JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DETECTING HARMFUL URL'S USING MACHINE LEARNING

K. Srinivasa Rao[1], A. Sai Teja[2], A. Siva Sankar Reddy[3], G. Sujith[4], J. Abhiram Reddy[5]

[1]Associate Professor, Department of Computer Science and Engineering, Vidya Jyothi Institute of Technology, Aziz Nagar, Hyderabad, Telangana, India

[2,3,4,5]Student, Department of Computer Science and Engineering, Vidya Jyothi Institute of Technology, Aziz Nagar, Hyderabad, Telangana, India

ABSTRACT - Currently, the number and severity of network information insecurity threats are rapidly increasing. Hackers today primarily target end-to-end technology and exploit human vulnerabilities. These Techniques include social engineering, phishing, and pharming, among others. One step in carrying out these attacks is to deceive users with malicious Uniform Resource Locators (URLs). As a result, harmful URL detection is becoming increasingly important. Several scientific studies have demonstrated a variety of methods for detecting malicious URLs using machine learning and deep learning techniques. Based on our proposed URL behaviours and attributes, we propose Detection of harmful URL's using machine learning techniques in this paper. Furthermore, bigdata technology is used to improve the detection of malicious URLs based on abnormal behaviour. To summarise, the proposed detection system consists of a novel set of URL features and behaviours, a machine learning algorithm, and bigdata technology. The results of the experiments show that the proposed URL attributes and behaviour can significantly improve the ability to detect malicious URLs. It is suggested that the proposed system be regarded as an optimised and user-friendly solution for detecting malicious URLs.

Key Words: Malicious, Harmful URL, Web Application, Phishing Attacks

I.INTRODUCTION

The term Uniform Resource Locator (URL) refers to Internet resources. Sahoo et al. presented about The URL's characteristics and two basic components are as follows: protocol identifier, which indicates 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 frequently attempt to change one or more components of the URL structure in order to deceive users into spreading their malicious URL. Malicious URLs are links that harm users. These URLs will redirect users to resources or pages where attackers can execute codes on users' computers, redirect users to unwanted sites, malicious websites, or other phishing sites, or malware download. Malicious URLs can also be hidden in seemingly safe download links and spread quickly via file and message sharing in shared networks. Drive-by Download, Phishing and Social Engineering, and Spam are some attack methods that use malicious URLs.

According to our research, attacks using the spreading malicious URL technique rank first among the ten most common attack techniques in 2019. Especially, According to our research, there are three main URL spreading techniques, which are malicious URLs, botnet URLs, and phishing URLs, which increase the number of attacks as well as the danger level.

Based on our research, the number of malicious URL distributions has increased over the years, indicating that there is a need to study and apply techniques or methods to detect and prevent these malicious URLs. There are now two primary tendencies when it comes to the challenge of detecting malicious URLs: malicious URL detection based on signals or sets of rules, and malicious URL detection based on behaviour analysis techniques. Malicious URLs can be identified fast and precisely via a method based on a collection of markers or rules. However, this strategy is incapable of detecting new dangerous URLs that do not match the specified indications or guidelines. The method for identifying malicious URLs is based on behaviour analysis approaches and uses machine learning algorithms to categorise URLs based on their actions. In this study, URLs are categorised according to their properties using machine learning techniques. Machine learning algorithms are used in our study to classify URLs based on their features and behaviours.

The features are novel in the literature and are extracted from static and dynamic URL behaviours. The primary contribution of the research is the newly proposed features. Machine learning algorithms are a component of the malicious URL detection system as a whole. We have used four machine learning algorithms in this project, Passive aggressive algorithm, Multinomial Naïve Bayes, Support Vector and Adaboost Classifier.

II.LITERATURE SURVEY

There are several related works available that have been already published. In this section, we will analyse the related survey approaches for the problems and their solutions and extend them to make the application. Below are some literature reviews:

1. **L. McCluskey, F. Thabtah, and R. M. Mohammad**, "Intelligent rule-based phishing website classification," IET Inf. The phishing is represented as art because it is a developing of website which is desire to fetch users information like pin number, password, username many more . The Phishing website contains a variety of cues throughout its content sections, but this is due to the browser-based security indicator provided in conjunction with the website. There are numerous solutions planned to combat phishing. Nonetheless, there is no specific cure or solution that can eliminate the threats. All of the presented techniques for predicting phishing attacks are predicted in data processing, significantly inducing categorification. Since the anti-phishing solution is aimed at predicting website accuracy, the results of the data-mining classification technique are specifically matched. In this paper, the author stores the lightweight in the vital options that differentiate phishing websites from other malicious websites and yet sensible rule based data processing classification technique which are in the prediction of phishing website and which of the classification techniques are developed to be a lot more authentic.

2. "Estimating malicious urls using a self-structuring neural network," **R. M. Mohammad, F. Thabtah**, and **L. McCluskey**, b 2014, Neural Compute Appl., vol. 25, no. 2, pp. 443-458. Every day, the Internet becomes a more important part of our lives. Even if continuous monitoring is carried out, information theft occurs. So, using a neural network, we attempted to detect the phishing website. So we tried machine learning, trained the machine with a few details, and attempted to identify the phishing website based on already existing features. A network structure is created in order to accurately predict phishing websites.

3. This paper discusses the concept of website phishing detection developed by **N. Sanglerdsinlapachai** and **A. Rungsawang**. This is where new methods like cantina are developed, as well as extra features like header comparison. This experiment is completed by working on two hundred data sets, one phishing website and one non-phishing website, and analysing the error rate of 7.50% and the f mesa of 0.9250.

4. **G. Canbek**, "A Review on Information, Information Security, and Security Processes," The phishing is said as art of creating or developing web link like trust worthy organisation and stealing the personal information and details of the victims. This has many cues in it that are browsed based security indicators. The answers are being developed for the phishing website, but even after many attempts, finding the solution is becoming difficult. As a result, data processing is a prominent method for detecting the phishing website developed in this paper research. In this study, they focused on lightweight characteristics that help distinguish between phishing and non-phishing websites. As a result, a rule-based classification method was developed to detect the attackers website.

5. **Y. Cao, W. Han, and Y. Le,** "Anti-phishing based on automated individual white-list," Corporations that provide online trading and will help to serve people all over the world. Even though marketing is taking place, there are numerous issues such as insecure online transactions and others. Phishing is considered an art form because it is the development of a website with the intent of obtaining user information such as a pin number, password, username, and so on. So, based on the url, ip address in the url, and some prefix and suffix details added to the url, a few features explain whether the website is phishing or not. So we examine one feature that will extract itself from the website and determine whether it is legitimate or not. By utilising this feature, you are attempting to protect the users' or clients' details from attackers.

6. **P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "**Phishnet: predictive blacklisting to detect phishing attacks," develop multi-labeled data from a single data set using an associative classification rule, which is considered a challenging task for the presented algorithm. The present associative classification algorithm will develop the large frequency classes that are attached to the rules that are present in the training data-sets and will remove all remaining classes even after all of these classes have data representations with the body's rules. The problem with which we are dealing in this paper is the associative classification algorithm known as Enhanced Multi labelled Classification based Associative Classification (eMCAC). The algorithms provided will discover

the rules associated with some set of classes from single label datas, as were the other current algorithms. The associative classification algorithm is incapable of inducing.Furthermore, the eMCAC will reduce the number of rules extracted using classified building methods. The proposed algorithms are tested using an already existing world application data set that is already related to phishing websites, and the results show that the eMCACs accuracy is very competitive when compared to all other associative classical algorithms that are contrasted and present in data-mining. Finally, the experimental results will show that our algorithm will be able to generate new rules from phishing data sets that the end user can use to make a decision.

III.Existing Methodology

The current system for malicious URL detection employs a number of methods, including blacklisting, whitelisting and heuristic analysis.

Blacklisting: Blacklisting entails keeping a list of known malicious URLs that are blocked from access by users. This method is effective at blocking known threats but has a limited scope because it cannot detect new or unknown threats.

Whitelisting: Whitelisting is the process of keeping a list of trusted URLs that are allowed access. This method is effective in mitigating the risk of threats from unknown sources, but it can be restrictive for users.

Heuristic: The process of examining a URL for suspicious patterns or characteristics that indicate malicious intent. This method employs algorithms to analyse the URL and identify potential threats.

IV.Proposed Methodology

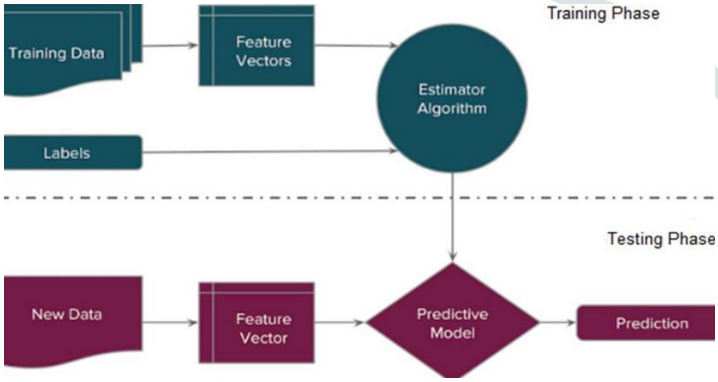

Fig 1:Architecture of Proposed System

The above architecture describes about the proposed system.

In the proposed methodology we introduced a web application where user can check if the URL is malicious or legitimate

By this verification we can prevent the user from the trap which is laid by the hackers. here with the help of machine learning algorithms and dataset

we predict the output that whether it is a malicious URL or a safe/legitimate URL, we implemented four algorithms they are passive aggressive classifier, multinomial NB, support vector and ABA Booster. By comparing their accuracy and the results while using the algorithms we selected a suitable algorithm i.e (passive aggressive classifier).

Here in this web application when we submit the URL that we want to check the application shows the results (i.e malicious or legitimate) with a confusion matrix and accuracy.

V.Result Screen

Fig 1: Detection Page

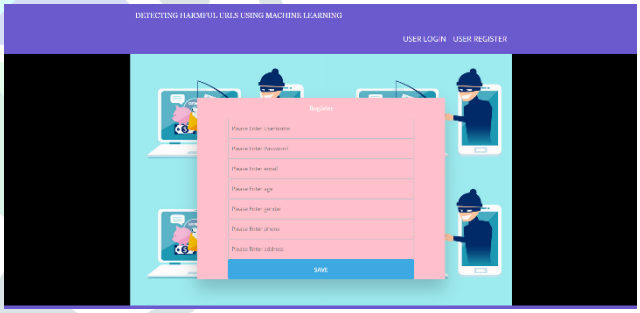

Fig 2: Registration Page

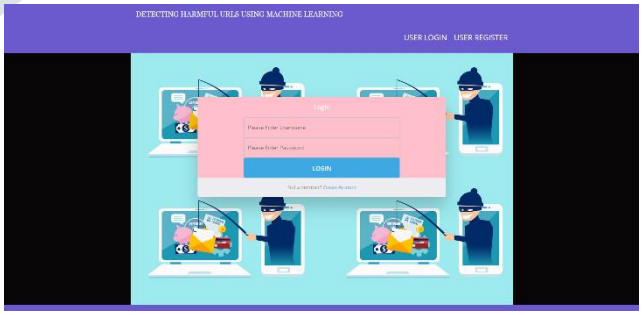

Fig 3: login Page

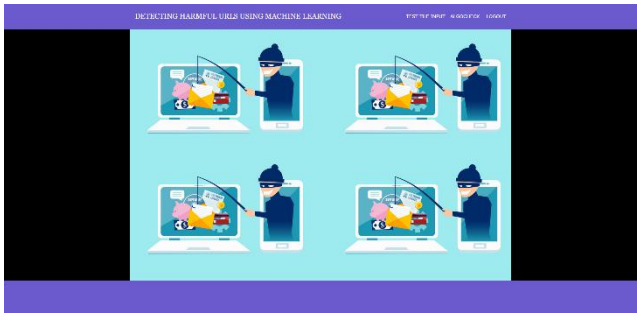

Fig 4: Home Page

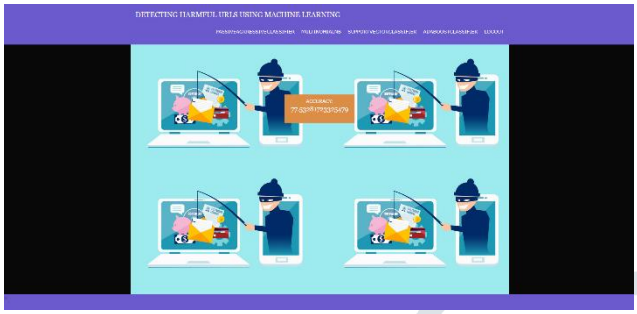

Fig 5: Passive Aggressive Accuracy

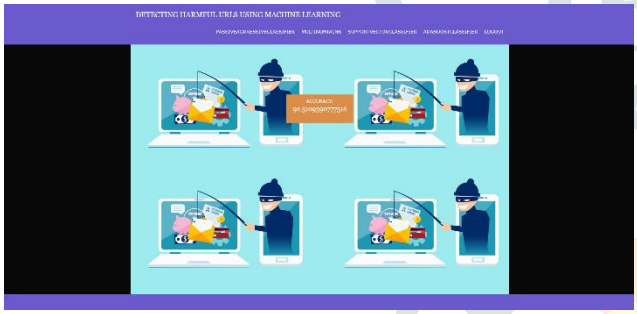

Fig 6: MultinominalNB Accuracy

Fig 7: Supportvector Accuracy

Fig 8: ADA Boost Accuracy

Fig 9: Input URL Page

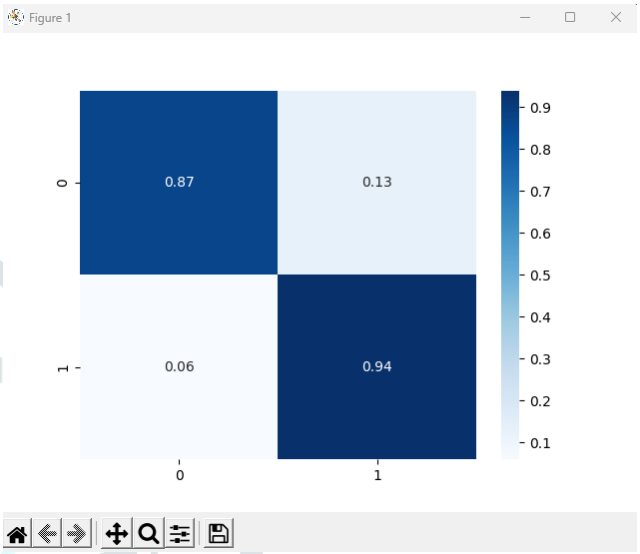

Fig 10: Confusion matrix of Result page

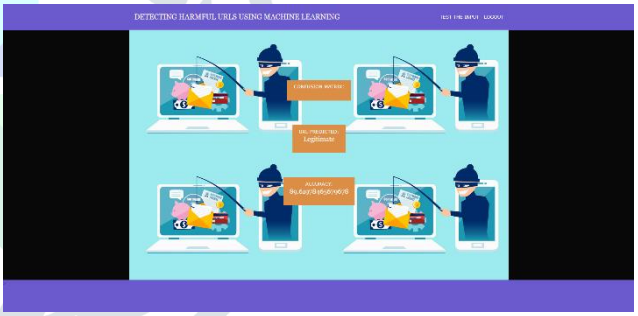

Fig 11: Result of the url we entered

VI.Conclusion

In this paper we conclude that this project uses machine learning technology to detect harmful URLs by extracting and analysing various features of legitimate and malicious URLs. To detect phishing websites, the Passive Aggressive Algorithm, Multinomial Nave Bayes, Support Vector, and Adaboost Classifier are used. The goal of this paper is to detect phishing URLs and narrow down the best machine learning algorithm by comparing each algorithm's accuracy rate, false positive and false negative rate. In experiments, the detection model produces the expected result. However, because network traffic in the test environment and the real network differ, and as the Internet evolves, so do the types of malicious URL. The model must be updated on a regular basis in the actual scenario. As a result, in the future, we intend to investigate how to simplify

the detection model's architecture and reduce training time while maintaining detection performance.

VII.References

1. **J. Lee, Y. Lee, D. Lee, H. Kwon and D. Shin**, "Classification of Attack Types and Analysis of Attack Methods for Profiling Phishing Mail Attack Groups", IEEE Access, vol. 9, pp. 80866-80872, 2021.

2. **J. Li and S. Wang, "PhishBox**: An Approach for Phishing Validation and Detection", 2017 IEEE 15th Intl Conf on Dependable Autonomic and Secure Computing 15th Intl Conf on Pervasive Intelligence and Computing 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), pp. 557-564, 2017.

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

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

5. **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.

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. **A. Oest, Y. Safaei, A. Doupé, G. Ahn, B. Wardman and K. Tyers**, "PhishFarm: A Scalable Framework for Measuring the Effectiveness of Evasion Techniques against Browser Phishing

8. **Xiao Han, Nizar Kheir and Davide. Balzarotti, "PhishEye**: Live Monitcoring of Sandboxed Phishing Kits", pp. 1402-1413, 2016.

9. **T. Nathezhtha, D. Sangeetha and V. Vaidehi**, "WC-PAD: Web Crawling based Phishing Attack Detection", 2019 International Carnahan Conference on Security Technology (ICCST), pp. 1-6, 2019.

10. **Liu Wenyin, Guanglin Huang, Liu Xiaoyue, Xiaotie Deng and Zhang Min**, "Phishing Web page detection", Eighth International Conference on Document Analysis and Recognition (ICDAR'05), vol. 2, pp. 560-564, 2005.

11. **Leo Breiman.:** Random Forests. Machine Learning 45 (1), pp. 5- 32, (2001).

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

13. **DeveloperInformatio**n. https://www.phishtank.com/developer_info.php. .

14. **URLhausDatabaseDump**. https://urlhaus.abuse.ch/downloads/csv/.

15. **DatasetURL**. http://downloads.majestic.com/majestic_million.csv.

16. **Malicious_n_Non-MaliciousURL**. https://www.kaggle.com/antonyj453/urldataset#data.csv.