LITERATURE SURVEY

Date	30 September 2022
Team Id	PNT2022TMID46988
Project Name	Project--5833-1658817449
Maximum Marks	4 Marks

Abstract:

Phishing is a common attack against Internet users that causes them to reveal their information using fake websites. The goal of the fake website is to steal personal information such as usernames, passwords and online banking transactions. Scammers use websites that are visually and semantically similar to the real ones.

As technology continues to advance, phishing techniques begin to advance rapidly, and this should be prevented by using anti-phishing mechanisms such as spoofed URL detection. Machine Learning is a powerful tool used to combat spoofing attacks. This report covers machine learning technology to detect fake URLs by extracting and analyzing different characteristics of legitimate and fake URLs. Random Forest, Logistic Regression and algorithms are used to detect fake websites.

Introduction:

Nowadays, the Internet plays an important role in communication, where people create an online environment to manage business functions, online activities of banks, social networks… However, the Internet also contains hidden things. a lot of risk because when users operate in an online environment they can be vulnerable to attackers. And their identity is often a fake URL. And spoofed URLs are often placed on popular websites or sent to user emails.

Literature Review:

Construction of Phishing Site. In the first step attacker identifies the target as a well-known organization. Afterward, attacker collects the detailed information about the organization by visiting their website. The attacker then uses this information to construct the fake website.

URL Sending. In this step, attacker composes a bogus e-mail and sends it to the thousands of users. Attacker attached the URL of the fake website in the bogus e-mail. In the case of spear phishing attack, an attacker sends the e-mail to selected users. An attacker can also spread the link of phishing website with the help of blogs, forum, and so forth [43].

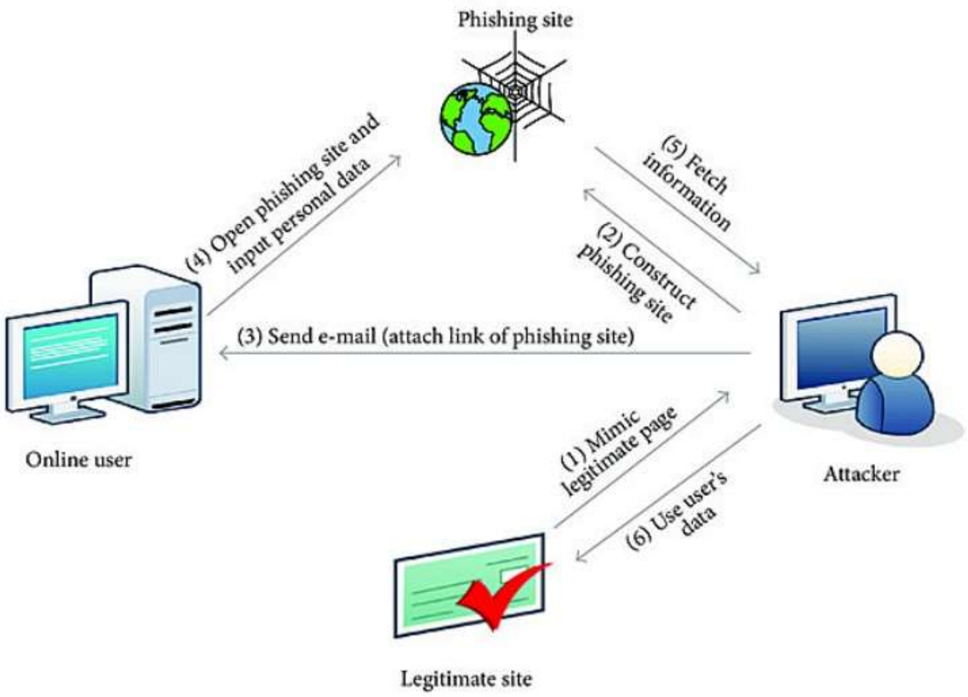
Stealing of the Credentials. When user clicks on attached URL, consequently, fake site is opened in the web browser. The fake website contains a fake login form which is used to take the credential of an innocent user. Furthermore, attacker can access the information filled by the user.

Identity Theft. Attacker uses this credential of malicious purposes. For example, attacker purchases something by using credit card details of the user.

Although attacks use different techniques to create phishing websites to deceive users, most have similarly designed phishing website features. Therefore, researchers have conducted extensive anti-phishing research using phishing website features. Current methods for phishing detection include black and whitelists, heuristics, visual similarity, and machine learning, among which heuristics and machine learning are more widely used. The following is an introduction to the aforementioned phishing detection techniques.

Black and whitelist

To prevent phishing attack threats, many anti-phishing methods have been proposed. Blacklisting methods are the most straightforward ways to prevent phishing attacks and are widely used in the industry. Google Safe Browsing uses a blacklist-based phishing detection method to check if the URL of the matching website exists in the blacklist. If it does, it is considered a phishing website.


Phishing site
(4) Open phishing site and input personal data
(5) Fetch information
(2) Construct phishing site
(3) Send e-mail (attach link of phishing site)
Online user
Attacker
(1) Mimic legitimate page
(6) Use user's data
Legitimate site

FEATURES OF PROPOSED SYSTEM:

1) **FUNCTIONAL CAPABILITIES:** The ultimate aim of this project is to detect phishing attacks in real-time. This model checks the website with machine learning server for any maliciousness in the accessed site.
2) **PERFORMANCE LEVEL:** At the client side, it takes 1-2 seconds to detect whether a site is phishing or not.
3) **DATA STRUCTURES:** The data in this project are maintained in the CSV form. It provides easy access to the user.

4) **SAFETY:** No data loss occurs in this system

5) **RELIABILITY:** We assure that the project is completely authenticated in order to enhance security and corruptions of database as well as the software.

1.phishing detection and protection scheme :

Developing with the anti-phishing methods, phishers use various phishing methods and more complex and hard-to-detect approaches. The most straightforward way for a phisher to swindle people is to make the phishing web page similar to their target. However, many distinctive and features can distinguish the original legitimate website from the clone phishing website like the spelling error, image alteration, long URL address and abnormal DNS records. The full list is revealed in Table 3 which is used later in our analysis and classification study. If an attacker clones a legitimate website as a whole or designed to look similar as they usually do in most attacks in recent times,our approach is that similar looking phishing web page con-tent is not left for the users to check for the indicator or the authenticity attentively, but can detect by automated methods. Our approach is based on website phishing detection using the features of the site, content and their appear-ance. These properties are stored in a local database (Excel table) as a knowledge model and first compared with the newly loaded site at the time of loading against the dangerous web page offline. After the comparison was unable to detect the similarity, then the critical approach to compare the legitimate and fake using the features of the website with machine learning for an intelligent decision. The critical contribution of our approach includes Result:The output is determined by the classifier, in the phishing detection stage which predicts if the web page is suspicious, legiti-mate or phishing. The knowledge model and plug-in development will be developed at a later stage

2. System detection related work :

Nowadays most people uses internet for various purposes such as online shopping like purchasing or selling products, chat with friends, sending mail. Internet users now spend more time on social networking sites Information can spread very fast and easily within the social media networks. Social media systems depend onusers for content contribution and sharing. Facebook had over 1.3 billion active users as of June 2014. there are over 1.3 billion (the number is keep growing) pages from various categories, such as company, product/service, musician/band, local business, politician, government, actor/director, artist, athlete, author, book, health, beauty, movie, cars, clothing, community. Fans not only can see information submitted by the page, but also can post comments, photos and videos to the page.

Result:

Domain anomaly features are used to identify possible malicious domains based on lexical and reputation factors, whereas social anomaly features represent anomalous user behaviors in social communications

3. Learning to Detect Phishing Emails :

An alternative for detecting these attacks is a relevant process of reliability of machine on a trait intended for the reflection of the besieged deception of user by means of electronic communication. This approach can be used in the detection of phishing websites, or the text messages sent through emails that are used for trapping the victims. Approximately, 800 phishing mails and 7,000 non-phishing mails are traced till date and are detected accurately over 95% of them along with the categorization on the basis of 0.09% of the genuine emails.

Result:

We can just wrap up with the methods for identifying the deception, along with the progressing nature of attacks.

4. Phishing websites machine learning:

Phishing URL is a widely used and common technique for cybersecurity attacks. Phishing is a cybercrime that tries to trick the targeted users into exposing their private and sensitive information to the attacker. The motive of the attacker is to gain access to personal information such as usernames, login credentials, passwords, financial account details, social networking data, and personal addresses. These private credentials are then often used for malicious activities such as identity theft, notoriety, financial gain, reputation damage, and many more illegal activities. This paper aims to provide a comprehensive and comparative study of various existing free service systems and research-based systems used for phishing website detection. The systems in this survey range from different detection techniques and tools used by many researchers. The approach included in these researched papers ranges from Blacklist and Heuristic features to visual and content-based features. The studies presented here use advanced machine learning and deep learning algorithms to achieve better precision and higher accuracy while categorizing websites as phishing or benign. This article would provide a better understanding of the current trends and existing systems in the phishing detection domain.

Result:

Phishing URL detection plays a pivotal role for many cybersecurity software and applications. In this paper, we researched and reviewed works based on the advanced machine learning techniques and approaches that promise a fresh approach in this domain.

5. Support vector machine :

The existing anti-phishing approaches use the blacklist methods or features based machine learning techniques. Blacklist methods fail to detect new phishing attacks and produce high false positive rate. Moreover, existing machine learning based methods extract features from the third party, search engine, etc. Therefore, they are complicated, slow in nature, and not fit for the real-time environment. To solve this problem, this paper presents a machine learning based novel anti-phishing approach that extracts the features from client side only. Below architecture diagram as shown in Fig. 1. represents mainly flow of training phase to Detection phase. First data need to be pre-processed and feature extraction using different feature sets and later we need to train this dataset with the corresponding algorithms and the output is displayed.

Result:

In future we can use a combination of any other two or more classifier to get maximum accuracy. We can also explore various phishing techniques that uses Lexical features.

QUALITY:

The project is developed with the help of Anaconda Navigator software which meets the requirement of the user, the project is checked whether the phases individually have a served its purpose.

REFERENCES:

[1] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor and J. Downs, "Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions", *Proceedings of the 28th international conference on Human factors in computing systems ser. CHI'10. New York NY USA:ACM*, pp. 373-382, 2010.

[2] B. Krebs, "HBGary Federal HACKED by Anonymous", December 2011,

[3] W. D. Yu, S. Nargundkar and N. Tiruthani, "A phishing vulnerability analysis of web based systems", Proceedings of the 13th IEEE Symposium on Computers and Communications (ISCC 2008). Marrakech Morocco: IEEE, pp. 326-331, July 2008.

[4] P. Kumaraguru, Y. Rhee, A. Acquisti, L. F. Cranor, J. Hong and E. Nunge, "Protecting people from phishing: the design and evaluation of an embedded training email system", *Proceedings of the SIGCHI conference on Human factors in computing systems ser. CHI'07. New York NY USA: ACM*, pp. 905-914, 2007.

[5] C. Yue and H. Wang, "Anti-phishing in offense and defense", *Computer Security Applications Conference 2008. ACSAC 2008. Annual*, pp. 8-12, 2008.

[6] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong and C. Zhang, "An empirical analysis of phishing blacklists", *Proceedings of the 6th Conference in Email and Anti-Spam ser. CEAS'09 Mountain view CA*, July 2009.

[7] Y. Zhang, J. I. Hong and L. F. Cranor, "Cantina: a content-based approach to detecting phishing web sites", *Proceedings of the 16th international conference on World Wide Web ser. WWW '07. New York NY USA:ACM*, pp. 639-648, 2007.

[8] H. Zhang, G. Liu, T. Chow and W. Liu, "Textual and visual content-based anti-phishing: A bayesian approach", *IEEE Transactions on Neural Networks*, vol. 22, no. 10, pp. 1532-1546, oct. 2011.

1. :