WEB PHISHING DETECTING

***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 URL’s by extracting and analyzing various features of legitimate and phishing URL’s. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URL’s as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

A phishing website is a common social engineering method that mimics trustful uniform resource locators (URL’s) and webpages. The objective of this project is to train machine learning models and deep neural nets on the dataset created to predict phishing websites. Both phishing and benign URL’s of websites are gathered to form a dataset and from them required URL and website content-based features are extracted. The performance level of each model is measures and compared

  There is a demand for an intelligent technique to protect users from the cyber-attacks. In this study, the author proposed a URL detection technique based on machine learning approaches. A recurrent neural network method is employed to detect phishing URL. Researcher evaluated the proposed method with 7900 malicious and 5800 legitimate sites, respectively. The experiments’ outcome shows that the proposed method’s performance is better than the recent approaches in malicious URL detection.

***LITERATURE SURVEY***

* Phishtank was proposed to carry out the inspection once a link has been pasted on the section given. This allow user to keep on track of faked website. They can copy and paste the link in order to identify whether the site that they are going to access is safe or not safe. User can use the website search feature directly or they can use information from PhishTank through its API. A search engine displayed on PhishTank website is to be used as the first method. Using its API will be the second method. API service can be avail by software builder after registering themselves on PhishTank website. Both methods mentioned above do not cost a single penny. The purpose of API’s usage is for user who has basis information on software development. Limitation of this project is there was no facility of displaying pop-up and email notification once user had access blacklisted website.
* PhishZoo was proposed to evaluate a new method for web phishing detection based on profiles of complex sites’ Appearance and content. PhishZoo makes profiles of sites comprising of the website contents and images displayed. These profiles are kept in a local folder and are either synchronized against the newly loaded sites at the time of loading or against risky sites for instance, links in email offline. Limitation of this project is there was no facility of displaying pop-up and email notification once user had access blacklisted website.

* GoldPhish was proposed to perceive and report phishing sites. This was done by using optical character recognition (OCR) to recite the text from an image of the page precisely from the company logo, grasping the top hierarchical areas from a search engine, and comparing them with the current web site. The forte of the tool lies in the user’s capability to recognize famous company logos. A phishing site cannot change a familiar company logo without the phishing target perceiving. Limitation of this project is there was no facility of displaying pop-up and email notification once user had access blacklisted website.
* AlexaRank is used as a benign and natural website benchmarking dataset. Alexa is a commercial enterprise which carries out web data analysis. It obtains the browsing habits of users from different sources and analyses them objectively for the reporting and classification of Internet web-based URLs. Researchers use the rankings provided by Alexa to collect a number of high standard websites as the normal dataset to test and classify websites. Alexa presents the dataset in the form of a raw text file where each line in the order ascended mentions the grade and domain name of a website.
* Kong j. et al., introduced a method for phishing URLs with innovative lexical features and blacklist. They collected a list of URLs using a crawler from URL repositories and collected 18 common lexical features. They implemented advanced ML techniques consisting of under/oversamples and classification. The automated approaches outperform other existing ML apporaches. The study has focused on content features and not lexical features, which was difficult to implement in real-world environments. The experimental results were better than the existing classification algorithms.