iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 Issue: V Month of publication: May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52342

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Phishing URL Detection using Machine Learning

Nematullah Noori[1], Vyenkatash Bawanthad[2], Mayur Pakhare[3], Ramashray Agrawal[4], Vinod Kimbahune[5]
[1, 2, 3, 4, 5]Department of Computer Engineering Dr D. Y. Patil Institute of Technology, Pimpri, Pune

Abstract: Phishing attacks continue to pose a major threat for computer system defenders, often forming the first step in a multi-stage attack. There have been great strides made in phishing detection; however,some phishing emails appear to pass through filters by making simple structural and semantic changes to the messages. We tackle this problem through the use of a machine learning classifier operating on a large corpus of phishing and legitimate emails. We design a system to extract features, elevating some to higher level feature, that are meant to defeat common phishing email detection strategies.

This paper presents an approach to detect phishing URLs in an efficient way based on URL features only. For detecting the phishing URLs SVM classifier is used. The performances are evaluated for different size of datasets using different number of features. The results are compared with other machine learning classification techniques. The proposed system is able to detect phishing websites using URL features only.

Keywords: Phishing, Phishing websites, Machine Learning, anti-phishing, phishing attack, security and privacy, phishing approaches

I. INTRODUCTION

With the steady acceleration in information technology, we are no longer immune to being victims of cybercrime. The use of the Internet has become essential in the modern era and an integral part of technological development, which leads to discoveries and reduction of time, effort, and costs.

Nevertheless, this provides a fertile ground for piracy expansion in exploiting the weaknesses to determine private and public interests. Although cybercrime does not differ much from traditional crimes in terms of its perpetrators' goal, because these crimes are based on unlawful targets, cybercrime has become more widespread than traditional crimes. It has become a core part in the world of digitization, as intercontinental crimes within cyberspace. Digital cybercrimes have no limits and are easy to implement. Creating a paperless environment has become a major focus in most countries worldwide, increasing dependency on these channels. On the other hand; unprotected websites may allow fake announcement exploits under circumstances that occupy public opinion (for instance new Corona pandemic (COVID-19)). This leads the victim to a phishing website. In this context, individuals' lack of awareness in information security plays a key role in increasing the number of victims of this crime.

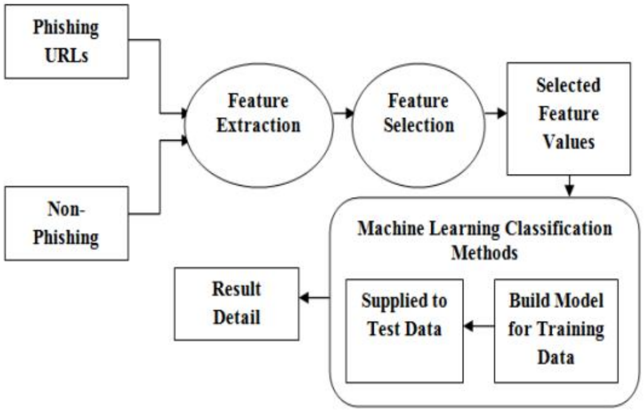
This work focuses on a URL phishing attack [1]. Phishing can be defined as impersonating a valid site to trick users by stealing their personal data comprising usernames, passwords, accounts numbers, national insurance numbers, etc. Phishing frauds might be the most widespread cybercrime used today. There are countless domains where phishing attack can occur like online payment sector, webmail, and financial institution, file hosting or cloud storage and many others. The webmail and online payment sector was embattled by phishing more than in any other industry sector. Phishing can be done through email phishing scams and spear phishing hence user should be aware of the consequences and should not give their 100 percent trust on common security application. Machine Learning is one of the efficient techniques to detect phishing as it removes drawback of existing approach[3].

II. RELATED WORKS

Many of the researchers proposed different techniques for detecting phishing URLs. Some of them have maintained a list of domain name or IP addresses of previously detected phishing websites. A system named Phishnet is proposed [Prakash et al., 2010] where a blacklists of phishing URL was maintained. It will check whether IP address, hostname or the URL itself belong to that blacklist or not. They have also proposed five heuristics to detect phishing URLs. An approach of maintaining whitelist method is proposed in [Jain and Gupta, 2016] containing the domain name and corresponding IP address of legitimate sites instead of blacklist techniques.[4] The system first checks whether a particular site is present in the list or not. If it is not, the system checks by extracting number of hyperlinks contained in the site. If the number of hyperlinks is NULL or zero or greater than certain threshold value, it is declared as phishing. Otherwise, it is declared as legitimate and added in whitelist

Proposed Methodology:

The design of our system is build up as shown in Figure. First, dataset of phishing and legitimate URLs are collected. Lexical features of these URLs are extracted. Feature selection method is used to find the important features only. This method provides ranking to each feature based on their contribution to detect phishing and non-phishing classes [2]. The performance by taking different number of features is compared using different algorithms. The features of lower ranks are removed which are found to have low contribution to detect the classes. Then, the performances of various classification methods are analyzed for different numbers of URLs [5].



III. MACHINE LEARNING ALGORITHMS

Machine learning technology consists of many algorithms such as Decision Tree, Naive Bayes, Random Forest, Logistic Regression, and K-Nearest-Neighbor (KNN), Support Vector Machine (SVM) forphishing detection. Among these, Support Vector Machine or SVM is a very popular algorithm that has proved to be very efficient and accurate compared to the other algorithms.

I. Algorithm	II. Time Complexity	III. Training Data Size	IV. Interpretability
V. SVM	VI. $O(n^2)$	VII. Small	VIII. Median
IX. Decision Tree	X. O(nd log n)	XI. Small	XII. High
XIII. Naïve Bayes	XIV. O(nd)	XV. Small	XVI. High
XVII. k-NN	XVIII. O(and)	XIX. Small	XX. Median
XXI. Random Forest	XXII. O(knd log n)	XXIII. Small	XXIV. Median

IV. SUPPORT VECTOR MACHINE(SVM)

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

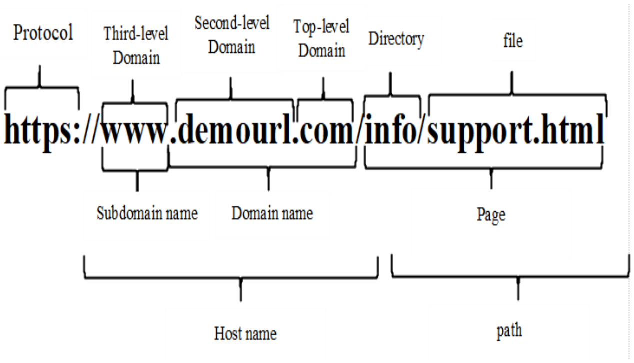
V. DATASETS

Typically, the phishing site information is gathered from kaggle.com. kaggle.com is a site where phishing URLs are recognized and can be gotten to through API call. Their information is utilized by organizations like Kaspersky, Mozilla, and Avast. Since it doesn't store the substance of website pages, it is a decent hotspot for URL-based examination.

VI. FEATURE EXTRACTION

URLs have specific qualities and examples that can be considered as its elements. Fig. 3 shows the pertinent pieces of a normal URL. On account of URL-based investigation for planning AI models, we really want to extricate these highlights to shape a dataset that can be utilized for preparing and testing. There are four classes of elements that are most usually considered for include extraction as in [10]. They are as per the following:

1) Address bar-based features
2) Abnormal based features
3) HTML and JavaScript-based features
4) Domain-based features



VII. PERFORMANCE EVALUATION METRICS

To evaluate the efficiency of a system, we use certain parameters. For each machine learning model, we compute the Accuracy, Precision, Recall, F1 Score, and ROC bend to decide its exhibition. Every one of these measurements is determined dependent on True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

On account of URL classification, True Positive (TP) is the quantity of phishing URLs that are accurately named phishing. True Negative (TN) is the quantity of real URLs that are accurately named authentic. False Positive (FP) is the quantity of genuine URLs that are named phishing. False Negative (FN) is the quantity of phishing URLs that are named genuine. These qualities are summed up in the table called Confusion Matrix.

TABLE OF CONFUSION MATRIX FOR PHISHING DETECTION

	Predicted Phishing	Predicted Legitimate
Actual Phishing	TP	FN
Actual Legitimate	FP	TN

Precision is the quantity of URLs that are phishing out of the multitude of URLs anticipated as phishing. It estimates the classifier's precision. The recipe to work out precision is given by Equation (1) beneath.

Precision = ……. (1)

$$\text{Precision}=\frac{TP}{(TP+FP)} * 100\%$$

Recall is the quantity of URLs that the classifier recognized as phishing out of the relative multitude of URLs that are phishing. It is likewise called sensitivity or True positive rate. It is a significant measure and ought to be pretty much as high as could be expected. The formula to compute Recall is given by Equation (2) beneath.

$$\text{Recall} = \frac{TP}{(TP+FN)} * 100\%$$

Recall = ……. (2)

F1-Score is the weighted normal of accuracy and recall. It is utilized to quantify accuracy and recall simultaneously. The formula to compute F1-Score is given by Equation (3) beneath.

F1Score=2* … (3)

$$\text{F-Score}=2*\left(\frac{Precision*Recall}{Precision+Recall}\right)$$

Accuracy is the quantity of cases that were accurately ordered out of the relative multitude of cases in the test information. The recipe to ascertain exactness is given by Equation (4) beneath

Accuracy = …. (4)

$$\text{Accuracy, ACC}=\frac{(TP+TN)}{(TP+TN+FP+FN)}*100\%$$

VIII. OBSERVATIONS

Phishing attack are continually advancing and the digital world is hit by new kinds of assaults frequently. Consequently, a specific location approach or calculation can't be labeled as the best one giving precise outcomes. Through the writing study, we discovered that Support Vector Machine gives better outcomes in many situations. However, at that point the exhibition of every calculation differs relying upon the dataset utilized, train-test split proportion, highlight determination strategies applied, and so forth Scientists like to make AI models that perform phishing location with the best incentive for assessment boundaries and least preparing time. Subsequently, our future works center around working on these parts of phishing identification.

IX. CONCLUSION

Phishing detection is currently an area of incredible interest among specialists because of its importance in ensuring the protection and giving security. Numerous techniques perform phishing location by characterization sites utilizing prepared AI models. In this paper, we depicted our precise study of existing URL-based phishing identification procedures from various perspectives. Albeit past overview papers exist, they by and large spotlight on in general phishing location methods, while we zeroed in on itemized URL-based discovery concerning highlights. Right off the bat, we audited the writing on by and large phishing identification plans. Second, we examined the design of URL-based phishing, and ordinarily utilized calculations and highlights. Third, normal information sources were recorded, and near assessment results and grids were displayed for a superior study. At long last, we closed with our idea to continue with the Support Vector Algorithm for more successful phishing URL identification in our venture.

REFERENCES

[1] Prajakta Patil, Rashmi Rane, Madhuri Bhalekar “Detecting spam and phishing mails using svm and obfuscation detection algorithm,”. 2017 International conference on inventive systems and control (ICISC).

[2] Bireswar Banik, Abhijit sarma “Phishing URL dectection system based on URl features using SVM,”. International journal of electronics and applied Research vol.5,issue 2, Dec 2018.

[3] Mohammed Abutaha, Mohammad Ababneh, Khaled Mahmoud, Sherenaz W. Al-Haj Baddar "URL phishing detection using machine learning techniques based on URLs lexical analysis,”. 2021 12th international conference on information and communication systems (ICICS).

[4] Almomani, B. B. Gupta, S. Atrawneh, A. Meulenberg and E. Almomani, “A Survey of phishing email filtering techniques,” in IEEE communications surveys and tutorials, vol. 15, no. 4, pp.2070-2090

[5] G. J. W. Kathrine P. M. praise, A. A. Rose and E. C. Kalaivani, “variants of phishing attacks and their detection techniques,” in 2019 3rd international Con ference on Trends in electronics and informatics, Tirunelvei.