IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

EFFICIENT EMAIL PHISHING USING MACHINE LEARNING

S. SREE VIDHYA[1], M. NAVEENKUMAR[2],
1 ASSISTANT PROFESSOR, 2 PG SCHOLAR
COMPUTER SCIENCE AND ENGINEERING
ERODE SENGUNTHAR ENGINEERING COLLEGE (AUTONOMOUS), ERODE, TAMILNADU, INDIA

ABSTRACT

Emails are frequently utilized as a way of personal and professional communication. Banking information, credit reports, login data, and other sensitive personal information are commonly transmitted over email. This makes them valuable to cybercriminals, who can exploit the knowledge for their gain. Phishing is a technique used by con artists to steal sensitive information from people by impersonating well-known sources. The sender of a phished email can persuade you to disclose personal information under pretenses. The detection of a phished email is treated as a classification problem in this research, and this paper shows how machine learning methods are used to categorize emails as phished or not. SVM classifier attains a maximum accuracy of 0.998 percent in email classification. This research examines the various state-of-the-art machine learning (ML) algorithms currently used to detect phishing emails at different stages of the attack. A comparative assessment and analysis of these methodologies are performed. This provides an overview of the topic, its immediate solution space, and potential future research possibilities.

Keywords: SVM classifier, Cybercriminals, Machine learning and Phishing emails

1.INTRODUCTION

1.1 PHISHING

Phishing is the most common type of cybercrime that involves persuading victims to submit sensitive information such as account numbers, passwords, and bank account numbers.

1.2 EMAIL PHISHING

Cyber-attacks are commonly launched using email, instant messages, and phone calls. Despite continual updates to the procedures for preventing such cyber-attacks, the result is insufficient. On the other hand, phishing emails have expanded tremendously in recent years, indicating the need for more effective and modern measures to combat them. Several approaches for filtering phishing emails have been developed. However, the problem still requires a comprehensive solution. This is the first poll we're aware of that focuses on applying Machine Learning (ML) approaches to detect phishing emails. This research examines the various state- of-the-art machine learning (ML) algorithms currently used to detect phishing emails at different stages of the attack. A comparative assessment and analysis of these methodologies are performed. This provides an overview of the topic, its immediate solution space, and potential future research possibilities. The rapid advancement of internet technologies has changed the way people interact online while also posing new security risks. Newly growing global dangers attack the user's computer and have the potential to steal their identity and money. Phishing is a term with thousands of references in scientific papers, a lot of press coverage, and scrutiny from banks and law enforcement agencies. However, this raises the question of what phishing is. In some publications, the phenomenon of phishing is expressly described; in

others, it is presented with an illustration, while others assume that the reader already understands what phishing is. Many academics have offered their definitions of phishing, resulting in a wide range of interpretations in the scholarly literature. Because the phishing issue is broad and covers a multitude of circumstances, the literature does not provide a detailed description of phishing attacks. The term phishing was coined in 1996 as a result of social engineering attacks by web scammers against America Online (AOL) accounts, according to the APWG. Detecting phished email in the proposed system can be regarded as a classification problem with two types, ham and phished. Machine learning is a branch of artificial intelligence. When a system is given the ability to learn, it is intelligent. Without explicitly programmed, supervised learning is a concept that we use in our model. For classification, machine learning techniques are utilized.

2. LITERATURE REVIEW

2.1 PHISHING ENVIRONMENTS, TECHNIQUES, AND COUNTERMEASURES: A SURVEY

Phishing has become an increasing threat in online space, largely driven by the evolving web, mobile, and social networking technologies. Previous phishing taxonomies have mainly focused on the underlying mechanisms of phishing but ignored the emerging attacking techniques, targeted environments, and countermeasures for mitigating new phishing types. This survey investigates phishing attacks and anti-phishing techniques developed not only in traditional environments such as e-mails and websites, but also in new environments such as mobile and social networking sites. Taking an integrated view of phishing, we propose a taxonomy that involves attacking techniques, countermeasures, targeted environments and communication media. The taxonomy will not only provide guidance for the design of effective techniques for phishing detection and prevention in various types of environments, but also facilitate practitioners in evaluating and selecting tools, methods, and features for handling specific types of phishing problems. This study reveals that anti-phishing research and development has focused on phishing in e-mails and websites, but paid little attention to that in IM, social networks, voice, blogs and web forums; further, phishing in mobile communication has yet to be explored from the technical perspective.

2.2 EXPLORING SUSCEPTIBILITY TO PHISHING IN THE WORKPLACE

Phishing emails provide a means to infiltrate the technical systems of organizations by encouraging employees to click on malicious links or attachments. Despite the use of awareness campaigns and phishing simulations, employees remain vulnerable to phishing emails. The present research uses a mixed methods approach to explore employee susceptibility to targeted phishing emails, known as spear phishing. In study one, nine spear phishing simulation emails sent to 62,000 employees over a six-week period were rated according to the presence of authority and urgency influence techniques. Results demonstrated that the presence of authority cues increased the likelihood that a user would click a suspicious link contained in an email. In study two, six focus groups were conducted in a second organization to explore whether additional factors within the work environment impact employee susceptibility to spear phishing. We discuss these factors in relation to current theoretical approaches and provide implications for user communities. Work-based norms and routines likely represent a primary factor impacting response Behaviour within the workplace, influencing the development of context-specific habits, expectations and perceptions of risk. Reflective of the combined findings of Study One and Two, considering aspects of the email that is received, the individual who receives it, and the context in which it is encountered, within theoretical approaches is vital if susceptibility within the workplace is to be truly understood. It is hoped that the findings of the current study will provide a basis for further theoretical development in this field, whilst also presenting an initial aid for user communities to consider, and begin to address, the range of potential susceptibility factors that may be present within organizational settings.

2.3 INTELLIGENT DEEP MACHINE LEARNING CYBER PHISHING URL DETECTION BASED ON BERT FEATURES EXTRACTION

Recently, phishing attacks have been a crucial threat to cyberspace security. Phishing is a form of fraud that attracts people and businesses to access malicious uniform resource locators (URLs) and submit their sensitive information such as passwords, credit card ids, and personal information. Enormous intelligent attacks are launched dynamically with the aim of tricking users into thinking they are accessing a reliable website or online application to acquire account information. Researchers in cyberspace are motivated to create intelligent models and offer secure services on the web as phishing grows more intelligent and malicious every day. In this paper, a novel URL phishing detection technique based on BERT feature extraction and a deep learning method is introduced. BERT was used to extract the URLs' text from the Phishing Site Predict dataset. Then, the natural language processing (NLP) algorithm was applied to the

unique data column and extracted a huge number of useful data features in terms of meaningful text information. Next, a deep convolutional neural network method was utilised to detect phishing URLs. It was used to constitute words or n-grams in order to extract higher-level features. Then, the data were classified into legitimate and phishing URLs. The experiments showed that the proposed method had achieved 96.66% accuracy in the results, and then the obtained results were compared to other literature review works. The results showed that the proposed method was efficient and valid in detecting phishing websites' URLs.

2.4 MACHINE LEARNING TECHNIQUES FOR DETECTION OF WEBSITE PHISHING: A REVIEW FOR PROMISES AND CHALLENGES

Websites phishing is a cyber-attack that targets online users to steal their sensitive information including login credentials and banking details. Attackers fool the users by presenting the masked webpage as a legitimate or trustworthy to retrieve their essential data. Several solutions to phishing websites attacks have been proposed such as heuristics, blacklist or whitelist, and Machine Learning (ML) based techniques. This paper presents the state of art techniques for detection of phishing websites using the ML techniques. This research identifies solutions to the website's phishing problem based on the ML techniques. The majority of the examined approaches are focused on traditional ML techniques. Random Forests (RFs), Support Vector Machines (SVMs), Naïve Bayes (NB), and Ada Boosting are the powerful ML models examined in the literature. This survey paper also identifies deep learning-based techniques to demonstrate better performance for detecting phishing websites compared to the conventional ML techniques. Challenges to ML techniques identified in this work include overfitting, low accuracy, and ML techniques' ineffectiveness in case of unavailability of enough training data. This research suggests that Internet users should know about phishing to avoid cyber-attacks. Phishing attacks present negative impacts on web owners and end-users. The reputation of website owners becomes questionable when attackers launch an attack, and as a result of the attack, website users lose their sensitive information.

2.5 PHIBOOST – A NOVAL PHISHING DETECTION MODEL USING ADAPTIVE BOOSTING APPROACH

Every day, cyberattacks increase and use different strategies. One of the most common cyberattacks is Phishing, where the attacker collects sensitive and confidential information by pretending as a trusted party. Different traditional strategies have been introduced for anti-phishing, such as blacklisted, heuristic search and visual similarity. Most of these traditional methods have a high false rate and take a long time to detect the phishing website. New modes have been introduced using machine learning techniques which improve the detection's accuracy. Machine learning techniques require a huge amount of data called features that are collected from different websites. These collected features are classified into four categories. This paper introduces a novel detection model by utilizing features' selection to pick up the highly correlated features with the class label. The phase of features' selection employs independent significance features library from MATLAB and heat-map from Python to find the highly correlated features. The results of this study explored the best splitting rate for the dataset to train the machine learning model, which was 70%. Then, the proposed model uses an adaptive boosting approach which consists of multiple classifiers to increase the model's accuracy. The proposed model produces an extremely high predictive accuracy of approximately 99%.

3. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

For identifying legal and fraudulent web pages, is-based phishing detection systems use two lists: white lists and blacklists. Phishing detection systems that use whitelists create secure and genuine websites that deliver relevant information. Every website that isn't on the white list is regarded as potentially dangerous. [12] built a system that creates a whitelist by logging the IP address of each site that the user has visited with a Login user interface. When a user accesses a website, the system will alert them if its registered information is incompatible. The authors of [13] classified phishing websites using URL parameters such as length, number of unique characters, directory, domain name, and file name to identify them. The system uses support Vector Machines to classify websites that are not online. Adaptive Regularization of Weights, Confidence Weighted, and Online Perceptron are utilized for online classification. According to the trials ' findings, using the Adaptive Regularization of Weights algorithm improves accuracy while reducing system resource requirements. Authors in [14] used a nonlinear regression technique to detect whether a website is phishing or not in a recent study. They train the system using harmony search and support vector machine meta-heuristic techniques. Harmony search, they claim, has a higher accuracy rate of 94.13 percent and 92.80 percent for train and test procedures, respectively, thanks to the use of around 11,000 web pages. In [15] created a phishing detection system that uses adaptive

self- structuring neural networks to classify the data. It has 17 features, some of which are reliant on third-party services. As a result, real-time execution takes substantially longer; yet, it can achieve higher accuracy rates. It only has 1400 items in its dataset, yet it has a reasonable acceptance rate for noisy data. Yank in [16] provides an anti-phishing strategy that employs machine learning to identify phishing websites from legal ones by extracting 19 features from the client side. They used Phish Tank (2018) and Openfish (2018) phishing pages and 1918 authentic web pages from Alexa popular websites, online payment gateways, and prominent banking websites. Their proposed approach achieved a 99.39 percent true positive rate using machine learning [4].

3.2 PROPOSED SYSTEM

The attackers add subdomains to the links to make them appear authentic. The number of dots in the link rose as subdomains were added. As suggested by in a valid email, the number dots should not be used. More than three. This is a binary feature, meaning it determines whether or not a link exists. It would be in the mail if the number of dots was more prominent than three. This is a phished email. The total number of links is: In general, phishing emails provide more information. In comparison to ham, the transmitter attempts to send many links. By tricking the user, you might direct him to an illicit website. This is a recurring feature. The presence of JavaScript in an email indicates that the sender is either trying to conceal information or activate specific browser changes [16]. This is a one-of-a-kind feature. The presence of the script> tag in an email indicates that it has been phished. Form tag: Phishing emails feature forms integrated into them to acquire information from users. This is a binary characteristic, meaning that the presence of a form tag indicates that the email is phished.

HTML emails allow the sender to include embedded graphics and URLs, which are not possible with plain text emails. If the email has an HTML tag, it is considered phishing. This is a one-of-a-kind feature. The use of action words in emails shows if the sender expects the recipient to do a specific action, such as clicking on a link, filling out a form, or submitting detailed information. This is a recurring feature. The word PayPal indicates that the sender is posing as a member of a recognized organization. The word "PayPal" appears in the mail's links or the "from" section, implying that the sender is affiliated with PayPal. This is a one- of-a-kind feature. The presence of the term bank is a binary indicator indicating the message is about banking. Either the sender is posing as a member of the financial organization, or the reader is looking at the reader's credentials. The word account appears in the email, indicating that it seeks emails tied to an account. It could be a social media account, a bank account, or something else entirely. It's a one of-a-kind feature.

3.3 LIST OF MODULES

Service Provider
View and Authorize Users
Remote User

3.3.1 SERVICE PROVIDER

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as

Login,
Browse E Mail Data Sets and Train & Test,
View Trained and Tested Accuracy in Bar Chart,
View Trained and Tested Accuracy Results,
View Predicted E Mail Phishing Detection Type,
Find E Mail Phishing Detection Type Ratio,
Download Predicted Data Sets,
View E Mail Phishing Detection Ratio Results,

3.3.2 VIEW AND AUTHORIZE USERS

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as,

User name,
Email,
Address and Admin authorizes the users.

3.3.3 REMOTE USER

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user's name and password. Once Login is successful user will do some operations like

Register and Login,

Predict Email Phishing Detection Type,

View your Profile.

4. SYSTEM DESIGN AND DEVELOPMENT

4.1 INPUT DESIGN

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations. This system has input screens in almost all the modules. Error messages are developed to alert the user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design. Input design is the process of converting the user created input into a computer-based format. The goal of the input design is to make the data entry logical and free from errors. The error is in the input are controlled by the input design. The application has been developed in user-friendly manner. The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with in an option to select an appropriate input from various alternatives related to the field in certain cases. Validations are required for each data entered. Whenever a user enters an erroneous data, error message is displayed and the user can move on to the subsequent pages after completing all the entries in the current page.

4.2 OUTPUT DESIGN

The Output from the computer is required to mainly create an efficient method of communication within the company primarily among the project leader and his team members, in other words, the administrator and the clients. A new user may be created by the administrator himself or a user can himself register as a new user but the task of assigning projects and validating a new user rests with the administrator only. The server has to be started and then the internet explorer in used as the browser. The project will run on the local area network so the server machine will serve as the administrator while the other connected systems can act as the clients.

5. CONCLUSION

This research proposes an intelligent approach for detecting phishing emails effectively. It examines the differences between Naive Bayes, Random Forests, and SVM. The goal is to find the most effective intelligent classification model for detecting email phishing. Different experiments were conducted on three benchmarking testing levels to evaluate the performance of the three classifiers. The presence of JavaScript in an email indicates that the sender is either trying to conceal information or activate specific browser changes [16]. This is a one-of-a-kind feature. The presence of the script> tag in an email indicates that it has been phished. Form tag: Phishing emails feature forms integrated into them to acquire information from users. This is a binary characteristic, meaning that the presence of a form tag indicates that the email is phished. We plan to test SVM's performance on different benchmarking datasets in the future. Performance comparison of SVM with various kernels, such as Gaussian or sigmoid kernels, will also be carried out.

6. REFERENCES

[1] A. Aleroud and L. Zhou (2018), "Phishing environments, techniques, and countermeasures: A survey," Computers & Security, vol. 68, pp. 160-196, 2017. [2] I. Vayansky and S. Kumar (2018), "Phishing–challenges and solutions," Computer Fraud & Security, vol. 2018, pp. 15-20.

[3] E. J. Williams (2018), "Exploring susceptibility to phishing in the workplace," International Journal of Human-Computer Studies, vol.120, pp .1-13.

[4] Ammar Odeh, (2021), "Machine Learning Techniques for Detection of Website Phishing: A Review for Promises and Challenges," in 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0813-0818.

[5] A. Odeh, (2021), "PHIBOOST-a novel phishing detection model using Adaptive boosting approach," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 7.

[6] K. L. Chiew, (2018), "A survey of phishing attacks: Their types, vectors and technical approaches," Expert Systems with Applications, vol. 106, pp. 1-20.

[7] M. Al-Fayoumi (2020), "Intelligent association classification technique for phishing website detection," International Arab Journal of Information Technology, vol. 17, pp. 488-496.

[8] Y. Kwak, (2020), "Why do users not report spear phishing emails?" Telematics and Informatics, vol. 48, p. 101343.

[9] G. Sonowal and K. Kuppusamy (2020), "Phi DMA–A phishing detection model with multi-filter approach," Journal of King Saud University-Computer and Information Sciences, vol. 32, pp. 99-112.

[10] I. Keshta and A. Odeh (2021), "Security and privacy of electronic health records: Concerns and challenges," Egyptian Informatics Journal, vol. 22, pp. 177-183.

[11] R. Faek (2021), "Exposing Bot Attacks Using Machine Learning and Flow Level Analysis," in International Conference on Data Science, E-learning and Information Systems 2021, pp. 99-106.

[12] A. ODEH (2020), "Efficient Prediction of Phishing Websites Using Multilayer Perceptron (Mlp)," Journal of Theoretical and Applied Information Technology, vol. 98.

[13] O. K. Sahingoz (2019), "Machine learning based phishing detection from URLs," Expert Systems with Applications, vol. 117, pp. 345-357.

[14] A. Oest (2018), "Inside a phisher's mind: Understanding the anti-phishing ecosystem through phishing kit analysis," in 2018 APWG Symposium on Electronic Crime Research (eCrime), pp. 1-12.

[15] A. Abbasi, et al. (2021), "The phishing funnel model: A design artifact to predict user susceptibility to phishing websites," Information Systems Research, vol. 32, pp. 410- 436.

[16] P. Yang (2020), "Phishing website detection based on multidimensional features driven by deep learning," IEEE Access, vol. 7, pp. 15196-15209.