LITERATURE SURVEY

DATE	**30 SEPTEMBER 2022**
TEAM ID	**PNT2022TMID03756**
PROJECT NAME	**WEB PHISHING DETECTION**

TITLE - Detecting Phishing Websites Using Machine Learning

AUTHOR : Amani Alswailem, Bashayr Alabdullah

Phishing websites are one of the internet security problems that target human vulnerabilities rather than software vulnerabilities. It can be described as the process of attracting online users to obtain their sensitive information such as usernames and passwords. The system acts as an additional functionality to an internet browser as an extension that automatically notifies the user when it detects a phishing website. The system is based on a machine learning method, particularly supervised learning.
We have selected the Random Forest technique due to its good performance in classification.It focuses 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, they conclude their paper with an accuracy of 98.8% and a combination of 26 features.

TITLE - Real Time Detection Of Phishing Websites

AUTHOR : Abdulghani Ali Ahmed , Nurul Amirah Abdullah

Web Spoofing lures the user to interact with the fake websites rather than the real ones. The main objective of this attack is to steal the sensitive information from the users.

The attacker creates a ‘shadow’ website that looks similar to the legitimate website. This fraudulent act allows the attacker to observe and modify any information from the user. This paper proposes a detection technique of phishing websites based on checking Uniform Resources Locators (URLs) of web pages.
The proposed solution is able to distinguish between the legitimate web page and fake web page by checking the Uniform Resources Locators (URLs) of suspected web pages. URLs are inspected based on particular characteristics to check the phishing web pages.The detected attacks are reported for prevention. The performance of the proposed solution is evaluated using Phistank and Yahoo directory datasets. The obtained results show that the detection mechanism is deployable and capable of detecting various types of phishing attacks maintaining a low rate of false alarms.

TITLE - Detection of Phishing Websites from URL’s by using Classification Techniques on WEKA

AUTHOR : *Buket Geyik; Kübra Erensoy; Emre Kocyigit*

Phishing is a type of fraud committed by intruders by using fake web pages to access people's private information such as user-id, password, credit card number and bank account numbers, etc. These scammers can also send email from many important institutions and organizations by using phishing attacks which imitate these web pages and act as if they are original. Traditional security mechanisms can not prevent these attacks because they directly target the weakest part of connection: end-users.
Machine learning technology has been used to detect and prevent this type of intrusions. The anti-phishing method has been developed by detecting the attacks made with the technologies used. In this paper, they combined the websites used by phishing attacks into a dataset, then obtained some results using 4 classification algorithms with this dataset. The experimental results showed that the proposed systems give very good accuracy levels for the detection of these attacks.

TITLE - An effective detection approach for phishing websites using URL and HTML features

AUTHOR : *Qingshan Jiang, Abdur Rasool ,Hui Chen, Qiang Qu & Yang Wang*

This paper proposes a new approach to solve the anti-phishing problem. The new features of this approach can be represented by URL character sequence without phishing prior knowledge, various hyperlink information, and textual content of the webpage, which are combined and fed to train the XGBoost classifier. One of the major contributions of this paper is the selection of different new features, which are capable enough to detect 0-h attacks, and these features do not depend on any third-party services. In particular, we extract character level Term Frequency-Inverse Document Frequency (TF-IDF) features from noisy parts of HTML and plaintext of the given webpage. Moreover, our proposed hyperlink features determine the relationship between the content and the URL of a webpage.

TITLE - A Desktop Application to Detect Phishing Webpages through Heuristic Approach

AUTHOR : Routhu Srinivasa Rao and Syed Taqi Ali

In this paper, we implemented a desktop application called PhishShield, which concentrates on URL and Website Content of phishing page. PhishShield takes URL as input and outputs the status of URL as phishing or legitimate website. The heuristics used to detect phishing are footer links with null value, zero links in body of html, copyright content, title content and website identity. PhishShield is able to detect zero hour phishing attacks which blacklists unable to detect and it is faster than visual based assessment techniques that are used in detecting phishing. The accuracy rate obtained for PhishShield is 96.57% and covers a wide range of phishing web sites resulting less false negative and false positive rate.