Trans. Japan Soc. Aero. Space Sci.,
Vol. XX, No. X, pp.1 - 4, 20XX
DOI: 10.2322/tjsass.XX.1

PHISHING DETECTION*

AI in cyber security

Hiba SHAHZAD

Phishing is a widespread and ever-growing type of cyberattack that aims to deceive victims into disclosing confidential information such as password, financial, and personal information. As the phishing model continues to improve, the classical phishing detection mechanism becomes ineffective. This paper experiments and compares two rule-based phishing detectors that are simple and easy to understand: URL Rule-Based System and Content Rule-Based System, utilizing the Phishing Websites Dataset that consists of publicly available data of 11,055 instances. The former phishing detector inspects the URLs using attributes such as excessive length, lacking the HTTPS protocol, the presence of special characters such as "@" and "–", the use of IP addresses, and the use of sub-domains. The latter phishing detector scrutinizes the text and HTML aspects of the URLs using attributes such as urgent language phrases, inconsistencies of the links, phishing domains of the senders, and password submission forms. Both phishing detectors are tested using accuracy, precision, recall, F1-measure, and confusion matrices.

The accuracy of the URL Rule-Based system was found to be 44.60

This paper proves that rule-based systems are appropriate for use as baseline approaches but are inadequate for practical phishing defense solutions. The conclusions of the paper stress the importance of adaptive approaches such as machine learning, natural language processing, and deep learning that are capable of detecting dynamic patterns to ensure effective phishing defenses.

Key Words: phishing

1. Introduction

Phishing is the term used to describe an attempt to obtain private information, usually passwords, bank account information, credit card numbers, usernames, or other sensitive data, with the intention of using or selling the stolen data. sometimes the attacker creates a website that is identical to an original website; when a user visits this page, the victim finds a website they think is authentic rather than a fake copy by clicking on a hyperlink within a forum or through a search engine.

1. The victim visits the phishing website after clicking on the email.
2. The attacker gathers the victim's login information.
3. The attacker gains access to a website using the victim's login credentials. The victim is giving their personal information to a spam website if he does. Malware can occasionally be downloaded to the target computer as well.

2. research question

"What is the difference between a content-based rule system and a URL-based rule system in terms of phishing attack detection?"

3. Literature Review

After reviewing paper [1], phishing attacks are explained in depth, including their history and the motivations behind attackers. The paper highlights that achieving high accuracy in phishing detection has always been a key challenge, and although many new techniques have recently emerged, no single method is sufficient because phishing can be carried out in many different ways.

Paper [2] focuses on deceptive phishing, which is the most common type of phishing attack. In this type of scam, attackers pretend to be a trusted source and send emails asking victims to verify information, re-enter login details, or provide passwords.

In paper [3], the authors discuss deceptive phishing on social networking sites and propose a solution using Data Mining and WordNet Ontology. The study notes that while some solutions exist for text-based phishing through instant messaging, detection of phishing carried out through voice chat has not yet been addressed. The authors propose using Association Rule Mining (ARM) combined with a speech recognition system to improve privacy in Instant Messengers (IM).

Paper [4] examines browser vulnerabilities. Similar to other for-profit software, browsers may have flaws that hackers use to initiate phishing scams. New vulnerabilities may be introduced with each new feature or third-party add-on,

*Presented at the XXth International Symposium on Space Technology and Science, June 4-9, 20XX, Matsuyama, Japan
Received XX June 20XX; final revision received XX August 20XX; accepted for publication XX September 20XX

†Corresponding author,

making it more challenging to identify and stop phishing attacks.

Paper [5] reviews several recent phishing attacks carried out using modern techniques.

Paper [6] discusses online phishing detection and prevention. The study shows that analyzing URL features and lexical patterns can improve accuracy. By combining URL analysis with email classification, the authors achieve highly precise anti-phishing results.

Paper [7] proposes a classification scheme based on requirements engineering—a field that focuses on defining real-world goals, system functions, constraints, and how these specifications evolve over time.

Paper [8] explains phishing as an email-based scam where attackers disguise themselves to steal sensitive information. The paper draws on deception theory and cognitive psychology, emphasizing how attention and information processing affect a user's ability to recognize phishing attempts.

Paper [9] presents a Bayesian framework for detecting phishing web pages. The model uses both textual and visual content to measure how similar a suspicious webpage is to a legitimate one.

Finally, paper [10] reports on a user study that examines how people react to common "trust indicators." The findings reveal why phishing emails and websites often appear convincing to users.

4. Methodology

The entire experimental procedure for identifying phishing websites using two distinct rule-based AI systems is described in this section. Preparing the dataset, extracting features, running two algorithms, and assessing the model using common performance metrics are all part of the experiment.

4.1. Dataset

A publicly accessible dataset called the "Phishing Websites Dataset" was used for this experiment. The 11,055 records in this dataset, which comes from the Kaggle/UCI Machine Learning Repository, are classified as phishing (1), legitimate (1), or suspicious (0). There are 31 predictive features in the dataset, including:

possessing an IP address
URL Length
Service for Shortining
Prefix Suffix
SSLfinal State
Request URL
Google Index

These attributes stand for domain attributes, page content, URL structure, and external link behaviour. The majority of the features in the dataset are binary or categorical, which makes it simple to convert them into logical rules, making it appropriate for rule-based detection.

Prior to testing, Python Pandas was used to load the dataset, clean it up, and standardise the column names. Since there were no missing values, the dataset was used straight away.

4.2. Algorithm 1 (URL rule-based systems

A rule system based on URLs is the first AI model. This method concentrates on structural elements that are taken straight from the URL. A number of manually created guidelines were put into effect, motivated by typical phishing traits:

URL Length: Obfuscation is frequently indicated by long URLs. The URL is flagged as risky if the length is abnormally long.

Phishing URLs frequently contain the characters "@" or "-" in order to deceive users or conceal the true domain.

Absence of "https": URLs that only use the http protocol, without secure protocol, are more likely to be harmful.

Raw IP addresses are frequently used in place of domain names in phishing links.

Excessive Subdomains: In misleading links, an excessive number of subdomains (such as login.bank.secure.verify.com.fake.ru) are frequently used.

These guidelines were applied to every URL. The algorithm categorised the URL as phishing if three or more dangerous conditions were met; if not, it was categorised as legitimate. This straightforward rule-based classifier serves as a standard against which more sophisticated methods can be evaluated.

4.3. Algorithm 2 – Content Rule-Based System

The second model analyzes the content, rather than the URL structure of the email or webpage. This approach exploits the linguistic and HTML-based cues that are common in phishing attacks. The rules include:

Presence of urgent or manipulative words:

Examples of phishing triggers are words and phrases like "verify", "update", "password", "urgent", "account suspended".

Suspicious Sender Domain:

The chance of phishing also rises with the use of fake e-mail domains or mismatched sender names.

Mismatched HTML Links: Links whose visited text is different from the real URL denotes deception. Forms requesting credentials Many phishing pages include forms requesting username, credit card, and password information. Each message and web page was analyzed using regular expressions. A score is assigned based on the number of patterns that appeared suspicious. If the score exceeded a threshold, it was classified as phishing.

4.4. Evalution metrics

The dataset was divided into train (70

Accuracy: Indicates how accurate a prediction is overall.

Precision: The proportion of phishing URL predictions that turned out to be phishing.

Recall: The capacity to accurately identify phishing attempts.

F1-Score: The harmonic mean of recall and precision.

True Positives, True Negatives, False Positives, and False Negatives are displayed in the confusion matrix.

Scikit-learn (sklearn) was used to calculate these metrics.

4.5. implementation of tools

Python was used to carry out the experiment using:
Pandas for managing datasets
For textual and pattern-based analysis, use re (regex).
scikit-learn for evaluation

NumPy for numerical operations

Every piece of code was created and run in a local Python environment.

5. Result

After the deployment of the two rule-based phishing detectors, URL Rule-Based AI and Content Rule-Based AI, the next task involved the evaluation of the detectors using the Kaggle Phishing Websites Dataset. This dataset consists of a total of 11,055 instances with 30 attributes related to URLs, which are labeled as phishing sites (-1) and genuine sites (1). The data underwent the operations of both algorithms. Algorithm 1 applied rules related to the structure of URLs, and Algorithm 2 applied text/content rules. After passing the data through both algorithms, the predictions were verified with the "Result" ground truth. The accuracy level of both algorithms came out to be around 44-45Both algorithms performed poorly because of the high diversity of the phishing URLs. These systems are not capable of adapting to novel kinds of attacks. The false positives were high since often the normal URLs may contain characteristics that appear fishy, such as very long domain names and sub-domains. False negatives occurred in phishing URLs employing modern evasion techniques such as: a. Using the Windows PowerShell interpreter short URLs, HTTPS with valid certificates, "legitimate-looking As a whole, the findings emphasize the limitation of rule-based algorithms when it comes to the detection of phishing attacks. These algorithms work well in understanding the behavior patterns but fail to be a match for algorithms that are implemented using the concept of machine learning.

6. Conclusion

Using the Phishing Websites Dataset, this experiment assessed two rule-based phishing detection systems. The objective was to determine the effectiveness of straightforward, interpretable rules in classifying phishing URLs in comparison to more contemporary methods. Accuracy scores of roughly 44–45

The findings unequivocally demonstrate that static rule systems can no longer handle the sophistication of phishing attacks. Traditional rules are unreliable because attackers commonly alter URLs, use URL shorteners, enable HTTPS, and imitate the patterns of legitimate websites. Consequently, there were a lot of false positives and false negatives in both algorithms. The study did, however, effectively illustrate the usefulness of rule-based systems as baseline models. They provide interpretability, transparency, and ease of implementation—all crucial attributes for security education and low-resource settings. The results also demonstrate why machine learning, natural language processing, and deep learning are essential components of contemporary phishing detection systems since they can adjust to changing attack patterns.

Future enhancements could consist of:

incorporating machine-learning classifiers like XGBoost, Random Forest, or SVM,

dynamically extracting host-based, content-based, and lexical features,

analysing webpage screenshots with deep learning

utilising live phishing feeds to update rules on a regular basis.

To sum up, this project demonstrated that although rule-based systems are quick and easy to understand, they are insufficient for phishing detection in everyday life. For robust protection, more sophisticated, adaptable models are needed.

7. Discussions

Phishing: What is it? a cyberattack that deceives people into divulging private information. Usually, phoney emails, URLs, or websites are used.

7.1. Why This Project?

Phishing is on the rise everywhere.

The ability of rule-based algorithms to identify phishing had to be tested.

7.2. Details of the Dataset

"Phishing Websites Dataset" was used.

30, features, 11,055 records.

Output label: -1 indicates phishing, 1 indicates legitimacy.

7.3. Algorithm 1: AI Based on URL Rules

Verifies textual elements such as:

IP rather than domain

Too long of a URL

An excessive number of subdomains

"@" or "-" in the domain

Missing HTTPS Decision rule: Phishing is classified if three or more dangerous conditions are met.

7.4. Algorithm 2: AI Based on Content Rules

Verifies textual elements such as:

Words: check, update, click

Untrustworthy domain

The displayed link and the actual link are not matching.

Form asking for a password Decision: Mark as phishing if several red flags are found.

7.5. Metrics for Evaluation

Precision

Accuracy

Remember

F1-Score

matrix of confusion

7.6. Final Precision

Rule-Based URL: approximately 44.6

Rule-Based Content: approximately 44.5

7.7. conclusion

Although rule-based systems are straightforward, they lack accuracy.

Attackers often alter their patterns.

In contemporary phishing detection, machine learning performs better.

8. References

[1]Gupta, B. B., Tewari, A., Jain, A. K., Agrawal, D. P. (2017). Fighting against phishing attacks: state of the art and future challenges. Neural Computing and Applications, 28(12), 3629-3654.

[2] Huang, H., Zhong, S., Tan, J. (2009, August). Browser-side countermeasures for deceptive phishing attack. In 2009 Fifth International Conference on Information Assurance and Security (pp. 352-355). IEEE.

[3] Ali, M. M., Siddiqui, O. A., Nayeemuddin, M., Rajamani, L. (2015, January). An approach for deceptive phishing detection and prevention in social networking sites using data mining and wordnet ontology. In Electrical, Electronics, Signals, Communication and Optimization (EESCO), 2015 International Conference on (pp. 1-6). IEEE.

[4] Raffetseder, T., Kirda, E., Kruegel, C. (2007, May). Building anti-phishing browser plug-ins: An experience report. In Proceedings of the Third International Workshop on Software Engineering for Secure Systems (p. 6). IEEE Computer Society.

[5] Yadav, S., Bohra, B. (2015, October). A review on recent phishing attacks in Internet. In 2015 International Conference on Green Computing and Internet of Things (ICGCIoT) (pp. 1312-1315). IEEE.

[6] Chen, J., Guo, C. (2006, October). Online detection and prevention of phishing attacks. In Communications and Networking in China, 2006. ChinaCom'06. First International Conference on (pp. 1-7). IEEE.

[7] Zave, P. (1995, March). Classification of research efforts in requirements engineering. In Proceedings of 1995 IEEE International Symposium on Requirements Engineering (RE'95) (pp. 214-216). IEEE.

[8] Wang, J., Herath, T., Chen, R., Vishwanath, A., Rao, H. R. (2012). Research article phishing susceptibility: An investigation into the processing of a targeted spear phishing email. IEEE transactions on professional communication, 55(4), 345-362.

[9] Zhang, H., Liu, G., Chow, T. W., Liu, W. (2011). Textual and visual content-based anti-phishing: a Bayesian approach. IEEE Transactions on Neural Networks, 22(10), 1532-1546.

[10] Jakobsson, M., Tsow, A., Shah, A., Blevis, E., Lim, Y. K. (2007, February). What instills trust? a qualitative study of phishing. In International Conference on Financial Cryptography and Data Security (pp. 356-361). Springer, Berlin, Heidelberg