Web Phishing Detection – Literature Survey

Team ID: PNT2022TMID27711

Team Leader :

Vignesh S – 311419104095

Team Members :

Selvam Narendiran M – 311419104070

Surya A – 311419104084

Venkatesh R - 311419104093

Abstract:

Phishing is a type of cybersecurity attack that involves stealing personal information such as passwords, credit card numbers, etc. To avoid phishing scams, we have used Machine learning techniques to detect Phishing Websites. Therefore, in this paper, we are trying to find the total number of ways to find Machine Learning techniques and algorithms that will be used to detect these phishing websites. We are using different Machine Learning algorithms such as KNN, Naive Bayes, Gradient boosting, and Decision Tree to detect these malicious websites. The research is divided into the following parts. The introduction represents the focused zone, techniques, and tools used. The Preliminaries section has details of the preparation of the information that is required to move further. Later the paper emphasizes the detailed discussion of the sources of information.

[1] Machine Learning-Based Phishing Detection from URLs,

Authors: Sahingoz, O. K., Buber, E., Demir, O., & Diri, B

The dataset used is self-constructed. Where phishing websites belong to PhishTank and legitimate URLS are from Yandex Search API. The main purpose was to detect the word which is similar to brand names, to detect keywords, the words, which are formed from random characters. Various classification algorithms such as Naive Bayes, Random Forest, kNN(n=3), Adaboost, K-star, SMO, and Decision Tree including some feature extraction types such as

NLPbased features, Word Vectors and Hybrid are used. This system gets high accuracy throughout the test.

[2] Detection of phishing URLs using machine learning techniques

Authors: J. James, Sandhya L. and C. Thomas

The system proposed used a method based on lexical features, host properties, and properties related to the page for the detection of phishing websites. For getting a proper understanding of the pattern of URLs, various data mining algorithms are used. The classification algorithms such as Naïve Bayes, J48 Decision Tree, K-NN, and SVM were considered for the detection of phishing websites. Decision Tree had better accuracy of 91.08% compared to other algorithms. So Tree-based classifiers are best suited for phishing URL classification.

[3] Performance study of classification techniques for phishing URL detection

Authors: Pradeepthi, K. V., & Kannan, A

The system recognizes Phishing URLs, by examining the URL structure without attending the Phishing URL using classification algorithms. The data collected is first passed through the training state where it undergoes feature selection and classification. The dataset used here contains 4500 URL records, on which classification is performed. Out of which 2500 URLs are genuine and the other 2000 are the phishing ones. The 2500 URLs were collected from the DMOZ repository. The 2000 phishing URLs have been picked from
PHISHTANK. Data classification after extraction of the relevant features was performed by Naive Bayes, Random Forest, Random Tree, Multi-layer Perceptron, C-RT, J 48 Tree, LMT, C 4.5, ID 3, and K-Nearest Neighbour. The Random Forest Algorithm had the highest classification accuracy.

[4] Phishing Website URL Detection using Machine Learning,

Authors: Dipayan Sinha, Dr. Minal Moharir, Prof. Anitha Sandeep,

Detection of phishing websites is performed by using machine learning techniques like Logistic Regression, Decision tree, Random Forest, Adaboost, Gradient Boosting, Gaussian NB, and Fuzzy pattern tree classifier. Data

collection involves phishing and legitimate websites. Extracting useful features has two steps: URL-based involves IP Address, '@' symbol in URL, dashes in URL, long URL, presence of unusual number, dot count, sub-domains in URL, etc. Domainbased includes Page Rank of the website, age of the Domain, and Validity of the Website. The dataset is split into training and testing set in the ratio 80:20. The Random Forest algorithm shows 96% of precision and recalls along with the highest F1 score of 95%.

[5] Phishing Websites Detection Using Machine Learning,

Authors: R. Kiruthiga, D. Akila,

9 A total of 15 research papers have been studied in this research paper. In this research, one method was discussed which uses five different algorithms that are Decision Tree, Generalized Linear Model, Gradient Boosting, Generalized Additive Model, and Random Forest. On comparing the results, the Random Forest algorithm had the highest accuracy of 98.4%, 98.59% recall, and precision of 97.70%. Dataset used is from the UCI machine learning repository.