Web Phishing Detection

Literature Review

Team ID: PNT2022TMID15890

Team Size: 4

Team Leader: Vasanthazhagan R A

Team member: Sathish S

Team member: Shriniwaaz K G

Team member: Tharun M V

Introduction

There are several sites on the internet that requires user's data to process their request. Sometimes these data at are at risk of being stolen. These cyber harassers use these data to threaten the users, access their confidential accounts. One such way of stealing data is Web Phishing. The harassers create a fraud website that resembles the original verified websites and make the users to type in the data by scamming them. Thereby, stealing the user's data. Web Phishing Detection is a tech where we use different machine learning algorithm to differentiate a fraudulent website from an authentic one.

Literature Review

Detecting Phishing Websites Using Machine Learning [1]

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. Accuracy of 98.8% and combination of 26 features. There are 36 features that can features that can be extracted from URL, page content and page rank. Using the combinations of these features irrelevant features are removed.

Result - Random combination of features and found it took the shape of normal distribution curve.

Review: Phishing Detection Approaches [2]

Uses different phishing detection approaches which include: Content-Based, Heuristic-Based, and Fuzzy rule-based approaches. Content-based approach does a deep analysis on pages' content. Heuristic Based Approach s discriminative features extracted by understanding and analyzing the structure of phishing web pages.

Result - Each approach has its advantages and disadvantages and improving these approaches is always required.

Real Time Detection of Phishing Websites [3]

Proposes a detection technique of phishing websites based on checking Uniform Resources Locators (URLs) of web pages by checking the Uniform Resources Locators (URLs) of suspected web pages. There are few features that used to identify fake site from a legitimate one. Some are URLs, domain identity, security & encryption, source code, page style & contents, web address bar and social human factor. Features of URL and domain names are checked using several criteria such as IP Address, long URL address, adding a prefix or suffix, redirecting using the symbol “//”, and URLs having the symbol “@”.

Result - The paper checks the authenticity of the Universal Resource Locator (URLs) based only on few characteristics for detecting phishing attacks.

Detection of Phishing Websites by using Machine Learning-Based URL Analysis [4]

Aimed to implement a phishing detection system by analyzing the URL of the webpage. detected 58 different features on the web URL which included words, digits, "=", "?", IP address, etc. They implemented the system by using 8 different algorithms Logistic Regression (LR), K-Nearest Neighborhood (KNN), Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), XGBoost, Random Forest (RF) and Artificial Neural Network (ANN).

Result - They used 3 datasets and obtained results in 8 different algorithms. Random Forest (RF) is the one seems to produce highest accuracy rate with 94.59%, 90.5%, 91.26% in the 3 datasets respectively.

Phishing Website Detection using Machine Learning Algorithms [5]

Deals with ML technologies for detection of phishing URLs by extracting and analyzing different features of legitimate and phishing URL. Python programming language is used to extract the features from URL. This paper talks about 3 machine learning algorithm Decision Tree, Random Forest and Support Vector Machine.

Result - Achieved 97.14% detection accuracy using Random Forest Algorithm with lowest false positive rate.

Phishing website detection using novel machine learning fusion approach [6]

Various machine learning algorithms like logistic regression, decision tree classifier, random forest classifier, AdaBoost, gradient boosting classifier for the phishing detection. A dataset from the UCI machine learning repository for the experiments. Two priority algorithms PA1, PA2. Based on the results of priority-based algorithms final fusion model was decided.

Result - A fusion classifier and achieved an accuracy of 97%. The proposed model was tested on one dataset only.

References

1. Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, Dr. Aram Alsedrani ‘s ***“Detecting Phishing Websites Using Machine Learning”*** published during 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), 01-03 May 2019.

2. AlMaha Abu Zuraiq, Mouhammd Alkasassbeh ‘s ***“Review: Phishing Detection Approaches”*** published during 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), 09-11 October 2019.

3. Abdulghani Ali Ahmed, Nurul Amirah Abdullah ‘s ***“Real Time Detection of Phishing Websites”*** published during 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 13-15 October 2016.

4. Mehmet Korkmaz, Ozgur Koray Sahingoz, Banu Diri ‘s ***“Detection of Phishing Websites by using Machine Learning-Based URL Analysis”*** published during 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 01-03 July 2020.

5. Rishikesh Mahajan, Irfan Siddavatam ‘s ***“Phishing Website Detection using Machine Learning Algorithms”*** published during International Journal of Computer Applications, October 2018.

6. A. Lakshmanarao,P.Surya Prabhakara Rao,M M Bala Krishna ‘s ***“Phishing website detection using novel machine learning fusion approach”*** published during 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 25-27 March 2021.