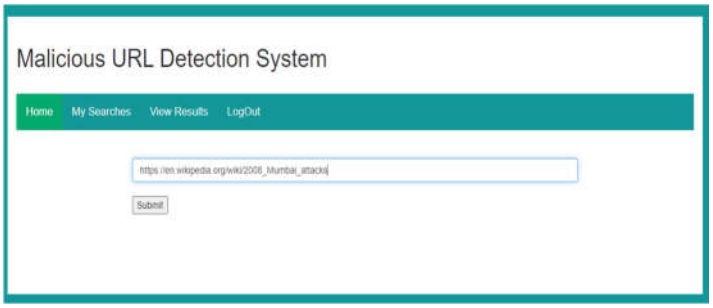
4.4 PRECISION The precision measurement is used to calculate the percentage of correctly recognised spam emails classified from a given collection of positive emails. This refers to the total number of emails accurately anticipated as positive out of the total number of emails forecasted as positive [35]. Equation identifies this.

Precision = TP / TP + FP

V. RESULTS

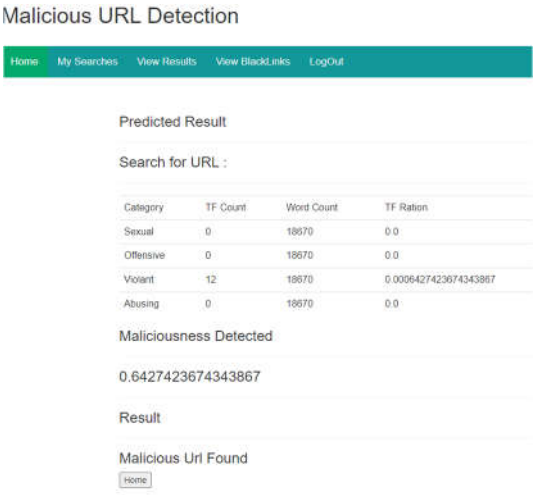
User Can Login in System or Register

User then can enter url in search box.

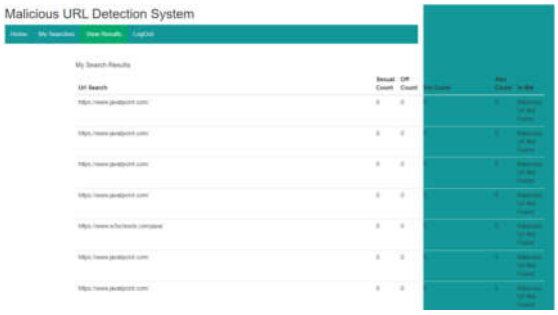


After web crawling and stop word removal system is ready to predict the result.

This is how it will show the result that malicious Url Found



In view result it will show all the searched urls history.



VI. CONCLUSION

In the proposed the web crawling based web page mining perform effectively which will find the malicious content more effectively as compared to other so that the proposed methodology will be highly effective as compared to previous. This system will

We came to a conclude that when we combine web crawling and sentimental analysis we get the better output and user can trust our system.

hope this paper will offer you exact knowledge of study and testing of malicious urls and means the accuracy of detecting urls.

VII. REFERENCES

[1] Immadisetti Naga Venkata Durga Naveen, Manamohana K, Rohit Verma, "Detection of Malicious URLs using Machine Learning Techniques" in proceedings of the International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-4S2 March, 2019

[2] Doyen Sahoo, Chenghao Liu, and Steven C.H.Hoi. "Malicious URL Detection using Machine Learning: A Survey. 1,1(August2019),37pages.

[3] Dharmaraj R. Patil and Jayantrao B. Patil, "Survey on Malicious Web Pages Detection Techniques" in proceedings of the International Journal of u- and e-Service, Science and Technology, August 2015 05, Volume No. 8, Number 05 (pp.195-206) http://dx.doi.org/10.14257/ijunesst.2015.8.5.18

[4] Dharmaraj R. Patil and Jayantrao B. Patil, "Feature-based Malicious URL and Attack Type Detection Using Multi-class Classification" in proceedings of The ISC Int'l Journal of Information Security, July 2018, Volume 10, Number 2 (pp. 141–162)

[5] Thomas G. Dietterich. Ensemble Methods in Machine Learning. International Workshop on Multiple Classifier Systems, pp 1-15, Cagliari, Italy, 2000.

[6] 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.

[7] Internet Security Threat Report (ISTR) 2019–Symantec. https://www.symantec.com/content/dam/symantec/docs/reports/istr-242019-en.pdf [Last accessed 10/2019].

[8] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, ―An empirical analysis of

[9] S. Sinha, M. Bailey, and F. Jahanian, ―Shades of grey: Refresh the trade, and check start date and end date

[10] MALWARE 2008. 3rd International Conference on. IEEE, 2008, pp. 57–64

[11] 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

[12] 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.

[13]]G. Canfora, E. Medvet, F. Mercaldo, and C. A. system calls sequences,‖ in Availability, Reliability, and Visaggio, ―Detection of malicious web pages using Security in Information Systems. Springer, 2014.

[14] 1) S. Purkait, ―Phishing counter measures and 382–420, 2012.

[15] Y. Tao, ―Suspicious url and device detection by log mining,‖ Ph.D. dissertation,Applied Sciences: School of Computing Science, 2014.