WEB PHISHING DETECTION

LITERATURE SURVEY

Machine learning based phishing detection from URLs

Abstract:

Due to the rapid growth of the Internet, users change their preference from traditional shopping to the electronic commerce. Instead of bank/shop robbery, nowadays, criminals try to find their victims in the cyberspace with some specific tricks. By using the anonymous structure of the Internet, attackers set out new techniques, such as phishing, to deceive victims with the use of false websites to collect their sensitive information such as account IDs, usernames, passwords, etc. Understanding whether a web page is legitimate or phishing is a very challenging problem, due to its semantics-based attack structure, which mainly exploits the computer users' vulnerabilities. Although software companies launch new anti-phishing products, which use blacklists, heuristics, visual and machine learning-based approaches, these products cannot prevent all of the phishing attacks. In this paper, a real-time anti-phishing system, which uses seven different classification algorithms and natural language processing (NLP) based features, is proposed. The system has the following distinguishing properties from other studies in the literature: language independence, use of a huge size of phishing and legitimate data, real-time execution, detection of new websites, independence from third-party services and use of feature-rich classifiers. For measuring the performance of the system, a new dataset is constructed, and the experimental results are tested on it. According to the experimental and comparative results from the implemented classification algorithms, Random Forest algorithm with only NLP based features

gives the best performance with the 97.98% accuracy rate for detection of phishing URLs.

PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning

Abstract:

Today, phishing is one of the most serious cyber-security threat in which attackers steal sensitive information such as personal identification number (PIN), credit card details, login, password, etc., from Internet users. In this paper, we proposed a machine learning based anti-phishing system (i.e., named as PHISH-SAFE) based on Uniform Resource Locator (URL) features. To evaluate the performance of our proposed system, we have taken 14 features from URL to detect a website as a phishing or non-phishing. The proposed system is trained using more than 33,000 phishing and legitimate URLs with SVM and Naïve Bayes classifiers. Our experiment results show more than 90% accuracy in detecting phishing websites using SVM classifier.

URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection

Abstract:

Malicious URLs host unsolicited content and are used to perpetrate cybercrimes. It is imperative to detect them in a timely manner. Traditionally, this is done through the usage of blacklists, which cannot be exhaustive, and cannot detect newly generated malicious URLs. To address this, recent years have witnessed several efforts to perform Malicious URL Detection using Machine Learning. The most popular and scalable approaches use lexical properties of the URL string by

extracting Bag-of-words like features, followed by applying machine learning models such as SVMs. There are also other features designed by experts to improve the prediction performance of the model. These approaches suffer from several limitations: (i) Inability to effectively capture semantic meaning and sequential patterns in URL strings; (ii) Requiring substantial manual feature engineering; and (iii) Inability to handle unseen features and generalize to test data. To address these challenges, we propose URLNet, an end-to-end deep learning framework to learn a nonlinear URL embedding for Malicious URL Detection directly from the URL. Specifically, we apply Convolutional Neural Networks to both characters and words of the URL String to learn the URL embedding in a jointly optimized framework. This approach allows the model to capture several types of semantic information, which was not possible by the existing models. We also propose advanced word-embeddings to solve the problem of too many rare words observed in this task. We conduct extensive experiments on a large-scale dataset and show a significant performance gain over existing methods. We also conduct ablation studies to evaluate the performance of various components of URLNet.

A Phishing Detection System based on Machine Learning

Abstract:

As the Internet has become an essential part of human beings' lives, a growing number of people are enjoying the convenience brought by the Internet, while more are attacks coming from on the dark side of the Internet. Based on some weaknesses of human nature, hackers have designed confusing phishing pages to entice web viewers to proactively expose their privacy, sensitive information.In this article, we propose a URL-based detection system - combining the URL of the web page URL and the URL of the web page source code as features, import Levenshtein Distance as the algorithm for calculating the similarity of strings, and supplemented by the

machine learning architecture. Due to the great accuracy in small sample numbers and binary classification, we implement Support-vector machine to be the machine learning algorithm model in our system. The system is designed to provide high accuracy and low false positive rate detection results for unknown phishing pages.

A New Method for Detection of Phishing Websites: URL Detection

Abstract:

Phishing is an unlawful activity wherein people are misled into the wrong sites by using various fraudulent methods. The aim of these phishing websites is to confiscate personal information or other financial details for personal benefits or misuse. As technology advances, the phishing approaches used need to get progressed and there is a dire need for better security and better mechanisms to prevent as well as detect these phishing approaches. The primary focus of this paper is to put forth a model as a solution to detect phishing websites by using the URL detection method using Random Forest algorithm. There are 3 major phases such as Parsing, Heuristic Classification of data, Performance Analysis in this model and each phase makes use of a different technique or algorithm for processing of data to give better results.

REFERENCES:

1. Sahingoz, O. K., Buber, E.., Demir, O. and Diri, B. (2019). Machine learning based phishing detection from URLs. Expert System Applications.(117), (pp. 345–357). doi: 10.1016/j.eswa.2018.09.029.
2. Jain A.K., Gupta B.B. “PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning”, Cyber Security. Advances in

Intelligent Systems and Computing, vol. 729, 2018, https://doi.org/10.1007/978-981-10-8536-9_44

3. Hung Le, Quang Pham, Doyen Sahoo, and Steven C.H. Hoi, “URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection”, Conference’17, Washington, DC, USA, arXiv:1802.03162, July 2017.
4. Wu CY, Kuo CC, Yang CS,” A phishing detection system based on machine learning” In: 2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA), pp 28–32, 2019.
5. Parekh, Shraddha & Parikh, Dhwanil & Kotak, Srushti & Sankhe, Prof. (2018). A New Method for Detection of Phishing Websites: URL Detection. 949-952. 10.1109/ICICCT.2018.8473085.