A MACHINE LEARNING BASED-APPROACH FORDETECTION OF PHISHING SITES

Dr. Mohamadi Begum Y

mohamadi.begum@presidencyuniversity.in

B. Uday Jaswanth Reddy.

201910101259@presidencyuniversity.in

B. Mounika

201910100910@presidencyuniversity.in

Ch. Raghu Ram Reddy

201910101077@presidencyuniversity.in

G. Yasaswitha

201910101120@presidencyuniversity.in

K.TEJA

201910101124@presidencyuniversity.in

ABSTRACT - Individuals and companies have suffered considerable losses as a result of the growth in phishing attempts, and is concerned about data privacy and confidentiality. Current phishing detection technologies are unable to keep up with the rising number of attacks, and more efficient solutions are required. This study presents a unique machine learning-based phishing detection technique to address this issue. This study examines the limitations of the already developed methods in detecting attacks and suggests a new innovative approach .In machine learning processing, Kaggle dataset is used. To detect phishing attempts, the proposed model utilises machine learning techniques. In the end, this study presents a promising new approach to phishing detection that has the potential to increase the efficiency and accuracy of existing detection approaches.

Keywords Phishing attack, Machine learning,Spam

1. INTRODUCTION

Companies lose $100 billion every year due to phishing, and phishing assaults are rising by 200% every year. This leads to the conclusion that the present solutions are insufficient, and that new ways for protecting businesses and end users should be developed. Many financial transactions are computerised, and there is less cash accessible, which has led to a new trend of phishing and other cybercrimes, scamming internet users to gather their bank credentials. In recent years, many criminal organisations have moved their focus from exploiting system weaknesses in information systems to abusing humans' incapacity to distinguish between legal and bogus internet resources such as email and websites. As a result, it is critical to propose a solution to the challenges.

The requirement for an effective solution makes phishing detection a popular research topic in recent years. The "blacklist" method, which involves adding blacklisted Sites and IP addresses (Internet Protocol) to the antivirus database, is frequently used to identify phishing websites.
Attackers use innovative deception tactics including encryption and a variety of other core methods, like quickly-flux, in which proxy are automatically built to host the website, algorithmic development of new URLs, and others, to avoid blacklists.

Zero hour of phishing efforts been identified using heuristicbased detection, which comprises characteristics known to appear in real-world phishing attacks. However, these traits are not always guaranteed to be present in such attacks, and the detection's false positive rate is extremely high.

To overcome the drawbacks of blacklisted and heuristic-based techniques, many security professionals are highly concentrating on machine learning (ML)methodologies. Multiple algorithms used in machine learning technologies provide predictions or inferences about future data based on historical data. To properly identify fake websites, especially zero-hour phishing websites, the programme will examine several prohibited and legitimate URLs and their properties.

The objective of this study is to identify malicious URLs and to identify the best machine learning technique by evaluating the accuracy rate, false positive and false negative rate of each algorithm.

2. DATASET

Name of our dataset is phishing_site_urls.csv The given data set is in comma separated values(.csv file).

5,49,346 unique entries have been used

Label column is prediction column which has 2 categories

Good variable - means the urls is safe and no malicious stuff

Bad variable - means the urls not safe and malicious stuff is present in the link or url

No null value in datasets .This dataset is taken from Kaggle

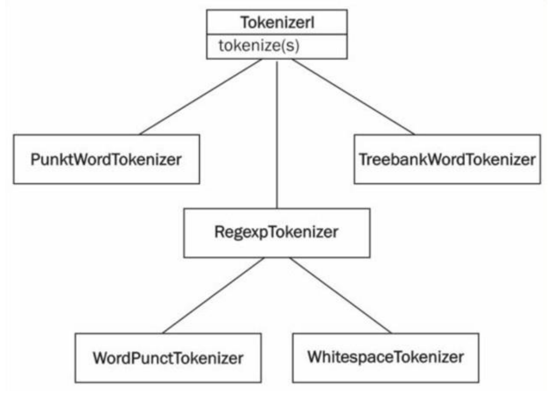
3. MACHINE LEARNING ALGORITHM

Five machine learning classification model Regexp Tokenizer , Snowball Stemmer Beautiful Soup, Logistic Regression ,Multinomial lNB has been selected to detect phishing websites.

IMPLEMENTATION:

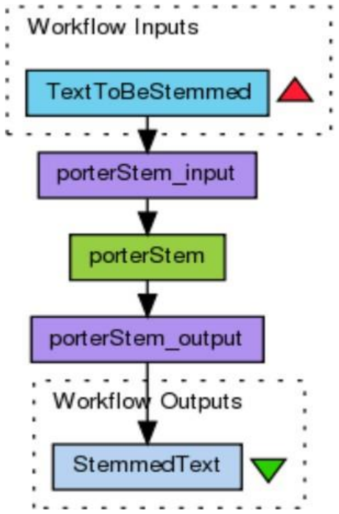
I. Regexp Tokenizer:

A tokenizer that splits a text into its component tokens and separators using a regular expression.



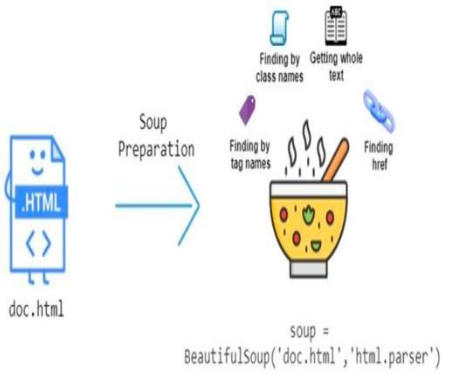
II. Snowball Stemmer:

Snowball is a string processing language that returns root words.
It is a stemming algorithm that is also known as the Porter2 stemming method since it is an improved version of the Porter Stemmer that has several faults solved in it.
Ada, java , java script ,go ,C#, Object Pascal, rust , and phyton are just a few of the languages that the Snowball compiler can translate.



III. Beautiful Soup:

It is used to extract data from HTML, XML, and other markup languages.
Using the Beautiful Soup library, extract only relevant hyperlinks for Google, such as links containing "a'>' tags with href attributes.
Make a Data frame out of the URLs.
Once you have a list of your websites with hyperlinks, convert it to a Pandas Data Frame with columns "from" (the URL where the link is located) and "to" (the link destination URL).



IV. Logistic Regression

In logistic regression, the dependent variable is a binary variable that stores data that can be represented as 1 or 0 1 = True and 0 = False

The logistic regression model, in other words, predicts P(Y=1) as a function of X.

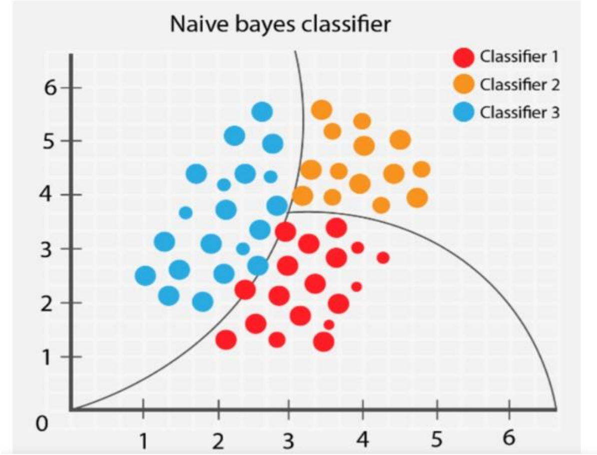
ML uses the classification method of logistic regression to calculate the probability of a dependent variable that is categorical.

V. MultinomialNB:

NLP Problems Using Multinomial Naive Bayes.

In NLP problems, the MNB algorithm is often used for tasks such as text classification, sentiment analysis, and spam filtering. It works by building a probability model for each class based on the frequency of each word in the training data. During the classification phase, the algorithm calculates the probability of each class given the input text and selects the class with the highest probability as the predicted class.

One of the advantages of the MNB algorithm is that it is simple and computationally efficient, which makes it well-suited for large-scale NLP problems. However, the "naive" assumption of independence between features may not always hold in real-world applications, which can lead to suboptimal performance. Nonetheless, MNB is a widely used algorithm in NLP and can serve as a good baseline model for many text classification tasks.



4. INNOVATIVE IDEA INTRODUCED IN DESIGN:

In this project we used a app called fastapi.

We can use this API by importing a library called uvicorn.

The API is connected to pishing detection model. In order to serialise objects to files on disc and then deserialize them back into the programme at run time, a PKL file must be created.

5. CONCLUSION :

This study uses machine learning technologies to improve the identification of phishing websites. A 98% accuracy rate is obtained. For a system to be able to identify a malicious URL, it must have extremely significant data. According on the above models' results, logistic regression performs the best. Therefore, we may draw the conclusion that logistic regression has a greater accuracy value than other methods.

REFERENCES :

Pujara, Purvi, and M. B. Chaudhari. "Phishing website detection using machine learning: a review." *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 3.7 (2018): 395-399.

Mahajan, Rishikesh, and Irfan Siddavatam. "Phishing website detection using machine learning algorithms." *International Journal of Computer Applications* 181.23 (2018): 45-47.

Kulkarni, Arun D., and Leonard L. Brown III. "Phishing websites detection using machine learning." (2019).

Kiruthiga, R., and D. Akila. "Phishing websites detection using machine learning." *International Journal of Recent Technology and Engineering* 8.2 (2019): 111-114.

Kumar, J., Santhanavijayan, A., Janet, B., Rajendran, B., & Bindhumadhava, B. S. (2020, January). Phishing website classification and detection using machine learning. In *2020 international conference on computer communication and informatics (iccci)* (pp. 1-6). IEEE.