Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024

DETECTIONN OF PHISHING WEBSITES THROUGH MACHINE LEARNING ALGORITHMS

[1] SAIRA BANU SHAIK , [2] MRS. L.LAKSHMI TEJASWI

[1] PG Scholar in the department of MCA at QIS College of Engineering & Technology (AUTONOMOUS), Vengamukkapalem, Ongole- 523272, Prakasam Dt., AP., India.

[2] Assistant Professor in the department of CSE/MCA at QIS College of Engineering & Technology (AUTONOMOUS), Vengamukkapalem, Ongole- 523272, Prakasam Dt., AP., India.

ABSTRACT:

Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

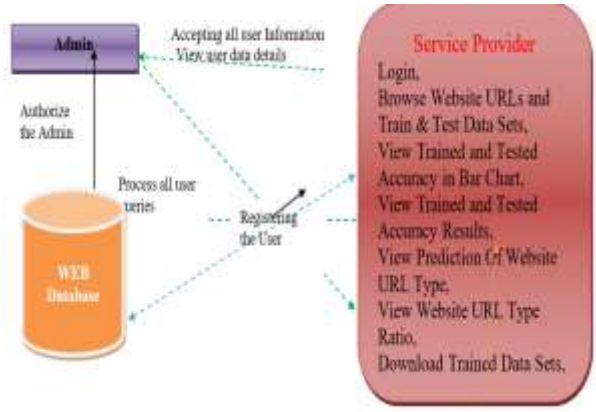
INDEX: Phishing attack, Machine learning

INTRODUCTION:

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of US$2billion per year because their clients become victim to phishing [1]. In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion [2]. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very

important to enhance phishing detection techniques. The general method to detect phishing websites by updating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is also known as “blacklist" method. To evade blacklists attackers uses creative techniques to fool users by modifying the URL to appear legitimate via obfuscation and many other simple techniques including: fast-flux, in which proxies are automatically generated to host the web-page; algorithmic generation of new URLs; etc. Major drawback of this method is that, it cannot detect zero-hour phishing attack. Heuristic based detection which includes characteristics that are found to exist in phishing attacks in reality and can detect zero-hour phishing attack, but the characteristics are not guaranteed to always exist in such attacks and false positive rate in detection is very high [3]. To overcome the drawbacks of blacklist and heuristics based method, many security researchers now focused on machine learning techniques. Machine learning technology consists of a many algorithms which requires past data to make a decision or prediction on future data. Using this technique, algorithm will analyze various blacklisted and legitimate URLs and their features to accurately detect the phishing websites including zero- hour phishing websites.

SYSTEM ARCHITECTURE



METHODOLOGY

Data set:

URLs of benign websites were collected from www.alexa.com and The URLs of phishing websites were collected from www.phishtank.com. The data set consists of total 36,711 URLs which include 17058 benign URLs and 19653 phishing URLs. Benign URLs are labelled as “0” and phishing URLs are labelled as “1”.

Table:1 The statistics of two datasets.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024

Dataset Split ratio	Classifiers	Accuracy Score	False Negative Rate	False Positive Rate
50:50	Decision Tree	96.71	3.69	2.93
	Random Forest	96.72	3.69	2.91
	Support vector machine	96.40	5.26	2.08
70:30	Decision Tree	96.80	3.43	2.99
	Random Forest	96.84	3.35	2.98
	Support vector machine	96.40	5.13	2.17
90:10	Decision Tree	97.11	3.18	2.66
	Random Forest	**97.14**	**3.14**	2.61
	Support vector machine	96.51	4.73	2.34

Confusion Matrix:

It is the most commonly used evaluation metrics in predictive analysis mainly because it is very easy to understand and it can be used to compute other essential metrics such as accuracy, recall, precision, etc. It is an NxN matrix that describes the overall performance of a model when used on some dataset, where N is the number of class labels in the classification problem.

$$Accuracy = \frac{TP + FN}{N}$$

Scikit-learn tool has been used to import Machine learning algorithms. Dataset is divided into training set and testing set in 50:50, 70:30 and 90:10 ratios respectively. Each classifier is trained using training set and testing set is used to evaluate performance of classifiers. Performance of classifiers has been evaluated by calculating classifier's accuracy score, false negative rate and false positive rate

$$Accuracy = \frac{TP + TN}{TP + FP + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Score = 2 \times \frac{precision + Recall}{Precision + Recall}$$

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 Browse Website URLs and Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Prediction Of Website URL Type, View Website URL Type Ratio, Download Trained Data Sets, View Website URL Type Ratio Results, View All Remote Users.

In future hybrid technology will be implemented to detect phishing websites more accurately, for which random forest algorithm of machine learning technology and blacklist method will be used.

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

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024

name, email, address and admin authorizes the users.

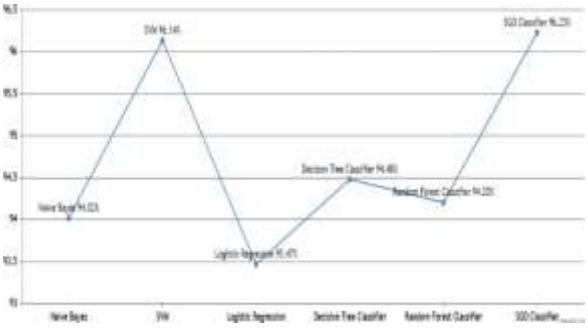
Remote User

This paper aims to enhance detection method to detect phishing websites using machine learning technology. We achieved 97.14% detection accuracy using random forest algorithm with lowest false positive rate.

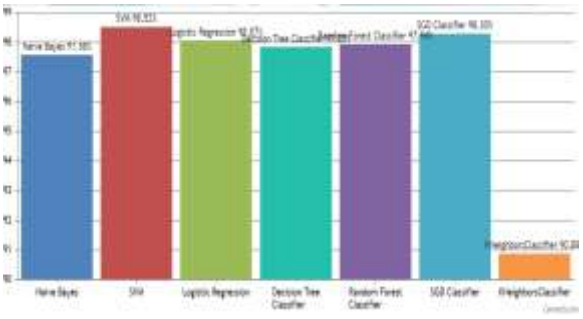
RESULT ANALYSIS

URL Type	Ratio
Phishing	70.58823529411765
Non Phishing	29.411764705882355

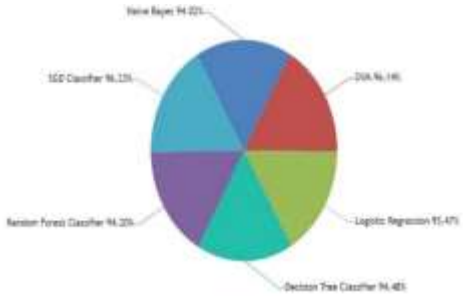
Prediction Ratio Details



Line Chart Prediction Results



Bar Chart Prediction Results



Pie Chart Prediction Results

CONCLUSION

This paper aims to enhance detection method to detect phishing websites using machine learning technology. We achieved 97.14% detection accuracy using random forest algorithm with lowest false positive rate. Also result shows that classifiers give better performance when we used more data as training data. In future hybrid technology will be implemented to detect phishing websites more accurately, for which random forest algorithm of machine learning technology and blacklist method will be used.

FUTURE ENHANCEMENT

The future of developing transformer-based models for detecting SMS spam holds tremendous potential for advancing the security and reliability of SMS communication platforms. One promising direction is the exploration of multimodal transformer architectures capable of

integrating diverse data modalities, such as text, images, and metadata, to enhance the accuracy and robustness of spam detection. By leveraging multimodal information, these models can capture additional contextual cues and semantic relationships, enabling more comprehensive analysis of SMS messages and improving detection performance. Furthermore, the integration of reinforcement learning techniques into transformer-based models offers exciting opportunities for dynamic adaptation and optimization of spam detection strategies. By continuously learning from user feedback and evolving spam patterns, reinforcement learning-based transformer models can adapt their detection policies in real-time, thereby enhancing their effectiveness in combating emerging spam threats and ensuring timely mitigation. Additionally, the future of transformer-based models for SMS spam detection lies in their application in diverse linguistic contexts and cultural settings. Multilingual transformer architectures, capable of processing and understanding SMS messages in multiple languages, hold promise for addressing the global nature of spam and enabling effective detection across linguistic barriers. Moreover, the development of domain-specific transformer models trained on domain-specific SMS datasets, such as financial transactions or healthcare communications, can further improve the accuracy and relevance of spam detection in specialized contexts. By tailoring transformer-based models to specific linguistic and domain-specific requirements, researchers can unlock new avenues for enhancing the security and trustworthiness of SMS communication platforms, ensuring users' privacy and safety in diverse communication environments.

REFERENCES

1) Gunter Ollmann, “The Phishing Guide Understanding & Preventing Phishing Attacks”, IBMInternet Security Systems, 2007.
2) https://resources.infosecinstitute.com/category/enterprise /phishing/the-phishing-landscape/phishing-data-attackstatistics/#gref
3) Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013
4) Mohammad R., Thabtah F. McCluskey L., (2015) Phishing websites dataset. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites Accessed January 2016

5) http://dataaspirant.com/2017/01/30/how-decision-treealgorithm-works/
6) http://dataaspirant.com/2017/05/22/random-forestalgorithm-machine-learing/
7) https://www.kdnuggets.com/2016/07/support-vectormachines-simple-explanation.html
8) www.alexa.com
9) www.phishtank.co

AUTHORPROFILE:

Mrs. L. Laskhmi Tejaswi currently working as an Assistant Professor in the Department of Computer Science and Engineering, QIS College of Engineering and Technology, Ongole, Andhra Pradesh. She did her BTech from Rao& Naidu Engineering college JNTUK, Kakinada, M.Tech from Qis College Of Engineering college And Technology JNTUK, Kakinada. Her area of interest is Machine Learning, Artificial intelligence, Cloud Computing and Programming Languages.

Ms. Saira Banu Shaik, currently pursuing Master of Computer Applications at QIS College of engineering and Technology (Autonomous), Ongole, Andhra Pradesh. She Completed BCA in from Sri Nagarjuna Degree College, Ongole, Andhra Pradesh. Her areas of interest are Machine learning & Cloud computing.