VIVA-Tech International Journal for Research and Innovation
ISSN(Online): 2581-7280
Volume 1, Issue 4 (2021)
Article No. X
PP XX-XX

VIVA Institute of Technology
9th National Conference on Role of Engineers in Nation Building – 2021 (NCRENB-2021)

VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL

VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

A Hybrid Approach For Phishing Website Detection Using Machine Learning.

Harsh Kansagara[1], Vandan Raval[2] ,Faiz Shaikh[3], Prof. Saniket Kudoo[4]
[1](*Department of Computer Engineering, Mumbai University, India)*
Email: 17301081harsh@viva-technology.org
[2](*Department of Computer Engineering, Mumbai University, India)*
Email: 17301044vandan@viva-technology.org
[3](*Department of Computer Engineering, Mumbai University, India)*
Email:17305018faiz@viva-technology.org
[4](*Department of Computer Engineering, Mumbai University, India)*
Email: saniketkudoo@viva-technology.org

Abstract: *In this technical age there are many ways where an attacker can get access to people's sensitive information illegitimately. One of the ways is Phishing, Phishing is an activity of misleading people into giving their sensitive information on fraud websites that lookalike to the real website. The phishers aim is to steal personal information, bank details etc. Day by day it's getting more and more risky to enter your personal information on websites fearing that it might be a phishing attack and can steal your sensitive information. That's why phishing website detection is necessary to alert the user and block the website. An automated detection of phishing attack is necessary one of which is machine learning. Machine Learning is one of the efficient techniques to detect phishing attack as it removes drawback of existing approaches. Efficient machine learning model with content based approach proves very effective to detect phishing websites.*

Our proposed system uses Hybrid approach which combines machine learning based method and content based method. The URL based features will be extracted and passed to machine learning model and in content based approach, TF-IDF algorithm will detect a phishing website by using the top keywords of a web page. This hybrid approach is used to achieve highly efficient result. Finally, our system will notify and alert user if the website is Phishing or Legitimate.

Keywords- *Content-based approach, Machine learning, Phishing detection, Random Forest, TF-IDF.*

I. INTRODUCTION

In this technical era, the people are interconnected with each other by means of internet with the help of the electronic devices like computers, laptops. One of the major cyber threats now is Phishing, where the attackers use illegitimate websites to obtain victims credentials and use it .Most phishing attacks work by creating a phony version of the real site's web interface to achieve the user's trust. Even for cautious users it's sometimes difficult to detect phishing attack. Because the attackers attempt to gain your trust by using the same user interface, almost same URL and cloned websites.

Machine learning is a system that provides the ability to automatically learn and improve from its experience without being overtly programmed. In our algorithm the process of learning begins from with training dataset in order to look for patterns in data and make better decisions in the future. The main benefit of the algorithm is to allow the system to learn and decide automatically whether the website is phishy or legitimate. In terms of machine learning larger number of data will increase the accuracy of the model significantly. So the prediction of phishing websites will deal with larger dataset to train the model for better accuracy.

Random forest is a machine learning algorithm which is used for both classification as well as regression. But still, it is mostly used for classification problems. We know that a forest is made up of many trees and more trees means more robust forest. It is an ensemble method which has many trees so it is better than a single

decision tree. Ensembles are a method to improve performance using divide-and-conquer approach. The main principle in ensemble methods is that we can form strong learner when a group of weak learners come together. Hence, we will use Random Forest algorithm for the classification of URL features.

There are various types of Phishing attacks which are been used by the attackers for various domains for different purpose. Since phishing is such a widespread problem in the cyber-security domain, there is a necessity of phishing website detection.

II. RELATED WORKS

Peng yang,Guangzhen Zhao, Peng zen [1] have proposed a multidimensional feature phishing detection approach based using deep learning.They use character sequence features of the given URL for quick classification by deep learning. Then they combine URL statistical features, webpage code features, webpage text features, and the quick classification result of deep learning into multidimensional features. The deep learning algorithms they used to train and test were CNN-LSTM algorithm. The classifier used in the multidimensional feature algorithm is the XGBoost (eXtreme Gradient Boosting) ensemble learning algorithm, which has high classification accuracy. The approach can limit the detection time for setting a threshold. They test on a dataset containing millions of phishing URLs and legitimate URLs. By reasonably altering the threshold, the experimental outcomes show that the detection efficiency can be improved.

Mohammed HazimAlkawaz, Stephanie Joanne Steven, Asif Iqbal Hajamydeen [2] has designed a software to show awareness of the extensive level of its functionality, features that can be displayed in the monitoring era. The system fosters many features in comparison of other software. Its unique features such as capturing blacklisted URL's from the browser directly to verify the validity of the website, notifying user on blacklisted websites while they are trying to access through pop-up, and also notifying through email. This system will assist user to be alert when they are trying to access a blacklisted website.

Huaping Yuan, Xu Chen, Yukun Li, Zhenguo Yang, Wenyin Liu [3] they propose to extract features from URLs and webpage links to detect phishing websites and their targets. Moreover to the basic features of a particular URL, such as suspicious characters, length, number of dots, a feature matrix is
also built from these basic features of the links in the given URL's webpage. Furthermore, they extract certain statistical features from each column of the feature matrix, such as mean, median, and variance. Lexical features are also removed from the given URL, the links and content in the webpage, such as title and textual content. A number of ML models have been explored for phishing detection, among which Deep Forest model which shows competitive performance, achieving a true positive rate of 98.3% and a false alarm rate of 2.6%. In particular, they aimed an effective strategy based on search operator via search engines to find the phishing targets, which attains an accuracy of 93.98%.

F.C. Dalgic, A.S. Bozkir and M. Aydos [4] have proposed vision based brand prediction approach that aimed to build a phishing web page detection and recognition system that is scalable and robust. Hence, use of numerous MPEG-7 and MPEG-7 like solid colour descriptors have been proposed. Furthermore, along with being invariant to input image size, they also recommended the extraction and utilization of color based information via coarse- to-fine multi-level spatial patch pyramid. The recommended approach presents a trivial schema serving competitive accuracy and higher feature extraction and inferring speed that can be used as a browser plugin or mobile device phishing protector.

SrushtiPatil, SudhirDhage [5] have reviewed various anti-phishing approaches. All methods are debated to give a clear conscience of already existing techniques, their limitations and possible improvements. Next they analyzed the in-use anti phishing tools available for free. Next they described the most important steps to build an efficient anti-phishing model with the help of architecture diagram. Finally they compared the models using all the 5 types of approaches based on the number of features used, accuracy and size of dataset.

S. Parekh, D. Parikh, S. Kotak and S. Sankhe [6] they have created a solution to detect phishing websites by using the URL detection method using Random Forest algorithm. They have used 3 major phases such as Parsing, Heuristic Classification of data, Performance Analysis in their model and each phase makes use of a different algorithm or technique for processing of data to give better results. They also used Rstudio for better analysis. The dataset used in this was from Phish Tank and it was spilt into 70% and 30%. The 70% data was considered for training and 30% for testing. They managed to get 95% accuracy and thus Random Forests was chosen for classification.

Jian Mao , Wenqian Tian, Pei Li , Tao Wei, and Zhenkai Liang [7] have propose a robust phishing detection approach, Phishing-Alarm, based on CSS features of web pages. They identify CSS features using different methods, as well as algorithms to efficiently evaluate page similarity. They prototyped Phishing-Alarm as an extension to the Google Chrome browser and demonstrated its effectiveness in evaluation using real-world phishing samples.

T. Nathezhtha, D. Sangeetha and V. Vaidehi [8] have proposed an attack detection method named as Web Crawler based Phishing AttackDetector (WC-PAD) which is a three step method. It takes the web traffics, URL as input features, based on the features classification it classifies website as phishing and non-phishing. Firstly, the DNS Blacklist based detection is done. Secondly a Web Crawler based detection is accomplished followed by Heuristics based detection. The trial analysis of the proposed WC-PAD is done with datasets contain actual phishing cases. The implementation results give 98.9% accuracy for the proposed approach in both phishing and zero-day phishing attack detection.

Altyeb Altaher [9] have proposed a hybrid approach for classifying the websites as Phishing, Legitimate or Suspicious, the proposed approach combines the K-nearest neighbors (KNN) algorithm with the support vector machine algorithm (SVM). The proposed approach combines the effectiveness and simplicity of KNN with SVM. Thus, the proposed approach gains the advantages of combining KNN with SVM to avoid the drawbacks when they are used separately. The experimental results give an accuracy of 90.04% for the proposed KNN with SVM approach.

S. Haruta, H. Asahina and I. Sasase [10] have proposed an approach of visual similarity-based phishing detection scheme using CSS and image. They focus on the fact that attackers often steal legitimate websites' CSS to mimic the legitimate websites. They regarded the websites which have hyperlinks from other websites as legitimate. By detecting the website which has similar appearance to legitimate websites or plagiarizes CSS of legitimate website, they detected phishing website and its target simultaneously. Moreover, by using CSS, they alleviated the shortcoming of locally different images which causes false negative in the conventional scheme. By the computer simulation, they showed their scheme achieves about 80% of detection accuracy and finds 72.1 % famous phishing target.

Christou, O. Pitropakis, N. Papadopoulos, P. McKeown, S. and Buchanan [11] have proposed a system where malicious domain name datasets were used to make predictions on their nature using the Random Forests and SVM algorithms. It uses an an automated filtration system that provide's a list of possibly malicious URLs, narrowing the list down and reducing the human input required to spot them,Machine Learning techniques use features extracted from the URLs and their DNS data to analyse and detect whether they are malicious or legitimate.

S. Roopak, A. P. Vijayaraghavan and T. Thomas [12] have extracted the relevant rules based on webpage source code and Secure Socket Layering (SSL) based features from a training dataset using Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm. Further, they check for the presence of these rules in a test dataset. Their experimental results show that this approach can identify phishing websites with an accuracy of 0.92%.

H. Yuan, X. Chen, Y. Li, Z. Yang and W. Liu [13] they propose to extract features from URLs and webpage links to detect phishing websites and their targets. Furthermore, they extract certain statistical features from each column of the feature matrix, such as mean, median, and variance. Lexical features are too extracted from the given URL. Furthermore, the content and links in its webpage such as textual title and content were also extracted. A number of machine learning models were explored for phishing detection, among which Deep Forest ML model which showed competitive performance, achieving a true positive rate of 98.3% and a false alarm rate of 2.6%. They designed an effective strategy based on search operators to find phishing URLs, which achieved an accuracy of 93.98%.

M. Sameen, K. Han and S. O. Hwang [14] they have designed an ensemble machine learning-based detection system called PhishHaven. It helps to identify AI-generated as well as human-generated phishing URLs. Their technique uses lexical analysis for feature extraction. To further enhance lexical analysis, they introduced 3 classification procedures. Firstly, they have URL HTML Encoding which helps to classify URL on-the-fly and proactively compare with some of the existing methods. Secondly, a URL Hit approach to deal with tiny URLs, which is an open problem yet to be solved. Lastly, the final classification of URLs is made on an unbiased voting mechanism in PhishHaven, which aims to avoid misclassification when the number of votes is equal. They have used a benchmark dataset of 100,000 phishing and normal URLs achieving 98.00% accuracy outperforming existing systems.

H. Chapla, R. Kotak and M. Joiser [15] in this paper they designed a framework of phishing detection using URL and Fuzzy Logic as Classifier. They used MatLab tool to code the program which can extract the features from the entered URL. They have different feature sets, out of those they extracted some features based on URL of website.They used 11 features for classification of phishing websites. The fuzzy classifier implemented using MatLab in this paper achieved an accuracy of 91.46%.

III. METHODOLOGY

Phishing websites can easily trick us into submitting private/financial information into their system dealing us a huge personal/financial loss. The aim of our project is to successfully classify websites as phishing websites or legitimate websites.

The system which we are going to use will be Random Forest Algorithm and TF-IDF approach for detecting phishing websites. Few websites can easily be classified as phishing just by looking at it, while few need a deeper analysis. We will carefully extract certain features from URLs which will be used for training and testing inputs. These inputs will be divided into four arrays namely training input, training output, testing input, and testing output. Training input array will be used for training our Random Forest Classifier. We'll then use testing input to test our classifier.

TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistical method which evaluates the significance of a particular word in a document. It is an information retrieval technique that weighs a term's frequency (TF) and its inverse document frequency (IDF). Each word or term will have its own unique TF and IDF scores. The product of TF and IDF scores of a term is called the TF-IDF weight of that term. We will calculate the TF-IDF scores of the words appearing in the webpage. The frequently occurring words or the words with the highest TDF-DF scores will be passed as a query in the search engine for results.

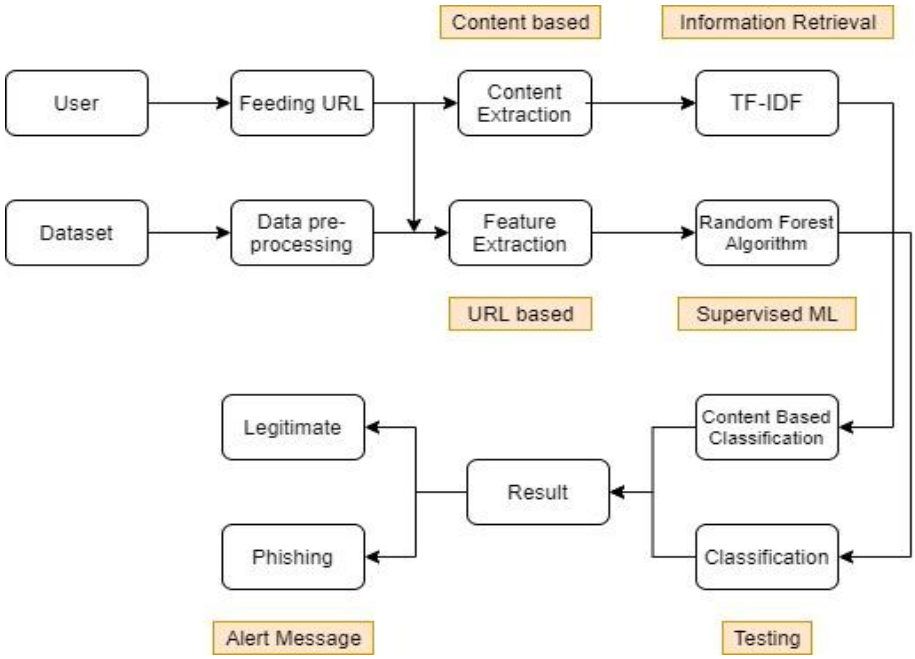

Figure 1: System Flow Diagram

IV. RESULT

This chapter provides the partial implementation and the screenshots of the results. For now, we have created the content based side of our project. The content based side will work in such a way that when we feed a URL into the program, it will scrape that webpage and find out the domain name of that URL, the title of that webpage and top 3 most frequently used words from the webpage. All of these things combined will make a query. This query will be fed into the search engine and will retrieve the top 10 URL's that pop up. Queries will be calculated for each of these 10 URL's and then will be compared to the original query of the URL which we gave the program. If the query matches, then the webpage can be considered as legitimate otherwise phishing.

V. CONCLUSION

This system tries to make a safe environment for browsing websites by detecting phishing websites keep the user safe. Or else, the user might end up giving his credentials to the phisher's which can lead to huge losses. This project also aims to implement the detection of the phishing websites using Machine Learning. This task will be done by using hybrid approach extracting the features of the website and notify the users that the website is legitimate of phishing. This system aims to achieve high detection performance.

Acknowledgements

We would like to express a deep sense of gratitude towards our guide Prof.Saniket Kudoo, Computer Engineering Department for his constant encouragement and valuable suggestions. The work that we were able to present is possible because of his timely guidance We would like to pay gratitude to the panel of examiners Prof. Sunita Naik, Dr. Tatwadarshi P. N., Prof. Dnyaneshwar Bhabad, & Prof. Vinit Raut for their time, effort they put to evaluate our work and their valuable suggestions time to time. We would like to thank Project Head of the Computer Engineering Department, Prof. Janhavi Sangoi for her support and co-ordination. We would like to thank Head of the Computer Engineering Department, Prof. Ashwini Save for her support and coordination. We are also grateful to teaching and non-teaching staff of Computer Engineering Department who lend their helping hands in providing continuous support.

REFERENCES

[1] P. Yang, G. Zhao and P. Zeng, "Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning" *IEEE Access, vol. 7,* 2019, pp. 15196-15209.
[2] M. H. Alkawaz, S. J. Steven and A. I. Hajamydeen, "Detecting Phishing Website Using Machine Learning," *IEEE International Colloquium on Signal Processing & Its Applications (CSPA),*, 2020, pp. 111-114.
[3] H. Yuan, X. Chen, Y. Li, Z. Yang and W. Liu, "Detecting Phishing Websites and Targets Based on URLs and Webpage Links," *International Conference on Pattern Recognition (ICPR),* 2018, pp. 3669-3674.
[4] F. C. Dalgic, A. S. Bozkir and M. Aydos, "Phish-IRIS: A New Approach for Vision Based Brand Prediction of Phishing Web Pages via Compact Visual Descriptors," *International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2018, pp. 1-8.
[5] S. Patil and S. Dhage, "A Methodical Overview on Phishing Detection along with an Organized Way to Construct an Anti-Phishing Framework," *International Conference on Advanced Computing & Communication Systems (ICACCS)*, 2019,pp. 588-593.
[6] S. Parekh, D. Parikh, S. Kotak and S. Sankhe, "A New Method for Detection of Phishing Websites: URL Detection," *International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018, pp. 949-952.
[7] J. Mao, W. Tian, P. Li, T. Wei and Z. Liang, "Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity," *IEEE Access, vol.5*, 2017, pp. 17020-17030.
[8] T. Nathezhtha, D. Sangeetha and V. Vaidehi, "WC-PAD: Web Crawling based Phishing Attack Detection," *International Carnahan Conference on Security Technology (ICCST),* 2019, pp. 1-6.
[9] Taha, Altyeb. "Phishing Websites Classification using Hybrid SVM and KNN Approach" *International Journal of Advanced Computer Science and Applications , Volume 8 Issue 6,* 2017.
[10] S. Haruta, H. Asahina and I. Sasase, "Visual Similarity-Based Phishing Detection Scheme Using Image and CSS with Target Website Finder," *IEEE Global Communications Conference*, 2017, pp. 1-6.
[11] Christou, O.; Pitropakis, N.; Papadopoulos, P.; McKeown, S. and Buchanan, "Phishing URL Detection Through Top-level Domain Analysis: A Descriptive Approach", *International Conference on Information Systems Security and Privacy, Volume 1: ICISSP*, 2020, pp. 289-298.
[12] S. Roopak, A. P. Vijayaraghavan and T. Thomas, "On Effectiveness of Source Code and SSL Based Features for Phishing Website Detection," *International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*, 2019, pp. 172-175.
[13] H. Yuan, X. Chen, Y. Li, Z. Yang and W. Liu, "Detecting Phishing Websites and Targets Based on URLs and Webpage Links," *International Conference on Pattern Recognition (ICPR),* 2018, pp. 3669-3674.

[14] M. Sameen, K. Han and S. O. Hwang, "PhishHaven—An Efficient Real-Time AI Phishing URLs Detection System," *IEEE Access, vol. 8*, 2020, pp. 83425-83443.
[15] H. Chapla, R. Kotak and M. Joiser, "A Machine Learning Approach for URL Based Web Phishing Using Fuzzy Logic as Classifier," *International Conference on Communication and Electronics Systems* , 2019, pp. 383-388.