IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PHISHING WEBSITE DETECTION AND PREVENTION BASED ON LOGISTIC REGRESSION

[1]Priyanka Gupta, [2]Amit Mahajan

[1]Assistant Professor, [2]Student

[1]School of Computer Science & Engineering,

[1]Lovely Professional University, Jalandhar, Punjab

Abstract: Phishing is becoming one of the most serious and rapidly growing threats, because phishing attackers obtain data that was entered by the client and use it without the customer's knowledge. In today's world, personal information is more valuable than money, hence phishing website attackers are focused on the user's personal information and exploiting it when the user submits personal information into phishing websites. As a result, the fundamental goal of these efforts is to prevent phishing attackers from misusing personal information. To solve this problem, a machine learning approach called logistic regression is utilized with a large dataset. This dataset is used to train the algorithm, which aids in the detection of new web connections that are fake. Attackers masquerade their website as legitimate in order to obtain data from users. They entice visitors to visit a website in order to obtain the personal information required. It is critical to determine whether the provided link is good or a phishing link before attempting to access such websites. We can protect ourselves from intruders and keep our data safe by verifying the link.

Index Terms **– Phishing, Machine Learning, Logistic Regression, and website Links.**

I. INTRODUCTION

The Internet has evolved into a critical foundation that provides unparalleled comfort to human society. Nonetheless, the Internet is also associated with some inescapable security difficulties, such as phishing, malicious programming, and security exposure, all of which have posed serious risks to customers' finances. Phishing, according to the APWG, is a criminal instrument that uses both social structure and focused deception to steal purchasers' unique character data and budgetary record capabilities. Clicking on a web link is one of the most dangerous and common ways for phishing websites to obtain the client's personal information. The attacker offers an appealing link to the client through email or another method in this phishing assault. If you click on that link and fill out all of the sensitive information, phishing attackers can use the information to steal money from your account or sell it to third parties for any purpose. Nowadays, there is a lot of software that identifies and rejects fraudulent transactions. However, most people disregard links, whether they are good or bad, therefore to identify links, we use a machine learning approach to determine whether they are good or bad. Phishing assaults are on the rise in today's digital world, thanks to the usage of technologies like email, mobile phones, and online connections by phishing attackers. The daily clients may receive a large number of links by email or other means, or they may click on a link without knowing what sort of link it is, thus the main goal of the project is to educate the client about the link using machine learning algorithms such as linear regression. This program is used to determine if a website's link is good or bad. This program extracts data from a data collection and trains a machine learning system using logistic regression. The system will be able to detect phishing URLs after it has been trained. This method splits the supplied link and checks the primary keywords from it, followed by a search of the database for good and bad links. If it matches one of those, it produces the appropriate result. If the output is satisfactory, we can continue; otherwise, we must terminate the procedure. When a new link is discovered, the list will be updated. First, this algorithm determines whether or not the link adheres to the web's principles. Checking is done by dissecting a link into its subdomains, as specified in the software's code. Once the connection has passed all of the regulations and processes, it can be used. Its categories the output as either good or bad.

II. RELATED WORK

Phishing is one of the types of assault that is becoming more prevalent in everyday life. To obtain the users' personal and sensitive information and use it against them. Around 36.29 percent of phishing websites have been discovered in the last six years. In the previous two years, 97.36 percent of those were discovered. This is related to an increase in internet usage. This is the most difficult duty for cyber security. To develop new techniques to entirely stop them, which is impossible, but we can halt them for a few years. By devising innovative methods for detecting and blocking certain URLs. In terms of approach and procedure, the present methods differ.

Component Page Phishing Detection: Social networking sites are the fastest-growing sites, with the majority of users using them. All of the user's personal information will be stored on this site. There are a lot of them on the market. As a result, there should be a lot of study done in this field, where the number of users is growing every day. Few people assault these sites in order to obtain data from them. Few will create a new social website to collect data that looks precisely the same. They appear to be the same website, and some people may be duped into providing personal information to these websites. We evaluate the webpage similarity components in the real webpage to see if the page is trustworthy or not in order to locate these websites.

Phishing website categorization using intelligent rules: Phishing is defined as the stealing of a user's sensitive information and using it against him. Phishing attacks come in a variety of forms. There are no two phishing attacks alike. Phishing, on the other hand, is a process that can be categorized. It's similar to how assaults may be categorized based on how they obtain data. One of these is through sending email links. Other methods include building a phishing website and stealing information, and there are a handful in which control is lost with a single click on a link, and the attacker has complete access. As a result, we cannot have the same response to every attack. However, we may get the same result by categorizing them. As a result, that strategy would be more convenient. The goal of this project is to categories attacks based on how they obtain data.

Detecting phishing attacks based on network performance indicators Network-based phishing is used to slow down a server so that more traffic may flow through it, giving the attacker an opportunity to access someone's information. This is the most common type of phishing. This can also be accomplished by phishing. It's like a bug in the link that does nothing but enter into your device and take control of the network and obtain the data when you click it. So, in order to detect such linkages, we have the four-lever based checker, which will take the process through these phases in order to provide an output. These will extract information from the link and assist us in identifying the phishing link.

Phishing and Malware Warnings Research: Phishing attacks can only happen if we allow them to. Our computer is capable of blocking specific URLs that are potentially detrimental to the system. Because it is ultimately in the hands of the user, this technique will examine human activities when clicking such links in order to detect phishing links. This research will assist us in identifying similar websites or connections so that we can quickly identify them, shut them down, and protect individuals from being assaulted.

III. PROPOSED METHODOLOGY

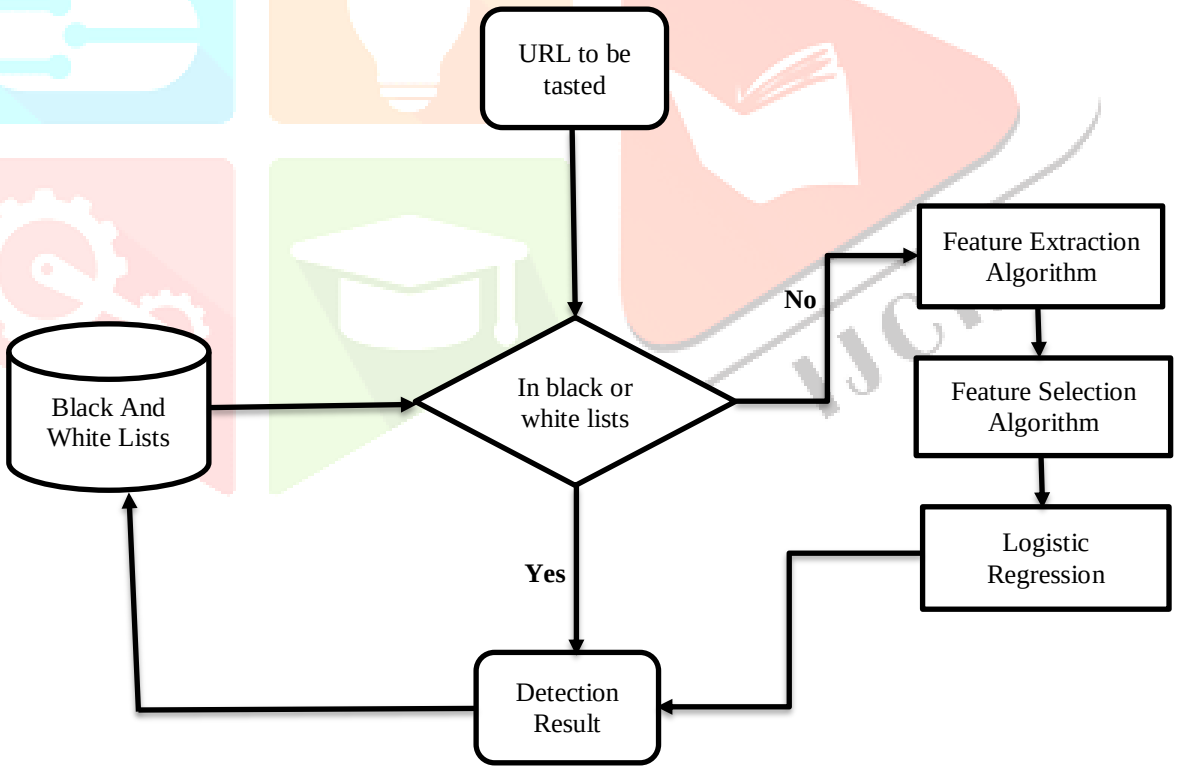

Fig 1. Flow Chart of the working Process

The phases in the implementation process are as follows:

3.1 Preprocessing and data collection

One of the most crucial jobs in the development of a machine learning model is data collection. It is the collecting of task-related data based on a set of defined factors in order to assess and provide useful results. Some e data, on the other hand, may be noisy, containing imprecise, incomplete, or wrong values. As a result, data must be processed before it can be analyzed and conclusions drawn. Data cleansing, data transformation, and data selection are all examples of data pre-processing.

3.2 Data Cleaning and transformation

The practice of removing noise from data is known as data cleaning. This noisy data, in the sense that it is incomplete and contains some missing values, will be completely written. This will be taken, and all of the data will be read, with the prospect of clean data being obtained so that the data may go to the next step. Smoothing, aggregation, generalization, and transformation are examples of data transformations that increase data quality.

3.3 Datasets Selection

Data selection refers to a set of procedures or functions that enable us to choose the most valuable data for our system. Input of data: We applied the best algorithm to discover the phishing website once we found the best algorithm. Then we'll feed data into the algorithm, and depending on the results, we'll calculate the output. A set is a collection of data from a single occurrence. We require sets for a variety of reasons, including training, in order to work on machine learning. Data set: A set of data that we feed into our machine learning algorithm to train our model. This is the data type that is *utilized* to offer an unbiased evaluation of the final product that has been produced and fit on the training data set.

3.4 Algorithm

Regression: It's a model for forecasting. It uses probability to forecast the connection between two values.
Dependent: What is the 14th of February's sale? The dependent variable is her sale.
Independent: 1) The number of products sold on February 14th.
2) Quantity (predictor).

Logistic Regression:

The outcome of logistic regression is expressed in 0's and 1's, and these values are used to forecast the output of the provided data, such as High (or) Low. Normally, the range of linear regression is 0-∞ and -∞ to +∞.
Because the sigmoid function is used in logistic regression, the linear line must be trimmed to 0 and 1. The threshold value represents the likelihood of 0 or 1, or good or bad, via sigmoid with this. To derived the logistic regression range 0→∞ to 0→1, these equations are obtained from the straight-line equation y = m1x1+ c (-∞ →+∞). So, we use **equation: -** If y=0, the equation is equal to 0. As a result, the y value varies between y=1→∞ and you may use the log final logistic regression equation to convert it further to get between -∞ →+∞.

IV. IMPLEMENTATION AND RESULT

To extract the link's features, you'll need a means to enter the link. A model generated with Python code will be used to carry out the task. After clicking the link, the input block redirects to python code, which uses machine learning methods to extract the features.

```
def predict_url(link):
    return loaded_model.predict(link)

link = input()
print(predict_url([link]))
```

Fig 2. Phishing Input Block

Phishing links are usually disguised as authentic. By using this method, an attacker can trick a victim into clicking on their fake web links that appear to be genuine. These phishing URLs have few distinguishing characteristics. These characteristics are divided into four types, as seen in fig 2 above.

1. Feature extraction
2. Address bar
3. Abnormal Features
4. Domain Features

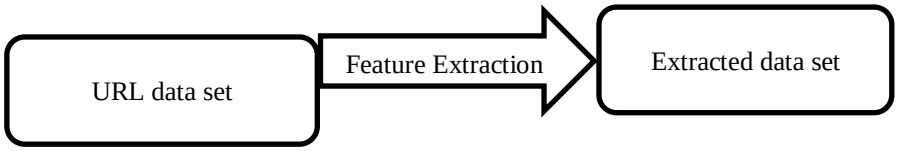

Fig 3. Feature Extraction

Address bar features: To correctly identify phishing attempts, special characters such as @,"", /, _, -, and the length of the input URLs must be given in this IP address. It also has a URL length, as well as the availability of Short URLs. The length of a domain registration is also examined. These features are reviewed and confirmed to see if the supplied link includes all of them. If the requirements are met, the link will be forwarded to the next one.

Abnormal Features: The requested URL and link tags like <META>, <Script>, <Link>, and the S F H status information may be used to check for abnormal characteristics. I'm getting feedback from the handler. Examine whether the web page contains a link that would send the data submitted to an email address. Finally, specify if it is an Abnormal URL or not.

Domain features: Because phishing websites have such a brief lifespan, this piece of the code examines the domain's age. It examines the DNS record, web traffic, and page rank. Because phishing pages have relatively few visits, the index will be 0 or 1 at most. It also examines the Google index for that, as well as the number of pages connecting to that page. All of the feature information is kept in the Statistical report.

The decision tree and logistic regression technique are used to forecast the outcome based on the data provided. The algorithm's accuracy rate may also be seen here. It will be useful in determining if a phishing website is good or not.

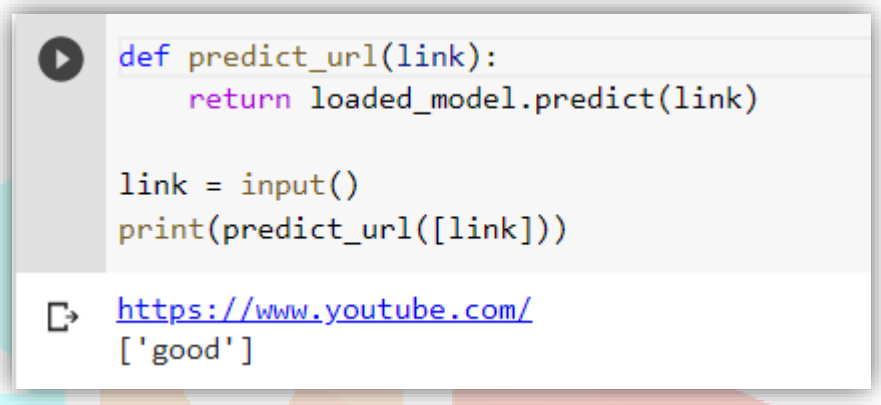

Fig 4. Prediction Output

V. Conclusion

In the world of the internet, phishing is an unfathomable danger. In this assault, the average person enters personal information onto a phoney website that appears to be a legitimate website. Based on current research, a survey of phishing strategies is conducted. This offered a thorough knowledge of the assault as well as a number of potential responses. In this study, several methodologies for detecting phishing have been discussed; nevertheless, most of the algorithms still have limitations such as accuracy, failure to differentiate objects, and so on. However, logistic regression approach is used to determine accuracy in this study because it is an open problem in the phishing industry. The accuracy gained is 95%, and based on the research, it may grow further.

References

[1] "Dynamical credibility assessment of privacy-preserving strategy for opportunistic mobile crowd sensing," vol. 6, pp. 37430-37443, IEEE Access Multidisiplinary, June 2018. D. Wu, L. Fan, C. Zhang, H. Wang, and A. Wang, "Dynamical credibility assessment of privacy-preserving strategy for opportunistic mobile crowd sensing," vol. 6, pp. 37430-37443, IEEE Access Multidisiplinary,

[2] "A survey on Privacy Preserving Data Aggregation Schemes in People Centric Sensing Systems and Wireless Domains," Indian journal of science and technology, Vol 9, 2016. K. R. Jansi and S.V. Kasmir Raja, "A survey on Privacy Preserving Data Aggregation Schemes in People Centric Sensing Systems and Wireless Domains," Indian journal of science and technology, Vol 9, 2016.

[3] "Design Perspectives of People Centric Sensing Systems," Indian Journal of Science and Technology, Vol 9(37), 2016. K. R. Jansi and S.V. Kasmir Raja, "Design Perspectives of People Centric Sensing Systems," Indian Journal of Science and Technology, Vol 9(37), 2016.

[4] K. Jansi, S.V. Kasmir Raja, and G.K. Sandhia, "Efficient privacy-preserving fault tolerance aggregation for people-centric sensing systems," Service Oriented Computing and Applications, Service Oriented Computing and Applications, Service Oriented Computing and Applications, 12:305–315 (2018).

[5] "A personalised whitelist approach for phishing webpage detection," in Proc. 7th Int. Conf. Availability, Rel. Security (ARES), Aug. 2012, pp. 249–254. A. Belabed, E. Aimeur, and A. Chikh, "A personalised whitelist approach for phishing webpage detection," in Proc. 7th Int. Conf. Availability, Rel. Security (ARES), Aug. 2012,

[6] "Anti-phishing based on automated individual white-list," in Proc. 4th ACM Workshop Digit. Identity Manage., 2008, pp. 51–60. Y. Cao, W. Han, and Y. Le, "Anti-phishing based on automated individual white-list," in Proc. 4th ACM Workshop Digit. Identity Manage., 2008, pp. 51–60.

[7] "Visually detecting identical Web pages: Application for phishing detection," Internet Technol., vol. 10, no. 2, pp. 1–38, May 2010. T.-C. Chen, S. Dick, and J. Miller, "Visually detecting identical Web pages: Application for phishing detection," Internet Technol., vol. 10, no. 2, pp. 1–38, May 2010.

[8] "Clientside protection against Web-based identity theft," by N. Chou, R. Ledesma, Y. Teraguchi, D. Boneh, and J. C. Mitchell. Netw.Distrib.Syst.Security Symp. (NDSS), 11th Annu., 2004, pp. 1–16

[9] "Cluster based Efficient Privacy Preserving Data Aggregation (CEPPDA) in People Centric Sensing Networks," International Journal of Advanced Science and Technology, Vol.29(6), pp.1450-1465, 2020.

[10] "QR-Code scanner based car sharing," ARPN Journal of Engineering and Applied Sciences, 2018, 13(10), pp. 3441-3448. Arulprakash, M., Kamal, A., and Manisha, A., "QR-Code scanner based vehicle sharing," ARPN Journal of Engineering and Applied Sciences, 2018, 13(10), pp. 3441-3448.