Phishing URL Detecting Website Using Machine Learning

Akshad Satkar[1], Pranay Nandagawali[1], Shubham Londhe[1], Tejas Tayade[1], Prof.Prakash.H.Patil[2]

[1] *Student, Department of Information Technology, DYPIET, Ambi, Maharashtra, India*
[2] *Head of Deparment, Department of Electronics and Telecommunication, DYPIET, Ambi, Maharashtra, India*

ABSTRACT

Phishing attacks remain a significant concern for computer system defenders, often serving as the initial phase in a multi-stage attack. Despite significant advancements in phishing detection, certain phishing emails can circumvent filters by altering the message's structure and meaning. To address this issue, we implemented a machine learning classifier on a large corpus of legitimate and phishing emails. Our system, SAFEPC (Semi-Automated Feature Generation for Phish Classification), extracts features, some of which are elevated to higher-level features, to outsmart conventional phishing email detection strategies. To evaluate SAFE-PC, we obtained a substantial corpus of phishing emails from a tier-1 university's central IT organization. Our implementation of SAFE-PC on the dataset revealed previously unknown insights into phishing campaigns targeted at university users. SAFE-PC surpasses a state-of-the-art email filtering tool, detecting more than 70% of phishing emails.

Keyword:- *Machine learning*

1. INTRODUCTION:

Phishing refers to the act of impersonating a legitimate website to deceive users into divulging their personal information, including usernames, passwords, account numbers, and national insurance numbers. Phishing scams are perhaps the most prevalent type of cybercrime today, and they can occur in numerous domains, such as online payment systems, webmail, financial institutions, file hosting or cloud storage, and others. Among all industry sectors, webmail and online payment systems are the most vulnerable to phishing attacks. These attacks can take the form of email phishing scams or spear phishing, so users should exercise caution and not entirely rely on common security applications. Machine learning is a highly efficient technique for detecting phishing attempts and overcomes the drawbacks of existing approaches.

2. SOFTWARE INFORMATION:

Python: Python is a general-purpose programming language that is interpreted and high-level. It was created by Guido van Rossum and was first released in 1991. Python prioritizes code readability, which is achieved through its unique usage of significant whitespace. Its object-oriented approach and language constructs aim to facilitate the writing of clear and logical code for projects of all scales. Python features dynamic typing and garbage collection and supports various programming paradigms, including procedural, object-oriented, and functional programming. Additionally, Python is renowned for its comprehensive standard library, which is often referred to as a "batteries included" feature.

Anaconda: Anaconda is a free and open-source distribution of Python and R, designed for scientific computing. Its main purpose is to make package management and deployment simpler. It contains data science packages that are compatible with Windows, Linux, and macOS. Anaconda, Inc. founded the software in 2012 and currently

maintains it. Additionally, Anaconda offers other products such as Anaconda Team Edition and Anaconda Enterprise Edition. Conda manages the package versions in Anaconda, and it can be used separately for other purposes besides Python. Miniconda, a smaller version of Anaconda, includes only the essential packages. Anaconda comes with more than 250 pre-installed packages, and users can access over 7,500 other open-source packages from PyPI and conda. It also includes a user-friendly graphical interface called Anaconda Navigator as an alternative to the command line.

SVM Algorithm: Support Vector Machines (SVM) is a powerful supervised machine learning algorithm used for classification and regression analysis. It works by finding the best possible boundary or hyperplane that separates data points of different classes in a high-dimensional space. The algorithm aims to maximize the margin or distance between the hyperplane and the nearest data points of each class. SVM can handle non-linearly separable data by transforming the input data into a higher-dimensional space using kernel functions. SVM is widely used in various applications such as image classification, text classification, and bioinformatics.

3. PROBLEM STATEMENT:

Phishing poses a significant threat as it involves the theft of sensitive user information such as banking details, emails, and other private data, through social engineering and technical deception. Due to its exploitation of human vulnerabilities, security measures are often insufficient in thwarting phishing attacks, which typically involve fraudulent emails and websites.

4. CONCLUSION:

In conclusion, machine learning has proven to be an effective tool in the fight against phishing attacks. By utilizing various machine learning algorithms and techniques, phishing detection websites can quickly and accurately identify potential threats and prevent users from falling victim to social engineering attacks. With the continued evolution of phishing tactics, it is essential to keep improving and refining these technologies to ensure the highest possible level of protection for internet users. By combining the power of machine learning with ongoing education and awareness efforts, we can continue to make significant strides in preventing phishing attacks and safeguarding sensitive information.

5. REFERENCES:

[1] N. Goel, A. Sharma, and S. Goswami, “A way to secure a qr code: Sqr,” in 2017 International Conference on Computing, Communication and Automation (ICCCA). IEEE, 2017, pp. 494–497.

[2] V. Mavroeidis and M. Nicho, “Quick Response Code Secure: A Cryptographically Secure Anti-Phishing Tool for QR Code Attacks,” in International Conference on Mathematical Methods, Models, and Architectures for Computer Network Security. Springer, 2017, pp. 313– 324

[3] K. Krombholz, P. Fruhwirt, T. Rieder, I. Kapsalis, J. Ullrich, and E. Weippl, “QR Code Security–How Secure and Usable Apps Can Protect Users Against Malicious QR Codes,” in 2015 10th International Conference on Availability, Reliability and Security. IEEE, 2015, pp. 230–237

[4] Denso Wave, “QR Code development story,” 2019, [Accessed: 28-Mar2019]. [Online]. Available: https://www.denso-wave.com/en/technology/ vol1.html

[5] A. Dabrowski, K. Krombholz, J. Ullrich, and E. R. Weippl, “QR inception: Barcode-in-barcode attacks,” in Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones Mobile Devices. ACM, 2014, pp. 3–10

[6] “Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection” Mahdieh Zabihimayvan and Derek Dora

[7] “Detection of Phishing Attacks with Machine Learning Techniques in Cognitive Security Architecture” Ivan Ortiz-Garces, Roberto O. Andrade, and MariaCazares

[8] “A Methodical Overview on Phishing Detection along with an Organized Way to Construct an Anti-Phishing Framework” Srushti Patil, Sudhir Dhage.