Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

An Improved Classification Model For Identifying The Phishing Attacks

Vinod Sapkal[1] , Dr. Ninad More[2]

[1]Department of Computer Science and Information Technology, CSMU, Navi Mumbai, Panvel. Maharashtra, India.

[2]Department of Computer Science and Information Technology, CSMU, Navi Mumbai, Panvel, Maharashtra, India.

Abstract — An attack known as phishing is a malicious scheme that is used to steal private information from users by tricking them into attacking via a fake website that has been designed to imitate and look very similar to an actual website. The perpetrator of the attack will steal the user's private information, including their username, password, and personal identification number (PIN), and then use that information to make fraudulent transactions. The credentials of the information holder, as well as any money they may have, will be seized. The phishing website and the legitimate website will have a high degree of understandable similarity, which will allow an attacker to steal the user's credentials from the legitimate website. There are a variety of methods available, including blacklisting, whitelisting, heuristics, and machine learning, which can be utilised in order to detect phishing attempts. Learning machines are used these days, and it has been proven that they are more effective. The phishing website's source code features, as well as its URL features and picture features, are all extracted by the proposed approach. In order to obtain the reduced features, the characteristics that have been extracted are input into the algorithm for ant colony optimization. The decreased features are then provided to the Naive Bayes classifier once more in order to determine whether the website in question is authentic or phished.

Keywords — Phishing, Ant colony Optimization, Naïve Bayes Classifier, Feature Extraction.

I. INTRODUCTION

People who utilised the internet are collectively referred to as "Netizens" because there are so many internet users today. These online services are convenient for all customers, and they also provide benefits in the form of cost savings, time savings, and labour savings.

Unfortunately, the rise in widespread phishing assaults against internet users has cast a shadow over the usability of online services. These attacks are becoming increasingly sophisticated. Phishing is a sort of identity theft in which the perpetrators attempt to get access to the private details and financial credentials of consumers who shop online.

Phishing attacks typically consist of four stages: the preparation stage, the mass broadcast stage, the mature stage, and the account hijacking stage. Phishing is a method of identity theft in which a fraudulent website poses as a genuine one in order to trick users into divulging important information like passwords, account details, or credit card numbers.

Phishing is a method of gathering sensitive information via the use of deceit. This information can be

obtained by pretending to be a reliable person or company in an internet contact. There were 18480 different websites, as stated by the anti-phishing working group. There were 9666 distinct phishing sites and unique phishing assaults recorded in the month of March in 2006. When compared to 2017, the total number of phishing attacks carried out in 2018 was 59% greater. It would appear that phishing was successful in setting another record year in attack volumes, with anticipated global losses from phishing in 2019 totaling $1.5 billion. This is a 22% rise from the previous year, 2017.

Internet access is becoming an essential part of almost everyone's day-to-day activities. Because technological advancements are being made at such a breakneck pace, users are forced to become savvier in their application of these tools. When technology advances, it inevitably has an effect on other areas of study as well. There are some vulnerabilities that are present on the internet, and because of this, they can be exploited to launch an attack against the user. Pharming is one of the many assaults that can take place over a network. In this attack, the attacker poses as a legitimate entity in order to steal user credentials. To entice the user by having a large number of visual similarities. According to the study that was compiled during the first three months of 2016, India was given the position of fifth place [11] among the top ten countries that were targeted by phishing attacks. Users who are completely oblivious to this attack are at risk of falling into the trap. This study examines the source code, URL, and image characteristics of a website. Then then selects the most important characteristics of the website through the process of ant colony optimization, and it uses a Bayesian classifier to determine whether or not the website engages in phishing.

II. RELATED WORK

Phishing and spamming are two of the most common types of attacks that may be carried out in cyberspace, and the author focuses on them here in [1]. In order to identify the spam attack, they used text mining and data mining techniques. For the purpose of finding phishing attacks, they retrieved information from the webpages, such as the source code and the URL. For these, they have gathered the dataset from phishtank containing faked websites and the Enron-spam corpus containing spam emails. Genetic Programming, Logistic Regression, Probabilistic Neural Network, Multi-Layer Perceptron, and Classification and Regression Tree are the training methods that are utilised for the classifiers.

The author of [2] offered a method for preventing phishing that involved extracting the source code of the URL, including meta, title, and body tags, among other things. They have focused more on the left-hand side of the URL rather than the right-hand side of the URL since the attacker is trying to spoof the phished website as the legitimate website. This is why they have focused on the left-hand side of the URL. Following completion of the matching process, the full URL is split up into tokens, and each token's identity keywords are then cross-referenced with Yahoo's search engine. Both the original domain name and the provided domain name are compared against each other and then compared against the country code top level domain. If the presence of a webpage fits with the country code top level domain, then the webpage is deemed to be authentic; otherwise, the webpage is considered to be phished.

The author places an emphasis in [3] on Chinese websites that are used for phishing. URL and web attributes, as well as a sequential minimal optimization technique, are the components that are utilised in the phishing detection process. They have utilised genetic algorithm in order to optimise the parameters of the characteristics. The WebZIP tool is used for gathering and downloading the source code of the e-commerce webpage, and the Weka data mining tool is used for training the suggested system. Both of

these tools may be found on this page.

An approach based on machine learning is presented by the authors of [4] as a method for the identification of phishing websites. This paper places an emphasis on aspects of the website, such as web images and document object models. In order to optimise the aspects that are extracted from the website, the authors make use of an evolutionary algorithm that is inspired by quantum mechanics. Following optimization, the features are fed into a transductive support vector machine, which determines whether the website in question is a phishing scam or not.

The authors of this study [5] proposed a model for identifying potentially malicious URLs; the model has three modules. The work of the modules responsible for data collection is to collect the URL links that have been posted. The feature extraction module is what's responsible for pulling the feature vectors out of the data. Within the classification module, they have implemented a Bayesian classifier in order to identify the potentially malicious URL. This model does not put any restrictions on blacklists; rather, it places an emphasis on URLs and behaviour in social links. Domain anomaly is what identifies potentially harmful URLs, while social anomaly looks at user behaviour to spot suspicious patterns.

The author of [6] uses a Neuro fuzzy approach to identify potential phishing websites. In addition, it utilises If...Then rules to distinguish between phishing websites, suspicious websites, and legitimate websites. It issues a warning, either in the form of a vocal alert or a colour signal, according to the seriousness of the situation.

The authors of [7] presented a method that may be used to identify the spammers. The spammers and non-spammers are separated into different categories according to the content and user characteristics. They have utilised a machine learning approach in order to extract the attributes in order to identify the spammers. After the features have been retrieved, they are fed into a support vector machine, which performs the function of a classifier by sorting the incoming data into spammers and non-spammers.

The author of this piece [8] proposed a method to classify the phishing websites that are used for electronic banking, and it makes use of methodologies such as fuzzy logic and data mining algorithm. For this purpose, they have used c4.5, Ripper, Part, Prism, and CBA as examples of fuzzy logic. The fuzzy logic identifies the keywords that are associated with phishing.

The author of [9] put forward the idea of a hybrid approach to the prevention of phishing. It functions as a plugin for web browsers. Whitelists and blacklists are the foundation of the tactics that are utilised for phishing websites. If the arriving URL has a previous match on the blacklist, it will immediately block that URL. When an incoming URL is checked against the whitelist, the page will load if it is a match. If the webpage does not match either of these, it will immediately forward the URL to a moderator so that they may determine whether or not the inbound webpage is a phishing attempt.

The author of the article [10] sheds some light on the feature selection process, specifically focusing on the correlation and wrapper approaches that can identify phishing. The results were analysed and compared using genetics and aggressive forward selection within the framework of learning algorithms for real-time datasets on phishing.

III. PROPOSED SYSTEM

To identify whether the incoming webpage is a validor false webpage the features of the webpage are extracted. These features are clustered as

- Source code featuresURL features
- Image features

A. Source Code Features

1) **Tracking of login screen:** This feature checks if itcontains any text box in order to get information from the user such as username, password, and PIN numbers.

2. **Disabling Right Click:** The attacker disables the right click so that the user cannot able to visualize the code of the website.

3. **Pop Up:** The phishing sites pop up with some messages to enter their credentials. The legitimate site does not ask them to enter their credentials.

B. URL Features

1) **IP address:** This feature examine whether thewebpage has IP address or not. Normally, the legitimate website uses its own domain name for verification. Occasionally, the attacker uses hexadecimal codes in which the IP address is converted into hexadecimal form thereby the attacker grabs the user's identity.

2) **Special Symbols:** Using @ symbol in the URL leads the browser to ignore everything preceding by @ symbol and the real address often follows @ symbol.

3) **Tracking Single slashes:** Normally, the phished websites contains more number of single slashes but the legitimate website contains not more than three slashes.

4) **Shortening services:** Tiny URL is considered to be an URL shortening services by which it produces redirecting of lengthy URL to some other page.

5) **Big domain URL:** Phishers can use long URL to hide the doubtful part in the address bar.

6) **Usage of prefix/suffix:** The legitimate webpage does not contain any special symbol in their URL but it uses rarely in legitimate page. The attacker uses dash symbol to separate domain name thereby itlooks like an original website.

7) **Domain Registration Length:** The phishingwebsite lives only for a short duration of time. But, the legitimate site renews their domain by paying regularly.

8) **Number of Dots:** The legitimate site containsdoes not exceed more than three dots. But the phishing site contains many dots in the URL.

C. Image Features

1) **Grayscale:** In this, the image contains only single value either 0 or 1. The value 0 is for black and 1 is for white. It just transmits only the strength of the information.

2) **Color Histogram:** In this, the pixels are categorizedaccording to the intensity of the colored image.

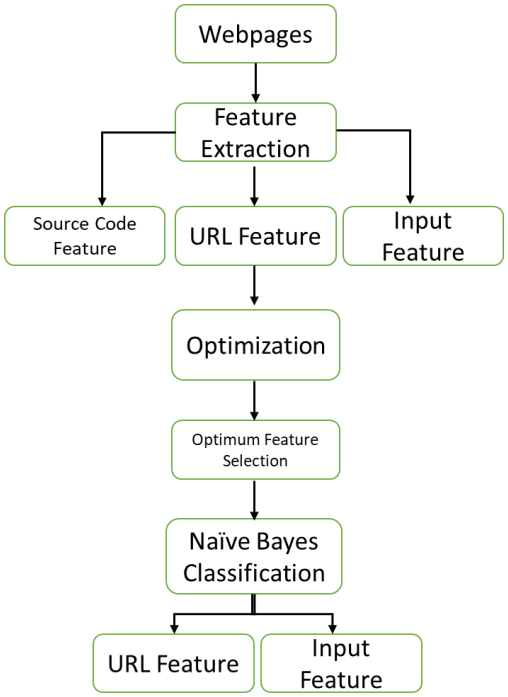

Fig. 1 Architecture of the Proposed System

Initially the ants are positioned on the nodes. The input for ant colony optimization are source code, URL and image features. The source code and URL features are passed in the form of links whereas the image features are given in the form of pixels. The ants are arbitrarily allocated for single feature, allowing the ant for visiting the feature and constructcomplete solution. Each and every ant returns the solution at the end of the cycle. The pheromone is updated by pheromone trail updating rule. The stopping criteria is the range of iteration. Finally the best features are extracted by ant colony optimization technique. The extracted features are given as the input to Naïve Bayes classifier. According to the Bayes theorem the Naïve Bayes classifier classifies whether a given webpage W is phishy or legitimate by using the formulas given in equation 1 and 2

$$p(W|P) = \frac{p(W \cap P)}{p(P)} \qquad (1)$$

and

$$p(P|W) = \frac{p(W \cap P)}{p(W)} \qquad (2)$$

Where,
P belongs to class phishy and W is Webpage takenfor consideration.

IV. EVALUATION PARAMETER

The prediction accuracy of the system is taken as theevaluation parameter and it is computed using the formula given in equation (3)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (3)$$

True Positive (TP): A positive occurrence isproperly classified as positive.
False Positive (FP): A negative occurrence ismistakenly classified as positive.

True Negative (TN): A negative occurrence isproperly classified as negative.
False Negative (FN): A positive occurrence ismistakenly classified as negative.

V. RESULTS

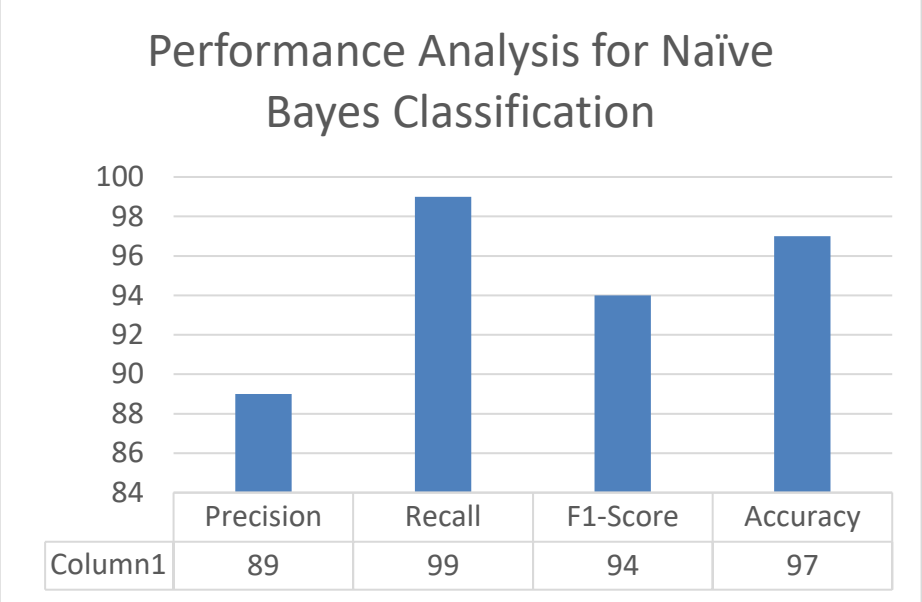

Fig. 2 Graphical Representation of EvaluationParameter

When compared to the present system, the proposed method has a greater rate of accuracy by 3%, producing results with a precision of 97%. It is very evident that it provides improved performance when identifying phishing webpages.

VI. CONCLUSION

This research demonstrates the impact that reducing the number of features in a feature set has on the detection of phishing websites. It has been demonstrated that the ant colony optimization algorithm is effective when applied to optimization issues. The purpose of the suggested system is to take use of this quality by employing the ant colony optimization technique in order to determine which traits are the most important, and then submitting that information to a Bayesian classifier in order to spot phishing scams.

REFERENCES

[1] Mayank Pandey, Vadlamani Ravi, Text and Data Mining to Detect Phishing Websites and Spam Emails, Swarm, Evolutionary, and Memetic Computing, Bijaya Ketan Panigrahi, Ponnuthurai Nagaratnam Suganthan, Swagatam Das, Shubhransu Sekhar Dash Eds., Springer International Publishing: Springer, 2013.

[2] Choon Lin Tan, Kang Leng Chiew, San Nah Sze , Phishing Webpage Detection Using Weighted URL Tokens for Identity Keywords Retrieval in 9th International Conference on Robotic, Vision, Signal Processing and Power Applications, Haidi Ibrahim, Shahid Iqbal, Soo Siang Teoh, Mohd Tafir Mustaffa Eds., Springer Singapore, 2017.

[3] Zhijun Yan, Su Liu, Tianmei Wang, Baowen Sun, Hansi Jiang, Hangzhou Yang, A Genetic Algorithm Based Model for Chinese Phishing E-commerce Websites Detection in HCI in Business, Government, and Organizations: eCommerce and Innovation, Fiona Fui-Hoon

Nah, Chuan- Hoo Tan, Springer International Publishing, 2016.

[4] Yuancheng Li, Rui Xiao, Jingang Feng, Liujun Zhao, "A semi-supervised learning approach for detection of phishing webpages," Optik-International Journal for Light and Electron Optics, vol.124, Issue 23, December 2013.

[5] Chia-Mei Chen, D.J. Guan, Qun-Kai Su, "Feature set identification for detecting suspicious URLs using Bayesian classification in social networks," Information Sciences, vol.289, December 2014.

[6] P.A. Barraclough, M.A. Hossain, M.A. Tahir, G. Sexton, N. Aslam, Intelligent phishing detection and protection scheme for online transactions, Expert Systems with Applications, vol. 40, Issue 11, September 2013.

[7] Xianghan Zheng, Zhipeng Zeng, Zheyi Chen, Yuanlong Yu, Chunming Rong, "Detecting spammers on social networks," Neurocomputing, vol. 159, pp. 27-34, July 2015.

[8] Maher Aburrous, M.A. Hossain, Keshav Dahal, Fadi Thabtah, "Intelligent Phishing Detection System for e- Banking Using Fuzzy Data Mining," Expert Systems with Applications, vol. 37, pp. 913-7921, December 2010.

[9] Gaurav Gupta, Josef Pieprzyk, "Socio-technological phishing prevention,"Information Security Technical Report, vol. 16, Issue 2, May 2011.

[10] Ram B. Basnet, Andrew H. Sung, Quingzhong Liu, "Feature Selection for Improved Phishing Detection" in Advanced Research in Applied Artificial Intelligence: Proc. of the 25th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2012, Dalian, China, June 9-12, 2012, He Jiang, Wei Ding ,Moonis Ali, Xindong Wu, Eds. Berlin: Springer, 2012.

[11] https://securelist.com/analysis/quarterly-spam- reports/74682/spam-and-phishing-in-q1-2016/