CLASSISFING VARIOUS ALGORITHM FOR PHISHING URL

[1] K MOSELIN ESTHER, [2]B JAYAPRADHA

[1] Research Scholar, Dr Ambedkar Government Arts College, [2] Assistant Professor, Dr Ambedkar Government Arts College

Abstract

The use of the Internet for routine activities including banking, mailing, gaming, and buying is on the rise. An attacker trying to steal sensitive information like passwords, credit card numbers, or personal information will use a fake website to resemble a legitimate one. This paper evaluated classification algorithm using a technique selected by a various feature and find the better accuracy to use for the future work to identify the fake URL.

Key word: Phishing, Classification, Features.

1. INTRODUCTION

Phishing URLs can be recognized by inspecting a number of factors, including the domain name, the HTTP or HTTPS protocol, the usage of misspelled or dubious phrases, and the existence of additional parameters in the URL. Machine learning algorithms, rule-based strategies, and browser extensions are a few of the ways that can be used to identify and stop phishing URLs.

A cyberattack known as phishing involves the attacker trying to trick the target into divulging sensitive information such as passwords, usernames, or credit card numbers. This is typically done using email, instant messaging, or other forms of electronic communication that appear to be provided by a trustworthy source, such a bank, a social media platform, or a company.

This type of attack has the goal of collecting private information that can be used for fraud involving your identity, finances, and other malicious acts. Social engineering techniques are frequently employed in phishing attacks to prey on the mental state of the victim and persuade them to do actions that benefit the attacker, including clicking on a malicious link, uploading a file that includes malware, or giving personal information.

Phishing, also is becoming a bigger risk. Phishing is a problem that affects people, businesses, and groups all around the world, so it's important to know the risks and take appropriate actions to protect your information as well as yourself from these attacks.

Phishing is another name for this. Since they commonly mirror authentic websites, these phishing websites, also known as fake websites, can be difficult to identify.

2. PROPOSED APPROACH

WEBSITE FEATURE

Some websites are explained below for phishing sites, which may be used randomly by hackers for stilling information from unknown people.

2.1 Address Bar-Based Feature

It is the top of the web browser in a user interface. The current webpage (URL) is displayed in it. we can also type the name or URL in the address bar. It is also known as the location bar and is designed like user friendly it might also list the previous history for the usage. Without the address bar, it will not perform.

2.1.1 URL Length

In the address bar, phishing can use long URL to hide the doubtful path

http://sitebulb.com/hints/dupllicate-content/technically-duplicate-urls//web-phishing-clone-website-and-host-fake-facebook-for-n00bs-4cf8bb8ca548? /phishing web/.com.

To our study, we have collected the URL length in the dataset and produced an average URL length. If the URL length is greater than or equal to 54 characters then the URL is classified as phishing in the results.[3]

$$\underline{\textbf{RULE}} = \text{IF}\begin{cases} \text{URL length } < 54 \rightarrow \text{legitimate} \\ \text{otherwise} \rightarrow \text{phishing} \end{cases}$$

Names	Length
Google Chrome	2mb (2,097,152)
Mozilla Firefox	65,5236
Safari	80000(After that error message will display)
Edge	2083
Apache	4000(After that error message will display)
Opera	unlimited

2.1.2. @ Symbol

One of the fake signs of fake websites is the use of the '@' symbol within the URL of the Address. This may lead users to neglect all characters before the '@'so attackers can guide the user to a fake website [4].

$$\underline{\textbf{RULE}} = \text{IF}\begin{cases} \text{URL having @ } \rightarrow \text{ phishing} \\ \quad \text{otherwise} \rightarrow \text{legitimate} \end{cases}$$

2.1.3. Redirecting using "//"

A "//" in the URL route indicates that the user will be forwarded to another website.

"http://www.legitimate.com//http://www.phishing.com" is an illustration of one of these URLs. We looked at the spot where the "//" appears. We discover that the "//" is implied if the URL begins with "HTTP." should come in at number six. However, the "//" should be in the seventh position if the URL uses "HTTPS.[3]

$$\underline{\textbf{RULE}} = \text{IF}\begin{cases} \text{The position of the last"}//\text{"}\ \mathit{in\ the\ URL} > 7 \ \rightarrow \text{ phishing} \\ \qquad\qquad \text{otherwise} \rightarrow \text{legitimate} \end{cases}$$

2.2 Domain-based features

2.2.1. Age of Domain

Most phishing websites can live for a short period of time but if the website has been for more than a year is a good sign of security. This feature can be extracted from the dataset [5].

A legitimate website may remain on a blacklist for a very long period even after the phishing website has left the domain, affecting whether or not its reputation is safe.

$$\underline{\textbf{RULE}} = \text{IF}\begin{cases} \text{age of domain} > 6\ @ \ \rightarrow \text{ phishing} \\ \qquad \text{otherwise} \rightarrow \text{legitimate} \end{cases}$$

2.2.2.DNS Record

A DNS record (Domain Name System record) is a type of data stored in a domain name server (DNS)that maps domain names to IP addresses or other information. DNS record is used by web browsers, email clients, and other Internet applications to find the server that hosts a particular domain. If the record is empty or not found then the website is phishing.

$$\underline{\textbf{RULE}} = \text{IF}\begin{cases} \text{no DNS record for the domain} \rightarrow \text{ phishing} \\ \qquad \text{otherwise} \rightarrow \text{legitimate} \end{cases}$$

2.2.3 Website Traffic

The number of users who visit a website is known as website traffic. It is an important metric for website owners and marketers, as it provides insight into how popular a website is and how well it is performing.

$$\underline{\textbf{RULE}} = \text{IF}\begin{cases} \text{website rank} < 100{,}000 \rightarrow \text{ legitimate} \\ \quad\quad\text{otherwise} \rightarrow \text{ phishing} \end{cases}$$

3. IMPLEMENTATION

Our dataset was divided into a training dataset and a test dataset. While the training dataset is used to fit a machine learning algorithm or model, a final model fits on the training dataset. We used 75% for training and 25% for testing from our dataset consisting of 11430 data.

Classification Algorithm

We have applied several classifiers for training, testing and evaluating the performance. Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), MLPClassifier, GaussianNB, Perceptron was applied.

We evaluated the performance of our suggested system using a variety of performance metrics, including Accuracy, Precision, Recall, and F1-score, which can be determined using Findings for a secure website

The four terms that a phishing website returned: Following are the definitions of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

CLASSIFIER	ACCURACY	PRECISION	RECALL	F1-score
NB	88.48	0.77	0.63	0.68
LR	91.90	0.91	0.91	0.92
PP	90.47	0.90	0.89	0.89
CNN ID	93.4	0.93	0.93	0.93
DT	85.9	0.86	0.89	0.89
KN	89.85	0.90	0.92	0.92
CNN+XGBOOST	96.32	0.98	0.98	0.97

DT+XGBOOST	89.98	0.9	0.98	0.97
DT+LR	91.98	0.90	0.97	0.90
DT+NB	94.09	0.96	0.92	0.90

4. CONCLUSION

Phishing attacks continue to be a significant threat to individuals and organizations alike. One of the most common forms of phishing is the use of phishing URLs, which are designed to find legitimate websites to trick users into providing sensitive information. Fortunately, various algorithms have been developed to detect and prevent these attacks.

5. REFERANCE

1. R. Das, M. Hossain, S. Islam, A. Siddiki et al., “Learning a deep neural network for predicting phishing website,” Ph.D. dissertation, Brac University, 2019
2. A. Alswailem, B. Alabdullah, N. Alrumayh, and A. Alsedrani, “Detecting phishing websites using machine learning,” in 2nd International Conference on Computer Applications & Information Security (ICCAIS). Riyadh, Saudi Arabia: IEEE, 2019, pp. 1–6.
3. R. M. Mohammad, F. Thabtah, and L. McCluskey, “Phishing websites features,” School of Computing and Engineering, University of Huddersfield, 2015.
4. Wenyin, Liu, et al. “Discovering phishing target based on semantic link network.” Future Generation Computer Security Application conf. (ACSAC’06), Dec 2006, pp. 381-392.
5. Canali, David et al.” Prophiler:a fast filter for the large scale detection of malicious web pages.” Processing of the 20th international conference on world wide web, ACMA,2011.