WEB PHISHING DETECTION

INTRODUCTION:

Phishing is a sort of social engineering in which an attacker delivers a phoney (e.g., spoofed, false, or other misleading) communication intended to dupe a person into giving the attacker access to sensitive information or to install dangerous software, such as ransomware, on the victim's infrastructure. Phishing attacks are getting more and more complex, and they frequently transparently mirror the site that is being attacked, allowing the attacker to watch everything the victim does there and to cross any further security barriers with them.

The term "phishing" was first used in Koceilah Rekouche's 1995 cracking toolkit AOHell, while it's probable that it was used earlier in a print version of the hacker magazine 2600. The phrase refers to the employment of more sophisticated lures to "fish" for users' sensitive information. It is a fishing-related word that was influenced by phreaking.

Some of the phishing attacks are email phishing, spear phishing, voice phishing, SMS phishing, clone phishing. Our project helps us to find out illegitimate websites by using machine learning algorithms, analysing the given datasets with help of some paramteres like ip address, url length, rightclick, domain registration, having @ symbol in the url etc and train the machine learning model with the given dataset. Also to create an website which traps the phishing websites and deploy the website in an cloud environment.

LITERATURE REVIEW:

Paper [1] Web content mining, web usage mining and web structure mining is also a part of the web mining. We follow the Web URL Mining. The source of phishing attacks are mostly from email, websites and malware. The links (URL) provided in phishing emails draws user into entering phishing website. In website based phishing, website is copy of the original website which looks like same but, the aim is different. To overcome the problem of phishing we design a framework to detect it using the fuzzy logic as a classifier. We used only URL based features for the extraction process. For that we collect our dataset from the Phish tank site, Open phish site and the URL of the website which can live on the web. We were collecting the 1000 URLs for our dataset purpose.

Paper [2] In our experiment we develop pages which are visually similar to original web sites. We upload these pages on free hosting site. After preparing that all we asked student to go to specified URL, which was www.myfb.comze.com Participants are given following scenario “Imagine that you receive an email message that asks you to click on one of the following links”. And one restriction was to use only from one of above three web browsers. After entering address on their browsers they have to report what happened in case of each web browser.

Paper [3] Numerous Anti-phishing tools are available which help to protect against phishing websites. Google chrome, Mozilla Firefox and Safari uses Google Safe Browsing (GSB) service which blocks the site if it is phishy. Other tools like Netcraft, Mcfee Site Advisor, Avast, Quick Heal are also in use. Google Safe Browsing service uses blacklist approach to analyze a URL. As the blacklist was not updated, Google Safe Browsing service could not detect the phishing site. Netcraft also detected a site. But a phishing site was reported and was not blocked by Netcraft. Netcraft does this when it is not sure if the website is 100%

phishing. Also, no warning is given unless user clicks on the icon on right to check the risk rating. This can be very risky if user does not checks the rating or decides to use the site even after seeing the risk rating. Anti-virus software like Avast and Quick Heal also provide protection against online security threats. We installed Avast anti-virus to check for its functioning. We found that the special Avast browser for secure browsing was unable to detect the phishing URL that Netcraft or Google Safe Browsing detected successfully. The higher security services in Avast were paid so we couldn't verify them. The above survey explicitly acknowledges a major necessity of an advanced Anti-phishing tool. It can also be significantly noted that these tools are installed separately. A layman may never install such tools if unaware of the phishing practices. In that case one is only relying on the Google Safe Browsing service. Thus, it is very important to create awareness regarding these tools and phishing attacks. Additionally, one must not completely rely on the Anti- phishing tools as well, as it can be seen that they may result into misclassification.

Paper[4], we proposed HTMLPhish, a deep learning based data-driven end-to-end automatic phishing web page classification approach. HTMLPhish receives the HTML content of a web page as input and applies CNNs to learn the semantic dependencies in both the characters and words in the HTML document in a jointly optimized network. Furthermore, we applied convolutions on a concatenation of the matrix of character and word embeddings in order to ensure the effective embedding of new words in the test HTML documents. Our approach can learn context features from HTML documents without requiring extensive manual feature engineering. We evaluated our model using a comprehensive dataset of HTML contents presented in a real-world distribution. HTMLPhish provided a high precision rate, showing a temporally stable result even when it was trained two months before being applied to a test dataset. The future work is to compare our model to feature engineering-based models that extract features only from the HTML document. Also, we intend to implement our model as a browser extension. This will enable HTMLPhish to recognise phishing websites in real-time.

Paper[5], The proposed system enables the internet users to have a safe browsing and safe transactions. Its helps users to save their important priivate details that should not be leaked. Providing our proposed system to users in the form of extension makes the process of delevering our system much easier. The results points to the efficiency that can be achieved using the hybrid solution of hueristic features, visual features and blacklist and whitelist approach and feeding these features to machine learning algorithms. A particular challege in this domains is that criminals are constantly making new strategies to countrer our defense meausres. To succeed in this context, we need algorithms that continually adapt to new examples and features of phishing URL’s. And thus we use online learning alogrithms. This new system can be designed to avail maximum accuracy. Using different approaches altogether will enhance the accuracy of the system, providing an efficient protection system. The drawback of this system is detecting of some minimal false positive and false negative results. These drawbacks can be eliminated by introducing much richer feature to feed to the machine learning

REFERENCES:

[1]. H.Chapla, R. Kotak and M. Joiser, "A Machine Learning Approach for URL Based Web Phishing Using Fuzzy Logic as Classifier," 2019 International Conference on Communication and Electronics Systems (ICCES), 2019, pp. 383-388, doi: 10.1109/ICCES45898.2019.9002145.

[2]. N. Mazher, I. Ashraf and A. Altaf, "Which web browser work best for detecting phishing," 2013 5th International Conference on Information and Communication Technologies, 2013, pp. 1-5, doi: 10.1109/ICICT.2013.6732784.

[3]. M. D. Bhagwat, P. H. Patil and T. S. Vishawanath, "A Methodical Overview on Detection, Identification and Proactive Prevention of Phishing Websites," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 1505-1508, doi: 10.1109/ICICV50876.2021.9388441.

[4]. C. Opara, B. Wei and Y. Chen, "HTMLPhish: Enabling Phishing Web Page Detection by Applying Deep Learning Techniques on HTML Analysis," 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9207707.

[5]. V. Patil, P. Thakkar, C. Shah, T. Bhat and S. P. Godse, "Detection and Prevention of Phishing Websites Using Machine Learning Approach," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697412.