[Tuning] Improve phishing URL algorithm

Harrison Watycha

Summary

In this report, and alongside the practical work that will be done, there will be improvement to understanding phishing and prevent phishing through the URL algorithm. Multiple features will be researched and studied, before there is an attempt to implement these features within the coding and it became an effective URL algorithm.

This report, the technical side will have some background research into the background of phishing URL algorithms and gain a better understanding of how the algorithm works. Following that, there will be a specific purpose of this report. Furthermore, case studies will be highlighted, and key findings will be expressed. Finally, the conclusion and implementation steps will be displayed.

Introduction

Background

In the research found, phishing is a type of cyber-attack that works based on essentially tricking the victim into providing their private information, under the assumption they are providing the information to someone who is trustworthy. Unlike smishing, it does not have to be just through an SMS message, it can be through a URL across multiple platforms. It also clears through the research that it can be difficult to detect this phishing as a lot of attackers will implement evasion techniques.

Purpose of the Research

The clear purpose of this report is to gain a foundational understanding of phishing URL algorithms, and then utilising that knowledge, be able to improve the algorithm. Improving by adding features and bolstering the engineering of the code and making sure the code can efficiently prevent phishing through URL. Additionally, another purpose of this research is to implement the highlighted improvements and make sure it will be an effective implementation.

Literature Review

In the review of the available literature, it became clear that through the academic journal conducted by Ponnusamy (2023, pg.18) phishing attacks are very serious attacks as when the victim is deceived, they often result in significant financial loss and or identify theft. Furthermore, the journal also highlights that attackers will commonly use domain squatting and adjusting/ shortening the URL, this is done to make it harder to prevent and track the attacker.

The next literature source is Chew (2007, pg.13) who indicates that feature engineering and especially features that will detect common features of URL phishing can be an effective way to locate the phishing. Also, Chew highlights that the content on the URL that is clicked can be dynamic and not always look like a scam, making it easy for the user to be deceived.

The final literature source is Li (2023, pg.2) who indicates the best methods to prevent phishing within URL'S is implementing an effective machine learning model (MLM), as this will learn the patterns of phishing and the URLS, can be able to detect it more accurately. Also, Tokenization is a great feature that would break down the URL and makes the attempted scam easier to analysis as it is all in their own features.

The source of Li (2023, pg.4) in their academic journal expresses when discussing the ethical considerations of improve phishing algorithms, if you are building an MLM that incorporates past phishing there must be a level of encryption as to protect past victims.

Discussion

Key Findings

Key findings of this report are that implementing any form of improvement to implement or update an effective MLM, also update and or update the effective way of Tokenization. Furthermore, another key finding has been the ability of an MLM to make the algorithm even more significantly improved, by providing the features of the algorithm in the early stages of the development, it will make the MLM have less vulnerabilities.

Additionally, another important finding has been the ability of keyword extraction, and how it can identify if a message is smishing, using AI.

Best Practices

Implementing a MLM and tokenization. Also, keyword feature analysis.

Implementing the features within an MLM, to incorporate it.

Take in ethical considerations, the usage of data provided by the clients.

Challenges and Solutions

Domain squatting and URL shortening (solution)

Creating an MLM. (Challenge)

A challenge would be having the properly trained personal working on the implementation of algorithm and features, not only will it take time but also some knowledge in the coder of choice (python in this case). (Challenge)

Conclusion

Summary of Key Points

The key takeaways from this report are the opportunities to improve the algorithm, this mostly around implementing new key features that improve the algorithm. Additionally, it is also clear that when detecting URL'S (via phishing) it can be hard to pinpoint if it is phishing's due to the attacker utilising squatting techniques and or URL shortening to hide their intentions.

Another key takeaway has been the implementation of improving the algorithm, and the importance of completing it during the construction in the MLM as the machine will then learn the patterns and common features of the phishing.

In addition to this, it is also clear that when specifically improving the features of an AI model it is important to consider the challenges that are posed by either technical challenges and or attackers avoiding detection.

References

List all sources cited in the report.

- Ponni Ponnusamy. (2023). An Optimized Bagging Learning with Ensemble Feature Selection Method for URL Phishing Detection. *Journal of Electrical Engineering and Technology*, 19(3), pp.1881–1889. doi:https://doi.org/10.1007/s42835-023-01680-z.
- Chew, M. (2007). A framework for detection and measurement of phishing attacks. *Proceedings of the 2007 ACM workshop on Recurring malcode - WORM '07*. doi:https://doi.org/10.1145/1314389.1314391.
- Li, T. (2023). A Survey of Intelligent Detection Designs of HTML URL Phishing Attacks. *IEEE Access*, pp.1–12. doi:https://doi.org/10.1109/access.2023.3237798.