# An Ideal Approach for Detection of Phishing Attacks using Naïve Bayes Classifier

R.Priya

*Assistant Professor, Department of Information Technology, RAAK College of Engineering and Technology, Puducherry.*

***Abstract* — Phishing attack is an aberrant trick to peculate user's private information by duping them to assail via a spurious website planned to mimic and resembles as an authentic website. The user's confidential information such as username, password, and PIN number will be grabbed by the attacker and creates a fraudulent transactions. The information holder's credentials as well as money will be seized. The phishing and legitimate website will have high intelligible resemblances by which the attacker will seize the credentials of the user. Inorder to detect the phishing attacks there exists various techniques such as blacklisting, whitelisting, heuristics and machine learning. Nowadays machine learning is used and found to be more effective. The proposed system extracts the source code features, URL features and image features from the phishing website. The features that are extracted are given to the ant colony optimization algorithm to acquire the reduced features. The reduced features are again given to the Naïve Bayes classifier inorder to classify the webpage as genuine or phished.***

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

## I. Introduction

Nowadays internet plays a vital role in everyone's day to day life. Everyday the technology are growing in tremendous speed it makes the user to use it in a smarter way. As the technology grows it leaves its impact in all the fields. There exists some loopholes that are available on the internet by this it acts as a back door to attack the user. Many attacks are exhibited over the network one of them is phishing in which the attacker impersonates himself as legitimate and grabs the user credentials. To tempt the user by high visual resemblances. According to the phishing report during the first quarter of the year 2016 India ranks fifth place [11] of the top 10 countries affected by phishing attack. The users who are all unaware of these attack may fall into the trap. This paper considers source code, URL and image features of a website and selects the optimum features by using ant colony optimization and finally classify the website as phishing and non-phishing by using Bayesian classifier.

## II. Related Work

Pandey et al. [1] emphases on two major attacks that are performed in cyberspace viz phishing and spamming. For the finding of phishing attack they have extracted features from the webpages such as source code and URL inorder to identify the spam attack they have applied text and data mining approaches. For these they have collected the dataset from phishtank for forged websites and spam emails from Enron-spam corpus. The techniques used for training the classifiers are Genetic Programming, Logistic Regression, Probabilistic Neural Network, Multi-Layer Perceptron, Classification and Regression Tree.

Tan et al. [2] proposed an anti-phishing technique in which the source code of the URL are extracted such as meta, title, body tags. They have concentrated more on left hand side of the URL rather than the right hand side of the URL this is carried out because the attacker tries to impersonate the phished website as the legitimate website. The entire URL is broken into tokens in which the identity keywords are compared with yahoo search engine after that matching is performed. The original domain name and the given domain name are matched also with the country code top level domain. If the presence of webpage matches with country code top level domain it is considered as genuine webpage otherwise it is phished webpage.

Yan et al. [3] emphases on Chinese phishing E-commerce websites. The feature used for the detection of phishing are URL and web features and sequential minimal optimization algorithm. To optimize the features parameters they have used genetic algorithm. WebZIP tool is used for collecting and downloading the source code of the E- commerce webpage and Weka a data mining tool is used for training the proposed system.

Li et al. [4] proposed a machine learning approach for the detection of phishing webpages. This paper emphases on features of the webpage such as web image and document object model to optimize the features that are extracted from the webpage they have used quantum inspired evolutionary algorithm. The optimized features are passed into transductive support vector machine to classify the webpage as legitimate or phishy.

Chen et al. [5] proposed a model for suspicious URL it contains three modules. The work of data collection modules is to collect the posted URL links. The feature vectors are extracted by utilizing the feature extraction module. In classification module, they have used Bayesian classifier to detect the suspicious URL. This model does not limits to blacklist instead they emphases on URL and behavior in social links. The malicious URL are detected by domain anomaly and user behavior are identified by social anomaly.

Barraclough et al. [6] detects phishing websites using Neuro fuzzy approach. Moreover, it uses If…Then rules to differentiate phishing, suspicious, legitimate websites. Based on the severity it gives caution in the form voice alarm or color indication.

Zheng et al. [7] offered a solution for finding the spammers. The spammers and non-spammers are categorized based on content and user features. To identify the spammers they have used machine learning approach to extract the features. The extracted features are passed in support vector machine which acts as a classifier to classify the spammers and non-spammers.

Aburrous et al [8] proposed a method to classify the phishing e- banking websites which uses the techniques such as fuzzy logic and data mining algorithm. The fuzzy logic identifies the keywords that are related to phishing for this they have used c4.5, Ripper, Part, Prism, and CBA.

Gupta et al. [9] proposed a hybrid model for phishing prevention. It can be used as browser plugin. The techniques used for phishing websites is based on whitelist and blacklist. If the incoming URL matches with blacklist it blocks the URL immediately. If the incoming URL matches with whitelist the webpage is loaded. If webpage is not matches with both it directly send the link to moderator to verify whether the incoming webpage is legitimate or phishy.

Basnet et al. [10] enlightens the feature selection namely correlation, wrapper methods which detects phishing. The results were compared with genetic and greedy forward selection in machine learning for real time datasets on phishing.

## III.Proposed System

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

- Source code features
- URL features
- Image features

### A. Source Code Features

*1) **Tracking of login screen:*** This feature checks if it contains 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 the webpage 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 it looks like an original website.

*7) **Domain Registration Length:*** The phishing website 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 contains does 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 categorized according to the intensity of the colored image.

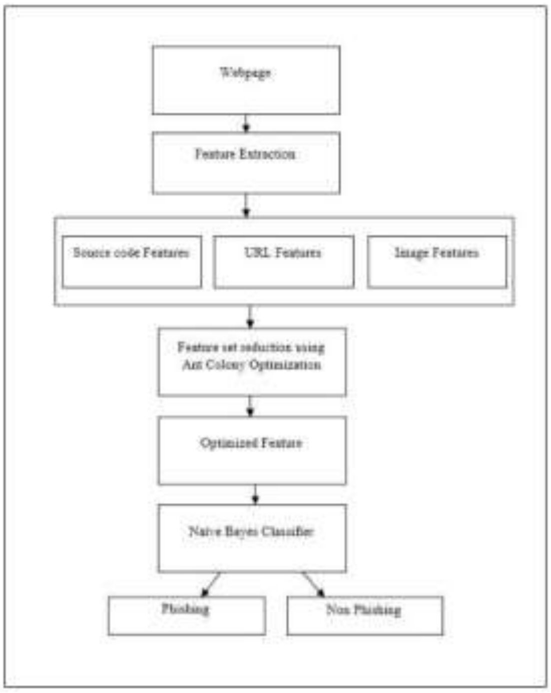

Fig. 1 Architecture of the Proposed System

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 construct complete 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 taken for consideration.

## IV. Evaluation Parameter

The prediction accuracy of the system is taken as the evaluation 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 is properly classified as positive.
False Positive (FP): A negative occurrence is mistakenly classified as positive.
True Negative (TN): A negative occurrence is properly classified as negative.
False Negative (FN): A positive occurrence is mistakenly classified as negative.

## V. Results

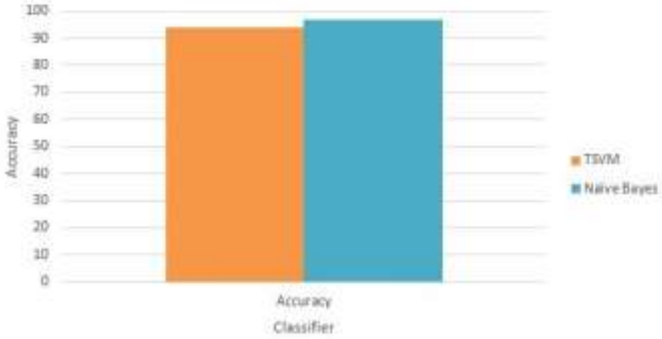

Fig. 2 Graphical Representation of Evaluation Parameter

The proposed system yields an accuracy of 97% compared to existing system its accuracy rate is higher by 3%. It is clear that it gives better performance in detecting phishing webpages.

## VI. Conclusion

This paper signifies the effect of feature set reduction for the detection of phishy webpages. Ant colony optimization algorithm has been proved to be suitable for optimization problems. The proposed system aims to take advantage of this quality and apply the ant colony optimization algorithm to identify the optimal features and then pass it to the Bayesian classifier for the detection of phishing attacks.

## 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/

## BIOGRAPHY

**R.Priya** completed her M.Tech (Information Security) degree in Computer Science and Engineering from Pondicherry Engineering College, Puducherry. She completed her B.Tech degree in Information Technology from Sri Ganesh College of Engineering and Technology, Puducherry. Currently she is working as an Assistant Professor, Department of Information Technology, RAAK College of Engineering and Technology, Puducherry. Her research interest are Information Security and Computer Networks.