International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

MACHINE LEARNING TECHNIQUES TO SEEK OUT MALICIOUS WEBSITES

Ankush Kulkarni[1], Nazmeen Khan[2], Srushti Kombde[3] ,Vidya A Nemade[4]

[1,2,3]*Student, Department of Computer Engineering, LES G.V. Acharya Institute of Engineering and Technology, Shelu, Maharashtra*

[4]*Asst. Professor and H.O.D, Department of Computer Engineering, GVAIET, Shelu, Maharashtra*

Abstract - *Detection of phishing websites are an awfully important precautions for several of the net platforms. Phishing of an internet site could also be a quite attack where the hacker generates a replica website to buffoons the users into submitting personal, financial or password data to what they think is their service providers website during this paper, we design and implement an intelligent model for detecting phishing websites. we've also discovered various techniques for detection and prevention of phishing. aside from that, we've introduced a replacement model for detection and prevention of phishing attacks. during this system we are using ensemble model with random forest, logistic regression and support vector machine which can provide accurate result whether the web site is legitimate or not. Ensemble model will find more accurate result than any of above mention algorithm.*

***Key Words*: Ensemble Algorithm, Logistic Algorithm, Support Vector Machine, Random Forest Classification, Phishing.**

1. INTRODUCTION

Phishing could also be a continuing threat, and thus the danger is even larger in social media like Facebook, Twitter, and bank websites. Hackers could create a similar to an internet site and tell you to enter personal information, which is then emailed to them .Hackers commonly cash in of those sites to attack folks that are in homes, or publicly so on to require personal and security data which can severely affect the user. Phishing takes benefits of the religion that the user may have with the web site or the corporate since the user might not be ready to tell that the location being visited or URLs getting used isn't real. Therefore, when this happens, the hacker has the probable chance to realize the private information

1.1 AIM OF THE PROJECT

The aim of the project is to make classifiers such as Random forest, logistic regression, support vector machine and ensemble learning technique by reviewing and refining parameters supported websites attribute.

1.2 The Objectives are:

- To spot a wide-ranging data supported a various data source.
- To spot appropriate set of parameters so algorithm can solve a given problem.
- Coach and validate the phishing URLs detection models in real-time environment.
- Provide a comparable study to show the effectiveness and capabilities of the model.

2. RELATED WORK

This section will attempt to illuminate the related works and research published recently also as in reference to concepts mentioned during this paper. Many researchers had applied their statistics to research the phishing URLs. Some research had motivated us to stabilized our approach in project.

Paper published by Ma et al [1,2] compared many batch-based learning algorithms for classifying phishing website and showed that the combination of host-based and lexical features results in the highest classification accuracy. They also compared the performance of batch-based algorithms to online algorithms using features and found that online algorithms outperform batch-based algorithms.

The work which was published by Garera et al [3] uses logistic algorithm over some selected features to classify phishing website. The features include the presence of keywords in the website features based on Google's Page Rank.

Classifier was not constructed by McGrath and Gupta[4] but performs analysis of phishing and non-phishing URLs with respect to datasets. They compared phishing URLs drawn from the Phish Tank[8] to non-phishing URLs from DMOZ Open Directory Project[9].

S.Parekh, D. Parikh, S. Kotak, and P. S. Sankhe[5] proposed a model with solution for detecting phishing sites by implementing website identification strategy using Random Forest algorithm. Accuracy obtained is of level 95.2% which shown by three stages, namely Parsing, Heuristic Classification of data, Performance Random forest method.

W. Fadheel, M. Abusharkh, and I. Abdel-Qader[6] proposed selection model to detect phishing URLs. They used Logistic Regression and Support Vector Machine as methods to validate the feature selection methodology. They also showed that SVM algorithm achieved best performance over LR algorithm.

L. MacHado and J. Gadge[7] proposes a way to detect phishing websites by making use of c4.5 decision tree. This technique extracts features and calculates heuristic values from the websites. These values were given to the c4.5 decision tree algorithm which determines whether the site is non-phishing or not. Dataset was collected from Phish Tank and Google.

3. METHODOLOGY

URLs are the one through which user try to access the internet and websites. URLs are commonly known as "Web Links". Our goal is to derive such classification model that will help us to detect the phishing and non-phishing website by analysis of lexical as well as by using some URLs features.

Earlier papers had used many algorithms to detect whether the website is legitimate or not but accuracy is still the question for prediction. For prediction of phishing website Random forest algorithm stands great as the prediction is 95.2% accurate so in this paper, we had used support vector machine, random forest algorithm and logistic regression and ensemble learning technique so that maximum accuracy must be yield out for prediction of phishing website

We have collected features of URLs of some websites from www.alexa.com. The phishing URLs features were collected from www.phishtak.com the info set consists of 19000 phishing URLs.
Below is system architecture of our system

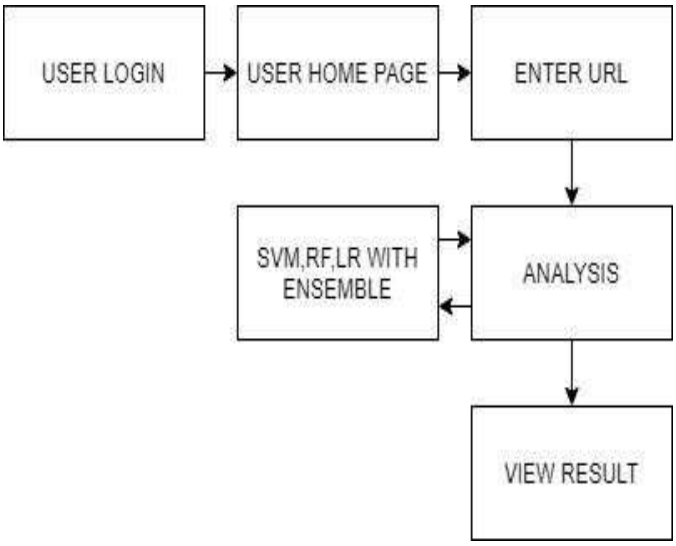

Fig 3.Sytem Architecture

3.1 Lexical feature analysis

Lexical features are nothing but textual property of the URL. URLs are human-readable text strings that are Parsed and converted in such a typical way by client programs. Browser have potential to translate each URL into the instruction that server hosting site through multistep resolution protocol.

3.2 URL Features

We had found some necessary features which can facilitate help us to sight and predict the website which are phishing and which aren't. Address bar-based features wherever Website or getting to be detected phishing using IP address, Tiny URL, sub Domain and multi sub domains. Abnormal Based features and domain-based features.

3.3 MACHINE LEARNING ALGORITHM

The four-machine learning algorithm that are used to analysis the features of URLs are as follows

1. Support Vector Machine: The SVM performs classification by finding the hyper plane that maximizes the margin between two classes. The vectors that outlines the hyper plane are the support vectors

2. Random Forest Algorithm: Random forest algorithm is a supervised classification algorithm. Because the name suggests, this algorithm creates the forest with a range of trees and highest number of trees denote a lot of accuracy in the prediction.

3. Logistic Regression: Logistic regression is a classification algorithm used to assign observations to a separate set of classes. A number of the samples of classification problems are Email spam or not spam, Online transactions Fraud or not Fraud.

4. Ensemble Learning: Ensemble learning uses multiple machine learning models to undertake to create higher predictions on a dataset. An ensemble model works by coaching totally different models on a dataset and having every model build prediction one by one. The predictions of those models are then combined within the ensemble model to form a final prediction.

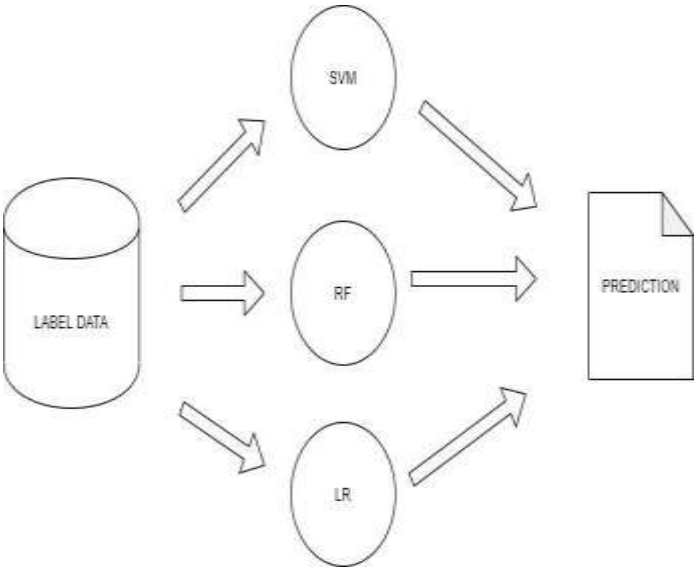

Figure 3.3 Ensemble Technique

4. SYSTEM FLOW CHART

Following is Flow chart of system will show extraction of features where the feature set comprises of host length, path length, number of slashes, number of path tokens. Flowchart shown below states many different functions that exist in the proposed system. For example, the user login, administrator login etc

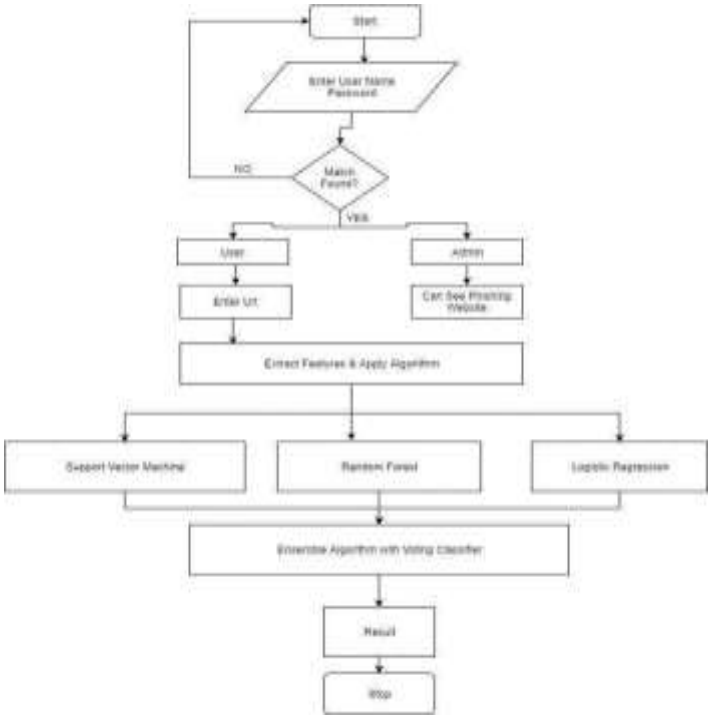

Figure 4 Sytem Flow Chart

5. CONCLUSIONS

We know that phishing is associate attack that aims at destroying weaknesses found throughout electronic communications like user unseaworthy their passwords to any unknown random websites. Hence awareness and defense both are required against these sites. Our projected system model will take the webpage through various levels of detection and user of this technique can prove beneficial for detecting a phishing website

In this particular domain challenge is that criminal is constantly changing and making strategies against our system so to overcome such strategies we need to adapt more and more algorithm, techniques and features for URLs.

Our project combination of SVM, RF and LR with provide accuracy, precision etc.

REFERENCES

[1] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker," on the far side Blacklists: Learning to sight Phishing Websites from malicious URLs", Proc.of SIGKDD2009.

[2] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker," Learning to sight Phishing website", ACM Transactions on Intelligent Systems and Technology, Vol. 2, No. 3, Article 30, Publication on April 2011

[3] Garera S., Provos N., Chew M., Rubin A. D., "A Framework for Detection and mensuration of phishing attacks", In Proceedings of the ACM Workshop on Rapid Malloced, Alexandria, VA.

[4] D. K. McGrath, M. Gupta, "Behind Phishing: An Examination of Phisher", In forwarding of the USENIX Workshops on Big-Scale Exploits and Emerging Threats.

[5] S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, "A New Methodology for Detection of Phishing Websites: URL Detection," in 2018 Second International Conference on ingenious Communication and Computational Technologies (ICICCT), 2018, vol. 0, no. ICICCT, pages. 949–952.

[6] W. Fadheel, M. Abusharkh, and I. Abdel-Qader, "On Feature choice for the Prediction of Phishing Websites," 2017 IEEE fifteenth International Conference Dependable, Authentication. Security. Computer. fifteenth Intl Conference Pervasive Intelligent. Computer. third International Conference Big Data Intelligent. Computer. Cyber Science. Technology. pp. 871–876, 2017.

[7] L. MacHado and J. Gadge, "Phishing Sites Detection Based on C4.5 Decision Tree Algorithm," in 2017 International Conference on Computing, Communication, management and Automation, published ICCUBEA 2017, 2018, pages. 1–5.

[8] Phish Tank.:http://www.phishtank.com.

[9] DMOZ Open Directory Project: http://www.dmoz.org.

BIOGRAPHIES

Mr. Ankush Kulkarni, Final Year Student of B.E (Computer Engineering) at G.V Acharya Institute of Engineering and Technology, Shelu, Maharashtra. Domain Of interest – Database, Data Structure and Algorithm, Machine learning, Data science.

Ms. Nazmeen Khan, Final Year Student of B.E (Computer Engineering) at G.V Acharya Institute of Engineering and Technology, Shelu, Maharashtra. Domain Of interest –Artificial Intelligence, Machine learning, IOT.

Ms. Srushti Kombde, Final Year Student of B.E (Computer Engineering) at G.V Acharya Institute of Engineering and Technology, Shelu, Maharashtra. Domain Of interest – Machine learning, Artificial Intelligence

Ms. Vidya A Nemade (M.E Computer Science and Engineering), Asst. Professor and Head of Department (H.O.D) at G.V Acharya Institute of Engineering and Technology, Shelu, Maharashtra. Domain Of interest - Image Processing, Database, Computer Networks, Compliers.