Literature Survey
Web Phishing Detection

Domain: Applied Data Science

Team Id: PNT2022TMID11399

Team Members:

Mohamed Faiz S (Lead)	-910619104045
Lavan R P	-910619104042
Mothish A	-910619104049
Naveen J	-910619104052

Paper 1:

Title: PHISHSIM: Aiding Phishing Website Detection With a Feature-Free Tool

Published on:2022

Authors: Rizka Widyarini Purwanto, Arindam Pal, Alan Blair, Sanjay Jha

Abstract:

In this paper, we propose a feature-free method for detecting phishing websites using the Normalized Compression Distance (NCD), a parameter-free similarity measure which computes the similarity of two websites by compressing them, thus eliminating the need to perform any feature extraction. It also removes any dependence on a specific set of website features. This method examines the HTML of webpages and computes their similarity with known phishing websites, in order to classify them. We use the Furthest Point First algorithm to perform phishing prototype extractions, in order to select instances that are representative of a cluster of phishing webpages. We also introduce the use of an incremental learning algorithm as a framework for continuous and adaptive detection without extracting new features when concept drift occurs. On a large dataset, our proposed method significantly outperforms previous methods in detecting phishing websites, with an AUC score of 98.68%, a high true positive rate (TPR) of around 90%, while maintaining a low false positive rate (FPR) of 0.58%. Our approach uses prototypes, eliminating the need to retain long term data in the future, and is feasible to deploy in real systems with a processing time of roughly 0.3 seconds.

Paper 2:

Title: Detection of Phishing Websites from URLs by using Classification Techniques on WEKA.

Published on:15 May 2021

Authors: Buket Geyik, Kubra Erensoy, Emre Kocyigit

Abstract:

The Internet is getting stronger day by day and it makes our lives easier with many applications that are executed on cyberworld. However, with the development of the internet, cyber-attacks have increased gradually, and identity thefts have emerged. It 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 ac-count numbers, etc. These scammers can also send e-mail from many important institutions and organizations by using phishing attacks which imitate these web pages and acts as if they are original. Traditional security mechanisms cannot 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, we combined the websites used by phishing attacks into a dataset, then we 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. Index Terms phishing attacks, machine learning, classification algorithms, phishing detection, cybersecurity

Paper 3:

Title: Detection of Phishing Websites by Using Machine Learning-Based URL Analysis

Published on:1-3 July 2021

Authors: Mehmet Korkmaz, Ozgur Koray Sahingoz, Banu Diri

Abstract:

In recent years, with the increasing use of mobile devices, there is a growing trend to move almost all real-world operations to the cyberworld. Although this makes easy our daily lives, it also brings many security breaches due to the anonymous structure of the Internet. Used antivirus programs and firewall systems can prevent most of the attacks. However, experienced attackers target on the weakness of the computer users by trying to phish them with bogus webpages. These pages imitate some popular banking, social media, e-commerce, etc. sites to steal some sensitive information such as, user-ids, passwords, bank account, credit card numbers, etc. Phishing detection is a challenging problem, and many different solutions are proposed in the market as a blacklist, rule-based detection, anomaly-based detection, etc. In the literature, it is seen that current works tend on the use of machine learning-based anomaly detection due to its dynamic structure, especially for catching the “zero-day” attacks. In this paper, we proposed a machine learning-based phishing detection system by using eight different algorithms to analyse the URLs, and three different datasets to compare the results with other works. The experimental results depict that the proposed models have an outstanding performance with a success rate.

Paper 4:

Title: Machine Learning Techniques for Detection of Website Phishing: A Review for Promises and Challenges

Published on:17 May 2021

Authors: Ammar Odeh, Ismail Keshta, Eman Abdelfattah

Abstract:

Websites phishing is a cyber-attack that targets online users to steal their sensitive information including login credentials and banking details. Attackers fool the users by presenting the masked webpage as legitimate or trustworthy to retrieve their essential data. Several solutions to phishing websites attacks have been proposed such as heuristics, blacklist or whitelist, and Machine Learning (ML) based techniques. This paper presents the state of art techniques for phishing website detection using the ML techniques. This research identifies solutions to the website's phishing problem based on the ML techniques. The majority of the examined approaches are focused on traditional ML techniques. Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), and Ada Boosting are the powerful ML techniques examined in the literature. This survey paper also identifies deep learning-based techniques with better performance for detecting phishing websites compared to the conventional ML techniques. Challenges to ML techniques identified in this work include overfitting, low accuracy, and ML techniques' ineffectiveness in case of unavailability of enough training data. This research suggests that Internet users should know about phishing to avoid cyber-attacks. This paper also points out the proposal for an automated solution to phishing websites.

Paper 5:

Title: Phishing Webpage Detection Method Based on Multidimensional Features Driven by Deep Learning

Published on:23 November 2020

Authors: JIAN FENG, LIANYANG ZOU,OU YE,JINGZHOU HAN,OU YE

Abstract:

Phishing is a kind of online attack that attempts to defraud sensitive information of network users. Current phishing webpage detection methods mainly use manual feature collection, and there are problems that feature extraction is complicated and the possible correlation between features cannot be avoided. To solve the problems, a new phishing webpage detection model is proposed, among which the main components are automatic learning representations from multi-aspects features through representation learning and extracting features by hybrid deep learning network. Firstly, the model treats URL, HTML page content, and DOM (Document Object Model) structure of webpages as character sequences respectively, and uses representation learning technology to automatically learn the representation of the webpages; then, sends multiple representations to a hybrid deep learning network composed of a convolutional neural network and a bidirectional long and short-term memory network through different channels to extract local and global features, and use the attention mechanism to strengthen the influence of important features; finally, the output of multiple channels is fused to realize classification prediction. Through four sets of experiments to verify the detection effect of the model, the results show that the overall classification effect of the model is better than the existing classic phishing webpage detection methods, the accuracy reaches 99.05%, and the false positive rate is only 0.25%. It is proved that the strategies of extracting webpage features from all aspects through representation learning and hybrid deep learning network can effectively improve the detection effect of phishing webpages.

Paper 6:

Title: Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions

Published on:17 February 2022

Authors: NGUYET QUANG DO, ALI SELAMAT, ONDREJ KREJCAR, ENRIQUE HERRERA-VIEDMA, HAMIDO FUJITA

Abstract:

Phishing has become an increasing concern and captured the attention of end-users as well as security experts. Existing phishing detection techniques still suffer from the deficiency in performance accuracy and inability to detect unknown attacks despite decades of development and improvement. Motivated to solve these problems, many researchers in the cybersecurity domain have shifted their attention to phishing detection that capitalizes on machine learning techniques. Deep learning has emerged as a branch of machine learning that becomes a promising solution for phishing detection in recent years. As a result, this study proposes a taxonomy of deep learning algorithm for phishing detection by examining 81 selected papers using a systematic literature review approach. The paper first introduces the concept of phishing and deep learning in the context of cybersecurity. Then, taxonomies of phishing detection and deep learning algorithm are provided to classify the existing literature into various categories. Next, taking the proposed taxonomy as a baseline, this study comprehensively reviews the state-of-the-art deep learning techniques and analyses their advantages as well as disadvantages. Subsequently, the paper discusses various issues that deep learning faces in phishing detection and proposes future research directions to overcome these challenges. Finally, an empirical analysis is conducted to evaluate the performance of various deep learning techniques in a practical context, and to highlight the related issues that motivate researchers in their future works. The results obtained from the empirical experiment showed that the common issues among most of the state-of-the-art deep learning algorithms are manual parameter-tuning, long training time, and deficient detection accuracy.