Journal of Science and Technology
ISSN: 2456-5660 Volume 8, Issue 07 (July -2023)
www.jst.org.in *DOI:https://doi.org/10.46243/jst.2023.v8.i07.pp158- 162*

Effective And Efficient Detection Of Phishing Emails Using Machine Learning

Moola.Akshitha[1] | Dr.D.Srinivas Reddy[2] | Dr.V.Bapuji[3]

[1]Department of MCA, Vaageswari college of Engineering,Karimnagar
[2] Professor,Department of MCA, Vaageswari college of Engineering,Karimnagar
[3] HoD,Department, of MCA, Vaageswari college of Engineering,Karimnagar

To Cite this Article
Moola.Akshitha| Dr.D.Srinivas Reddy| Dr.V.Bapuji, "Effective And Efficient Detection Of Phishing Emails Using Machine Learning" *Journal of Science and Technology, Vol. 08, Issue 07,- July 2023, pp158-162*
Article Info
Received: 06-06-2023 Revised: 07-07-2023 Accepted: 16-07-2023 Published: 27-07-2023

ABSTRACT

Emails are widely used for personal and professional communication,often involving the transmission of sensitive information like banking details,credit reports,and login data.Consequently,these emails become valuable targets for cyber criminals who seek to exploit such knowledge for their own malicious purposes.Phishing, a deceptive technique employed by these individuals,involves impersonating well-known sources to deceive and extract sensitive information from unsuspecting individuals.The sender of a phishing email uses false pretenses to persuade recipients into disclosed personal information.In this work,the detection of phishing emails is learning methods to categorize emails as either genuine or phishing attempts.LMT classifiers have proven highly effective in accurately classifying emails,achieving optimal accuracy in email classification tasks.

KEYWORDS: Phishing,Emails,Efficient,Detection,Effective,Transmission.

INTRODUCTION

Phishing stands as the most prevalent form of cybercrime, involving the manipulation of victims to disclose sensitive information like account numbers, passwords,and bank details. Cyber-attacks commonly exploit email,instant messages,and phone calls[1,2].Despite ongoing efforts to update preventive measures,the outcomes have proven insufficient.Conversely,there has been a significant increase in phishing emails in recent years, underscoring need for more effective and modern countermeasures[3,4].Numerous approaches have been developed to filter phishing emails,but a comprehensive solution to the problem is still required.This study represents the first known survey focusing on the application of Machine Learning[ML] algorithms currently employed to detect the phishing emails at different stages of an attack[5].

It includes a comparative assessment and analysis of these methodologies ,offering an overview of the topic,its immediate solution space,and potential future research directions[6-8]The rapid advancement of internet technologies has transformed online interactions while introducing new security risks .Despite phishing being extensively referenced in scientific papers receiving press coverage and drawing attention from banks and law enforcement agencies the question of what phishing truly entails arises[10].

Some publications explicitly describe the phenomenon of phishing while others provide illustrations or assume prior reader familiarity. The variation in academic definitions has resulted in a wide range of interpretations in the scholarly literature. Due to the broad nature of phishing issue, which encompasses diverse circumstances, the existing literature lacks a detailed description of phishing attacks [1,2]. The term "phishing"originated in 1996 when web scammers conducted social engineering attacks against America online (AOL) accounts,as reported by the APWG. Detecting phished emails within the proposed system can be seen as a classification problem with two types:legitimate (ham) and phishing. Machine learning, as a branch of artificial intelligence, endows a system with the ability to learn and exhibit intelligence without explicit programming. Our model employs supervised learning concepts for classification, utilizing various machine learning technique [3,4].

1. RELATED WORK

Phishing detection system that aim to identify legal and fraudulent web pages utilize two lists: Whitelists and blacklists. Whitelists are employed by phishing detection system to designate secure and authentic websites that provide relevant information. Any website not listed on the whitelist is considered potentially dangerous. [5] developed a system that generates a whitelist by logging the IP addresses of visited sites through a login user interface. When a user accesses a website, the system alerts them if the registered information incompatible.

In a separate study, the authors of [5] classified phishing websites by analyzing URL parameters such as length, number of unique characters, directory,domain name and file name. Support Vector Machines are used for offline classification, while Adaptive Regularization of Weights,Confidence Weighted, and Online Perceptron are employed for online classification. The trials reveal that the use of the Adaptive Regularization of Weights algorithm improves accuracy while reducing system resource requirements.

Another recent study [6] employed a nonlinear regression technique to detect phishing websites. The system was trained using harmony search and support vector machine meta-heuristic techniques.With a dataset of around 11,000 web pages,harmony search achieved hing accuracy rates of 94.80% for the training and testing procedures,respectively.

In [7], a phishing detection system was developed using adaptive self-structuring neutral networks to classify the data. It utilize 17 features,some of which rely on third-party services.Although real-time execution takes longer,this approach higher accuracy rates and demonstrates reasonable acceptance for noisy data despite having only 1,400 items in its dataset.

Yank [8] proposed an anti-phishing strategy that utilizes machine learning to identify phishing websites from legitimate ones by extracting 19 features from the client side.Their approach involved analyzing PhishTank (2018) and Openfish (2018) phishing pages alongside 1,918 authentic web pages from popular Alexa websites,online payment gateways,and prominent banking websites.The proposed approach achieved a remarkable 99.39% true positive rate using machine learning techniques[4].

2. PROPOSED WORK

This research presents an intelligent neutral network model for the effective detection of phishing websites on the Internet. The model incorporates a classification algorithm to enhance efficiency in the detection process. To evaluate the performance of the intelligent algorithm, a web phishing dataset is utilized, with a specific focus on assessing its classification accuracy.

Phishing emails tend to include a higher number of links, providing excessive information compared to legitimate emails. The sender deliberately attempts to deceive the recipient my directing them to illicit websites. This pattern is commonly observed in phishing emails. The presence of JavaScript in an email signifies an attempt by the sender to either hide information or trigger specific browser modifications [8],making it a distinct characteristic. If an email contains the "Script>" tag, it is an indication of a phishing

attempt. Phishing emails often incorporate forms to extract information from users, and the presence of a form tag is a binary characteristic that signifies a phishing email.

HTML emails enable senders to include embedded graphics and URLs,which is not possible in plain text emails. If an email contains an HTML tag, it is considered indicative of a phishing attempt, representing another unique feature. The presence of the term “bank” serves as a binary indicator, suggesting that the message is related to banking. It indicates that the either the sender is impersonating a member of a financial organization or the email is targeting the recipient`s banking credentials. Additionally, if the word account” is found in the email, it implies a search for emails associated with an account, which could pertain to various types such as social media, bank, or others. This characteristic represents a unique feature.

3.1 Support Vector Machine

(SVM) is widely employed supervised technique for next categorization due to its notable speed and accuracy. By utilizing the training data, SVM generates a hyper-plane a two dimensional line that effectively separates different categories. This hyper-plane, known as the decision boundary, plays a crucial role in SVM`s classification process [9]. In the context of phishing detection, SVM takes a collection of features representing the input, such as the presence or absence of specific terms. The output, either 1 or -1, indicates whether the email is classified as a phishing attempt or not.

3.2 The naive Bayes

The Bayes theorem is a probabilistic tool that is used by the naive Bayes classifier[10] to categorize sample data. Given a hypothesis H and evidence E, the Bayes theorem states that the
The following relationship exists between the likelihood of the hypothesis P(H|E) after having the evidence and the probability of the hypothesis P(H) before having the evidence :

$$P(H \setminus E) = \frac{\mathbf{P(E \setminus H)}}{\mathbf{P(E)}} P(H)$$

The probability of each category is determined, and the one with the greatest likelihood is selected.

3.3 Random forests

It is also referred to as random decision forests, are an ensemble learning approach utilized for tasks such as classification, regression, and more. The method involves training a significant number of decision trees and subsequently determining the class that occurs most frequently (in classification) or calculating the average prediction (in regression) from the individual trees. Random decision forests effectively tackle the issue of decision trees overfitting to their training set [10].

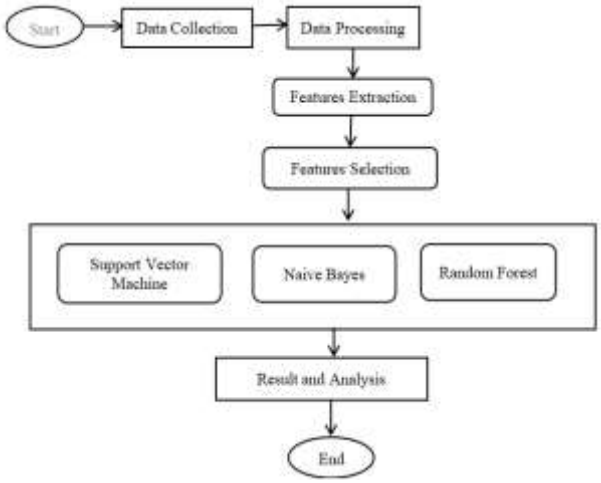

Figure 1: Block diagram of The Proposed System

3. RESULT AND ANALYSIS

The performance of the three intelligent classification algorithms is evaluated using the confusion matrix. The matrix represents the classifier`s performance on the input dataset, allowing for the derivation of various performance metrics, including accuracy and F-measure.

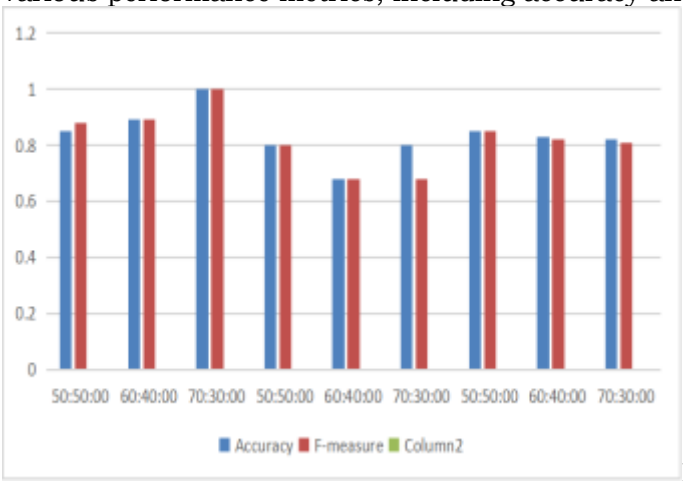

Figure 2:The histogram comparison between three classifier

It is crucial to consider that increasing the percentage of training data results in higher computation complexity and longer training time.To achieve a balance between the accuracy of categorization and minimizing incorrect classifications,a training dataset consisting of 30% of the total data is used.

4. CONCLUSION

In this work,an intelligent approach is proposed to enhance the effectiveness the phishing email detection.The study investing and compares the distinction among Naive Bayes,Random Forest,and Support Vector Machines (SVM) as intelligent classification models,with the objective of identifying the most efficient model for detecting email phishing.A serial of experiments is conducted on three

benchmarking testing levels to access the performance of these classifiers. Additionally,future plans include testing the performance of SVM using various kernels,such as Gaussian or sigmoid kernels. For future research,our intention is to access the performance of additional machine learning classifiers and compare them to identify the most effective one for enhancing URL security.By exploring different classifiers, to find the optimal approach that enhances the overall security of URLs.

5. REFERENCES

[1] Aleroud,A.,&zhou,L.(2017).Phishing environments,techniques,and countermeasures:A survey.Computers &security,68,160-196.

[2] Vayansky,I.,&Kumar,S.(2018).Phishing-challenges and solutions.Computer Fraud & Security,2018,15-20.

[3]. E. J. Williams and colleagues, "Exploring susceptibility to phishing in the workplace," International Journal of Human-Computer Studies, vol. 120, no. 1, 2018, pp. 1–13.

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

[5]. Effective Detection of Phishing Websites Using Multilayer Perceptron, A. Odeh et al., 2020.

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

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

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

[9] Why do consumers not report spear phishing emails? [9] Y. Kwak et al., Telematics and Informatics, vol. 48, p. 101343, 2020.

[10] "PHISHING WEBSITE DETECTION USING MULTILAYER PERCEPTRON," A. Odeh et al.