IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 6, June 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306152|

Detection of Phishing Website using Machine Learning

Selvi P, Devika K

Assistant Professor, Department of CSE, Jayam College of Engineering and Technology, Dharmapuri, Tamil Nadu, India

Student, Department of Master of Computer Application, Jayam College of Engineering and Technology, Dharmapuri, Tamil Nadu, India

ABSTRACT: Malicious Web sites largely promote the growth of Internet criminal activities and constrain the development of Web services. As a result, there has been strong motivation to develop systemic solution to stopping the user from visiting such Web sites. We propose a learning based approach to classifying Web sites into 3 classes: Benign, Spam and Malicious. Our mechanism only analyzes the Uniform Resource Locator (URL) itself without accessing the content of Web sites. Thus, it eliminates the run-time latency and the possibility of exposing users to the browser based vulnerabilities. By employing learning algorithms, our scheme achieves better performance on generality and coverage compared with blacklisting service.

URLs of the websites are separated into 3 classes:

- Benign: Safe websites with normal services
- Spam: Website performs the act of attempting to flood the user with advertising or sites such as fake surveys and online dating etc.
- Malware: Website created by attackers to disrupt computer operation, gather sensitive information, or gain access to private computer systems.

I. INTRODUCTION

While the Internet has brought unprecedented convenience to many people for managing their finances and investments, it also provides opportunities for conducting fraud on a massive scale with little cost to the fraudsters. Fraudsters can manipulate users instead of hardware/software systems, where barriers to technological compromise have increased significantly. Phishing is one of the most widely practised Internet frauds. It focuses on the theft of sensitive personal information such as passwords and credit card details. Phishing attacks take two forms:

- Attempts to deceive victims to cause them to reveal their secrets by pretending to be trustworthy entities with a real need for such information
- Attempts to obtain secrets by planting malware onto victims' machines.

The specific malware used in phishing attacks is subject of research by the virus and malware community and is not addressed in this thesis. Phishing attacks that proceed by deceiving users are the research focus of this thesis and the term 'phishing attack' will be used to refer to this type of attack

1.1 Background:

Phishing websites, fraudulent sites that impersonate trusted third party to gain access to private data, continue to cost Internet users over a billion dollars each year. In this paper, we describe the design and performance characteristics of a scalable machine learning classifier we developed to detect phishing websites. We use this classifier to maintain Google's phishing blacklist automatically. Our classifier analyses millions of pages a day, examining their and the contents of a page to determine whether or not a page is phishing. Unlike previous work in this field, we train the classifier on a noisy dataset consisting of millions of samples from previously collected live classification data. Despite the noise in the training data, our classifier learns a robust model for identifying phishing pages which correctly classifies more than 90% of phishing pages several weeks after training concludes.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306152|

Phishing is a fraudulent technique that uses social and technological tricks to steal customer identification and financial credentials. Social media systems use spoofed e-mails from legitimate companies and agencies to enable users to use fake websites to divulge financial details like usernames and passwords. Hackers install malicious software on computers to steal credentials, often using systems to intercept username and passwords of consumers' online accounts. Phishers use multiple methods, including email, Uniform Resource Locators (URL), instant messages, forum postings, telephone calls, and text messages to steal user information. The structure of phishing content is similar to the original content and trick users to access the content in order to obtain their sensitive data. The primary objective of phishing is to gain certain personal information for financial gain or use of identity theft. Phishing attacks are causing severe Economic damage around the world. Moreover, most phishing attacks target financial/payment institutions and webmail, according to the Anti-Phishing Working Group (APWG) latest Phishing pattern studies.

In order to receive confidential data, criminals develop unauthorized replicas of a real website and email, typically from a financial institution or other organization dealing with financial data. This e-mail is rendered using a legitimate company's logos and slogans. The design and structure of HTML allow copying of images or an entire website. Also, it is one of the factors for the rapid growth of Internet as a communication medium, and enables the misuse of brands, trademarks and other company identifiers that customers rely on as authentication mechanisms. To trap users, Phisher sends "spooled" mails to as many people as possible. When these e-mails are opened, the customers tend to be diverted from the legitimate entity to a spoofed website.

There is a significant chance of exploitation of user information. For these reasons, phishing in modern society is highly urgent, challenging, and overly critical. There have been several recent studies against phishing based on the characteristics of a domain, such as website URLs, website content, incorporating both the website URLs and content, the source code of the website and the screenshot of the website. However, there is a lack of useful anti-phishing tools to detect malicious URL in an organization to protect its users. In the event of malicious code being implanted on the website, hackers may steal user information and install malware, which poses a serious risk to cyber security and user privacy. Malicious URLs on the Internet can be easily identified by analyzing it through Machine Learning (ML) technique. The conventional URL detection approach is based on a blacklist (set of malicious URLs) obtained by user reports or manual opinions. On the one hand, the blacklist is used to verify an URL and on the other hand the URL in the blacklist is updated, frequently. However, the numbers of malicious URLs not on the blacklist are increasing significantly. For instance, cybercriminals can use a Domain Generation Algorithm (DGA) to circumvent the blacklist by creating new malicious URLs. Thus, an exhaustive blacklist of malicious URLs is almost impossible to identify the malicious URLs. Thus new malicious URLs cannot be identified with the existing approaches. Researchers suggested methods based on the learning of computer to identify malicious URLs to resolve the limitations of the system based on the blacklist. Malicious URL detection is considered a binary classification task with two-class predictions: malicious and benign. The training of the ML method consists of finding the best mapping between the d-dimensional vector space and the output variable. This strategy has a strong generalization capacity to find unknown malicious URLs compared to the blacklist approach.

II. LITERATURE SURVEY

1. Detecting phishing websites using machine learning technique

Author-Ashit Kumar Dutta

Year-2021

In recent years, advancements in Internet and cloud technologies have led to a significant increase in electronic trading in which consumers make online purchases and transactions. This growth leads to unauthorized access to users' sensitive information and damages the resources of an enterprise. Phishing is one of the familiar attacks that trick users to access malicious content and gain their information. In terms of website interface and uniform resource locator (URL), most phishing webpages look identical to the actual webpages. Various strategies for detecting phishing websites, such as blacklist, heuristic, Etc., have been suggested. However, due to inefficient security technologies, there is an exponential increase in the number of victims. The anonymous and uncontrollable framework of the Internet is more vulnerable to phishing attacks. Existing research works show that the performance of the phishing detection system is limited. There is a demand for an intelligent technique to protect users from the cyber-attacks. In this study, the author proposed a URL detection technique based on machine learning approaches. A recurrent neural network method is employed to detect phishing URL. Researcher evaluated the proposed method with 7900 malicious and 5800

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306152|

legitimate sites, respectively. The experiments' outcome shows that the proposed method's performance is better than the recent approaches in malicious URL detection.

2. Phishing Website Detection Using Machine Learning
Author- AdarshMandadi; SaikiranBoppana; Vishnu Ravella; R Kavitha
Year-2022
Phishing is an internet scam in which an attacker sends out fake messages that look to come from a trusted source. A URL or file will be included in the mail, which when clicked will steal personal information or infect a computer with a virus. Traditionally, phishing attempts were carried out through wide-scale spam campaigns that targeted broad groups of people indiscriminately. The goal was to get as many people to click on a link or open an infected file as possible. There are various approaches to detect this type of attack. One of the approaches is machine learning. The URL's received by the user will be given input to the machine learning model then the algorithm will process the input and display the output whether it is phishing or legitimate. There are various ML algorithms like SVM, Neural Networks, Random Forest, Decision Tree, XG boost etc. that can be used to classify these URLs. The proposed approach deals with the Random Forest, Decision Tree classifiers. The proposed approach effectively classified the Phishing and Legitimate URLs with an accuracy of 87.0% and 82.4% for Random Forest and decision tree classifiers respectively.

3. Phishing Website Detection using Machine Learning Algorithms
Author- Rishikesh Mahajan, IrfanSiddavatam
Year-2018
Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

4. Detecting Phishing Websites Using Machine Learning
Author- SafaAlrefaai; GhinaÖzdemir; Afnan Mohamed
Year-2022
Phishing, a cybercriminal's attempted attack, is a social web-engineering attack in which valuable data or personal information might be stolen from either email addresses or websites. There are many methods available to detect phishing, but new ones are being introduced in an attempt to increase detection accuracy and decrease phishing websites 'success to steal information. Phishing is generally detected using Machine Learning methods with different kinds of algorithms. In this study, our aim is to use Machine Learning to detect phishing websites.

III. EXISTING SYSTEM

Phishing causes billions of dollars in damage every year and poses a serious threat to the Internet economy. Email is still the most commonly used medium to launch phishing attacks. In this paper, we present a comprehensive natural language based scheme to detect phishing emails using features that are invariant and fundamentally characterize phishing. Our scheme utilizes all the *information* present in an email, namely, the header, the links and the text in the body. Although it is obvious that a phishing email is designed to elicit an action from the intended victim, none of the existing detection schemes use this fact to identify phishing emails. Our detection protocol is designed specifically to distinguish between "actionable" and "informational" emails. To this end, we incorporate natural language techniques in phishing detection.

DISADVANTAGES

- We also utilize contextual information, when available, to detect phishing: we study the problem of phishing detection within the contextual confines of the user's email box and demonstrate that context plays an important role in detection.
- To the best of our knowledge, this is the first scheme that utilizes natural language techniques and contextual information to detect phishing.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306152|

- We show that our scheme out performs existing phishing detection schemes. Finally, our protocol detects phishing at the email level rather than detecting masqueraded websites.
- This is crucial to prevent the victim from clicking any harmful links in the email. Our implementation called Phish Net-NLP, operates between a user's mail transfer agent (MTA) and mail user agent (MUA) and process search arriving email for phishing attacks even before reaching the inbox.

IV. PROPOSED SYSTEM

This article surveys the literature on the detection of phishing attacks. Phishing attacks target vulnerabilities that exist in systems due to the human factor. Many cyber-attacks are spread via mechanisms that exploit weaknesses found in end users, which makes users the weakest element in the security chain. The phishing problem is broad and no single silver-bullet solution exists to mitigate all the vulnerabilities effectively, thus multiple techniques are often implemented to mitigate specific attacks. This paper aims at surveying many of the recently proposed phishing mitigation techniques.

ADVANTAGES

1. All of URLs in the dataset are labelled.
2. We used two supervised learning algorithms random forest and support vector machine to train using scikit-learn library.

DATASET COLLECTION

To train our deep learning architecture, we collected images. The architecture of the learning technique highly depends on CNN. Data from source is collected for training and testing the model. Dataset contains images of faces only. It consists of about 1,315 images in which 658 images containing people with face masks and 657 images containing people without face masks. For training purposes, 80% images of each class are used and the rest of the images are utilized for testing purposes. Fig. 2 shows some of the images of two different classes.

DEEP LEARNING

Deep learning is a subset of machine learning that is particularly well-suited to tasks such as image classification, object detection, and face recognition. Deep learning involves the use of deep neural networks, which are composed of many layers of interconnected nodes that process and transform data. In the context of face mask detection, deep learning can be used to train a neural network on a dataset of images that contain people wearing masks and people not wearing masks. The trained neural network can then be used to classify new images as containing faces wearing masks or faces not wearing masks.

PRE-PROCESSING

Masks play a significant role in protecting the health of individuals against virus spread in air, as it is one of the few precautions available for COVID-19 in the absence of immunization. Hence, it is very important for us to detect whether an individual wears a mask or not and ensure they avail public services only under the former condition. The images with face mask and without masks, used as datasets in the proposed system are availed from Kaggle and various other sources. We can obtain high accuracy depending on the number of images collected. The dataset for training contains a lot of noise and duplicates. The accuracy of the model depends on the dataset chosen for training. The dataset hence has to be pre-processed before being fed as input.

V. CONCLUSION

Finally, phishing attacks are a major problem. It is important that they are countered. The work reported in this thesis indicates how understanding of the nature of phishing may be increased and provides a method to identify phishing problems in systems. It also contains a prototype of a system that catches those phishing attacks that evaded other defences, i.e. those attacks that have "slipped through the net". An original contribution has been made in this important field, and the work reported here has the potential to make the internet world a safer place for a significant number of people.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306152|

VI. FUTURE WORK

In the future we provide some technical solution by improve the efficiency of spam filters. By which too many mails are classified correctly and properly. By this legitimate user can surf internet with less fear. The user-phishing interaction model was derived from application of cognitive walkthroughs. A large-scale controlled user study and follow on interviews could be carried out to provide a more rigorous conclusion. The current model does not describe irrational decision making nor address influence by other external factors such as emotion, pressure, and other human factors. It would be very useful to expand the model to accommodate these factors. We have theoretically and experimentally evaluated of Phish Limiter. We have evaluated the trustworthiness of each SDN flow to identify any potential hazards based on each deep packet inspection.

REFERENCES

[1] Almutairi, W., Saeed.F, Al-Sarem.M, Al-Shamani.M," Hybrid Filter and Wrapper Feature Selection Method for Enhancing Detection Process of Phishing Websites" https://doi.org/10.1007/978-981-16-5559-3_34 (Springer 2022).
[2] CagatayCatal, GörkemGiray, BedirTekinerdogan, Sandeep Kumar &Suyash Shukla "Applications of deep learning for phishing detection: a systematic literature review" https://doi.org/10.1007/s10115-022-01672 (Springer 2022).
[3] Frank Cremer, Barry Sheehan, Michael Fortmann, Arash N. Kia, Martin Mullins, Finbarr Murphy & Stefan Materne "Cyber risk and cybersecurity: a systematic review of data availability" https://doi.org/10.1057/s41288-022-00266- 6 (Springer 2022).
[4] GregaVrbancic, IztokFister Jr., ViliPodgorelec " Datasets for phishing websites detection" Volume 33, December 2020, 106438 https://doi.org/10.1016/j.dib.2020.106438 (Elsevier 2020).
[5] HuseyinAhmetoglu ,Resul Das "A comprehensive review on detection of cyber-attacks: Data sets, methods, challenges, and future research directions" https://doi.org/10.1016/j.iot.2022.100615 (Elsevier 2022).
[6] I. Kara, M. Ok and A. Ozaday, "Characteristics of Understanding URLs and Domain Names Features: The Detection of Phishing Websites With Machine Learning Methods," vol. 10, pp. 124420-124428 doi: 10.1109/ACCESS.2022.3223111 (IEEE Access 2022)
[7] ImenSouiden , Mohamed NazihOmri , ZakiBrahmi "A survey of outlier detection in high dimensional data streams" https://doi.org/10.1016/j.cosrev.2022.100463 (Elsevier 2022).
[8] Jain, A.K., Debnath, N. & Jain, A.K. APuML: "An Efficient Approach to Detect Mobile Phishing Webpages using Machine Learning" https://doi.org/10.1007/s11277-022-09707 (Springer 2022).
[9] K. Sushma, M. Jayalakshmi and T. Guha "Deep Learning for Phishing Website Detection" doi:10.1109/MysuruCon55714.2022.9972621 (IEEE Explore 2022).
[10]KibreabAdane&BerhanuBeyene "Phishing Website Detection with and Without Proper Feature Selection Techniques: Machine Learning Approach" https://doi.org/10.1007/978-3-031-24475-9_61 (Springer 2023).
[11]M. Aljabri and S. Mirza, "Phishing Attacks Detection using Machine Learning and Deep Learning Models" Riyadh, Saudi Arabia, 2022, pp. 175-180, doi: 10.1109/CDMA54072.2022.00034 (IEEE Explore 2022).
[12]Manuel Sánchez-Paniagua, Eduardo Fidalgo, Enrique Alegre, RocíoAlaiz-Rodríguez" Phishing websites detection using a novel multipurpose dataset and web technologies features" https://doi.org/10.1016/j.eswa.2022.118010 (Elsevier 2022).

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details