JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

"STUDY OF FAKE WEBSITES DETECTION USING MACHINE LEARNING"

Shubham Jadhav[1], Mr. S. K. Patil[2], Ms. M. A. Waghmode[3], Rushikesh Kadam[4], Aakash Nirhali[5],
[2,3]Assistant Professor,Dept. of E&TC Engg., SKNCOE,SPPU, Pune.
[1,4,5]UG Student,Dept. of E&TC Engg., SKNCOE,SPPU, Pune.

[1]jadhavshubham2024@gmail.com

[2]skpatil_skncoe@sinhgad.edu

[3]meghali.waghmode_skncoe@sinhgad.edu

[4]rushikadam311998@gmail.com

[5]aakashnirhali076@gmail.com

Abstract:

Phishing is the most straight forward method of obtaining sensitive information from unsuspecting consumers. The goal of phishers is to obtain sensitive information such as usernames, passwords, and bank account numbers. People working in cyber security are now looking for reliable and consistent detection strategies for phishing websites. The purpose of this work is to use machine learning to detect phishing URLs by extracting and evaluating various aspects of authentic and phishing URLs. Phishing websites are detected using Decision Tree, Random Forest, and Support Vector Machine algorithms. The goal of the study is to detect phishing URLs and to narrow down the best machine learning method by analysing each algorithm's accuracy rate, false positive rate, and false negative rate. Phishing, Feature Classification, Random Forest Classifier, and other terms are used in this article.

Keywords: Phishing, Extreme Learning Machine, Feature Classification, etc.

(I)Introduction

Technology is advancing at a breakneck pace, and the internet has become an indispensable part of people's daily lives as a result. Because of the rapid advancement of technology and the widespread use of digital systems, internet use has increased, and data security has become increasingly important. The basic goal of information technology security is to ensure that adequate precautions are taken against threats and dangers that may be encountered by users when using these technologies. Phishing is the deception of a trustworthy person in an electronic connection in order to get sensitive information such as usernames, passwords, and credit card numbers. It's usually done through email spoofing or instant messaging, and it often urges consumers to enter personal information on a bogus website that looks and feels exactly like the real one. Information security dangers have been observed and developed over time as the internet and information systems have evolved. The impact is a breach of information security due

to the compromise of private data, with the victim potentially losing money or other assets as a result. Internet users are vulnerable to a variety of cyber risks, including the theft of personal information, identity theft, and financial losses. As a result, internet use in the home and at work may be questionable. To lessen security vulnerabilities, users must be able to identify and defend against privacy leakage using effective analytical tools. At the time of an attack, effective systems that can improve self-intervention must be built utilizing an artificial intelligence based information security management system.

Phishing is the most straightforward method of obtaining sensitive information from unsuspecting consumers. The goal of phishers is to obtain sensitive information such as usernames, passwords, and bank account numbers. People working in cyber security are now looking for reliable and consistent detection strategies for phishing websites. The purpose of this work is to use machine learning to detect phishing URLs by extracting and evaluating various aspects of authentic and phishing URLs. Phishing websites are detected using Decision Tree, Random Forest, and Support Vector Machine algorithms. The goal of the study is to detect phishing URLs and to narrow down the best machine learning method by analyzing each algorithm's accuracy rate, false positive rate, and false negative rate. Phishing, Feature Classification, Random Forest Classifier, and other terms are used in this article.

(II)Literature Survey

1. Detecting Phishing Websites Using Machine Learning Amani Alswailem Bashayr Alabdullah Norah Alrumayh Dr.Aram Alsedrani 2019 IEEE The system is based on a machine learning method, particularly supervised learning. Here is selected the Random Forest technique due to its good performance in classification. The focus is to pursue a higher performance classifier by studying the features of phishing websites and choose the better combination of them to train the classifier. As a result, the conclusion is the paper is with accuracy of 98.8

2. A Machine-Learning Framework for Supporting Intelligent Web-Phishing Detection and Analysis Alfredo Cuzzocrea, Fabio Martinelli, Francesco Mercaldo 2019 ACM In particular the system makes use of state-of-the-art decision tree algorithms for detecting whether a Web site is able to perform phishing activities. If this is the case, the Web site is classified as a Web-phishing site. Experimental evaluation confirms the benefits of applying machine learning methods to the well-known web-phishing detection problem.

3. Phishing Web Sites Features Classification Based on Extreme Learning Machine. Yasin S¨onmez1 T¨urker Tuncer2 H¨useyin G¨okal 3 Engin Avc4 2018 IEEE The purpose of this study is to perform Extreme Learning Machine (ELM) based classification for 30 features including Phishing Websites Data in UC Irvine Machine Learning Repository database. For results assessment, ELM was compared with other machine learning methods such as Support Vector Machine (SVM), Na¨ıve Bayes (NB) and detected to have the highest accuracy of 95.34 percentage

4. Intelligent Phishing Website Detection using Random Forest Classifier Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, Touseef J. Chaudhery IEEE 2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA) In this paper, an intelligent system to detect phishing attacks is presented. We used different data mining techniques to decide categories of websites: legitimate or phishing. Different classifiers were used in order to construct accurate intelligent system for phishing website detection. Classification accuracy, area under receiver operating characteristic (ROC) curves (AUC) and F- measure is used to evaluate the performance of the data mining techniques. Results showed that Random Forest has outperformed best among the classification methods by achieving the

highest accuracy 97.36 percentage. Random forest runtimes are quite fast, and it can deal with different websites for phishing detection.

5. Phishing Website Detection Framework Through Web Scraping and Data Mining Andrew J. Park Ruhi Naaz Quadari Herbert H. Tsang 2017 IEEE The focus of this research is to establish a strong relationship between those identified heuristics(content-based) and the legitimacy of a website by analyzing training sets of websites (both phishing and legitimate websites) and in the process analyze new patterns and report findings. Many existing phishing detection tools are often not very accurate as they depend mostly on the old database of previously identified phishing websites. However, there are thousands of new phishing websites appearing every year targeting financial institutions, cloud storage/file hosting sites, government websites, and others. This paper presents a framework called Phishing-Detective that detects phishing websites based on existing and newly found heuristics. For this framework, a web crawler was developed to scrape the contents of phishing and legitimate websites. These contents were analyzed to rate the heuristics and their contribution scale factor towards the illegitimacy of a website. The data set collected from Web Scraper was then analyzed using a data mining tool to find patterns and report findings. A case study shows how this framework can be used to detect a phishing website. This research is still in progress but shows a new way of finding and using heuristics and the sum of their contributing weights to effectively and accurately detect phishing websites.

(III)System Architecture

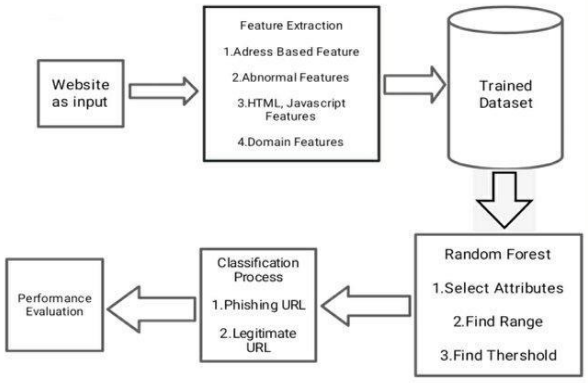

Fig 1 : System Architecture

The proposed methodology involves preprocessing imported data and importing a dataset of phishing and legal URLs. Phishing website detection is based on four types of URL features: domain-based, ad-dress-based, abnormal-based, and HTML and JavaScript features. With the processed data, certain URL features are extracted, and values for each URL attribute are generated. The URL is analysed using a machine learning algorithm that calculates the range value and threshold value for URL attributes. The URL is then classed as phishing or authentic. The attribute values are computed using phishing website feature extraction and are used to determine the range and threshold value. The values for each phishing attribute range from -1 to 1, with low, medium, and high values based on the phishing website feature. The classification of phishing and legitimate website is based on the values of attributes extracted using four types of phishing categories and a machine learning approach.

Algorithm:

Random Forest:

Working of Random Forest Algorithm

We can understand the working of Random Forest algorithm with the help of following steps

Step 1 First, start with the selection of random samples from a given dataset.

Step2 Next, this algorithm will construct a decision tree for every sample.

Then it will get the prediction result from every decision tree.

Step 3 In this step, voting will be performed for every predicted result.

Step 4 At last, select the most voted prediction result as the final prediction result.

(IV)Conclusion

As a result, we'll use Ml to build a prototype model for detecting phishing websites. We're planning to create a system that can quickly identify phishing sites. Python will be utilised as the programming language.

Future Scope:

There are billions of people using social media all around the world. However, little is known about how people use social media and the factors that influence how vulnerable they are to phishing attempts on social media platforms. Nonetheless, cyber thieves frequently exploit social networking platforms to defraud their victims. It is a tremendous source of earnings for cybercriminals because there are billions of users signing into their favourite social media accounts. So, after successfully completing this planned job for bachelor of engineering, we will try to extend our work for social media platforms such as Facebook, Instagram, and others in the future.

(V)References

[1] Matthew Dunlop, Stephen Groat, David Shelly (2010) "GoldPhish: Using Images for Content- Based Phishing Analysis"

[2] Rishikesh Mahajan (2018) "Phishing Website Detection usingMachine Learning Algorithms"

[3] Purvi Pujara, M. B.Chaudhari (2018) "Phishing WebsiteDetection using Machine Learning : A Review"

[4] David G. Dobolyi, Ahmed Abbasi (2016) "PhishMonger: A Freeand Open Source Public Archive of Real-World Phishing Websites"

[5] Satish.S, Suresh Babu.K (2013) "Phishing Websites Detection Based On Web Source Code And Url In The Webpage"

[6] Purvi Pujara, M. B.Chaudhari (2018) "Phishing Website Detection using Machine Learning : A Review"

[7] Satish.S, Suresh Babu.K (2013) "Phishing Websites Detection Based On Web Source Code And Url In The Webpage"

[8] Tenzin Dakpa, Peter Augustine (2017) “Study of Phishing Attacks and Preventions”

[9] Ping Yi (2018) “Web Phishing Detection Using a Deep Learning Framework”

[10] Jalil Nourmohammadi Khiarak (2017) “What is Machine Learning”

[11] Sadia Afroz, Rachel Greenstadt (2018) “PhishZoo: An Automated Web Phishing Detection Approach Based on Profiling and Fuzzy Matching”

[12] Arun Kulkarni, Leonard L. Brown (2019) “Phishing Websites Detection using Machine Learning”

[13] Rohan Saraf , Mayur Khatri , Mona Mulchandani (2014) “Phish Tank-A Phishing Detection Tool”

[14] Sadia Afroz, Rachel Greenstadt (2017) “PhishZoo: Detecting Phishing Websites By Looking at Them”

[15] Matthew Dunlop, Stephen Groat, David Shelly (2010) " GoldPhish: Using Images for Content-Based Phishing Analysis