INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 05 ISSUE: 03 | MARCH - 2021 ISSN: 2582-3930

MALICIOUS WEBSITE DETECTION USING MACHINE LEARNING

[1.] Dhanalakshmi.S, [2.] MuppudathiAishwarya.B,
[3.] A.Shanthakumari, [4.] Dr.P.Veeralakshmi
[1,2]Students, Prince Shri Venkateshwara Padmavathy Engineering College
[3,4]Faculty, Prince Shri Venkateshwara Padmavathy Engineering College

Abstract— Now-a-days people are using digital technologies which have advanced more rapidly than any innovation. Machine learning algorithm is a part of an Artificial Intelligence (AI). Machine learning algorithm build a sample data, known as training data, in order to make prediction or decision without being explicitly programmed to do so. Major problem facing in machine learning is lack of good data. Data quality is essential for the algorithm to function as intended. The phishing website can be detected based on some important characteristics like URL and Domain Identity, and security and encryption criteria in the final phishing detection rate. In order to detect and predict phishing website, we proposed an intelligent, flexible and effective system that is based on using Machine learning technique. We implemented classification algorithm and regression techniques to extract the phishing data sets criteria to classify their legitimacy. Here we use two machine learning techniques there are Logistic regression and Decision tree algorithm. Logistic regression is a go to method for binary classification problem. It is another technique borrowed from the field of statistics. Logistic regression gives 95% of accuracy for trained data sets. Whereas Decision tree are a type of Supervised Machine Learning. Where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split. Which gives 85% accuracy for trained data.

1 INTRODUCTION

Machine learning is a very hot topic for many key reasons, and it provides the ability to automatically obtain deep insights, recognize unknown patterns, and create high performing predictive models from data, all without requiring explicit programming instructions. This high level understanding is critical if ever involved in a decision-making process surrounding the usage of machine learning, how it can help achieve business and project goals, which machine learning techniques to use, potential pitfalls, and how to interpret the results. Machine learning is the application of artificial intelligence and based on the idea of the system that will learn data with less human intervention.

Supervised Learning:

Supervised learning in ML is the task of learning a function which maps an input to an output that is based n the sample input-output pairs. It also refers a function from training data which contain a set of training samples. Supervised Learning is classified into two types namely 1. Classification 2. Regression. It works with or learns with labeled data that implies some data which is already tagged with the correct answer. It also allows collection of data and to produce data output from the previous experience. It also helps to improve the performance learning and training needs a lot of computation skills.

Unsupervised Learning:

Unsupervised learning in ML is a process where the user need not to supervise the model. To discover patterns and information, it allows the model to work on its own that was previously detected. It mainly cope-up with unlabeled data. UL algorithm users are allowed to perform more complex processing tasks then supervised learning. Also it is unpredictable the any other natural learning methods. Unsupervised Learning algorithms has clustering, anomaly detection and neural networks. Storing of precise information is difficult in unsupervised learning. It gives less accurate values as the input is not labeled data.

Reinforcement Learning:

Reinforcement learning in ML to make a sequence of decisions. It mainly used for game-like situation. To get the solution of the problem the computer employs the trial and error method. The reward-policy is also set as a rule in case of game by the designer. If reinforcement learning algorithm is run on a sufficiently powerful computer infrastructure then the artificial intelligence can gather experience from thousands of parallel game plays.

APPLICATIONS OF MACHINE LEARNING

As we move forward into the digital age, one of the modern innovations we've seen is the creation of **Machine Learning**. This incredible form of artificial intelligence is already being used in various industries and professions.

Image Recognition:

Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, Automatic friend tagging suggestion:

Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook

friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's face detection and recognition algorithm.

Speech Recognition:

While using Google, we get an option of "Search by voice," it comes under speech recognition, and it's a popular application of machine learning.

Speech recognition is a process of converting voice instructions into text, and it is also known as "Speech to text", or "Computer speech recognition." At present, machine learning algorithms are widely used by various applications of speech recognition. Google assistant, Siri, Cortana, and Alexa are using speech recognition technology to follow the voice instructions.

Traffic prediction:

If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions. It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways:

- Real Time location of the vehicle form Google Map app and sensors
- Average time has taken on past days at the same time.

Medical Diagnosis

In medical science, machine learning is used for diseases diagnoses. With this, medical technology is growing very fast and able to build 3D models that can predict the exact position of lesions in the brain. It helps in finding brain tumors and other brain-related diseases easily.

RELATED WORKS:

In this [1] paper, AI Meta-Learners and Extra-Trees Algorithm for the Detection of phishing websites which is developed as the artificial intelligence schemes have been the cornerstone of modern counter measures used for mitigating phishing attacks. Here the cyber security refers to the management and development to the technologies tools and techniques required for protecting the data, devices and information. The proposed AI-based Meta-Learners were fitted on a phishing website datasets and their performances were evaluated. The models achieved a detection accuracy not lower than 97% with a drastically low false-positive rate of not more0.028. Hence, we recommend the adaptation of meta-learners when building phishing attack detection models. Here the false positive is detected so, to overcome these we just include the security purpose for the detection for phishing methods and to secure the higher accuracy rate based on the algorithm.

In this [2], The Particle Swarm optimization –Based feature Weighting for improving intelligent phishing website detection ,Over the last few years website phishing attacks have been constantly evolving causing customers to lose trust in e-commerce and online services. Various tools and systems based on the backlist of phishing websites are applied to detect the phishing websites. In this paper intelligent phishing website detection using Particle Swarm Optimization is proposed to enhance the detection of phishing websites.

In this [3], Phishing website Detection Based on Multi-dimensional Features Driven by Deep Learning, Phishing is currently a feature threat facing the internet, and losses due to phishing are growing steadily. Feature engineering is very important in phishing website detection solutions, but the accuracy of detection critically depends on prior knowledge of features. This approach can reduce the detection time for setting a threshold. Testing on dataset containing millions of phishing URLs and legitimate URLs, the accuracy reaches 98.99%, and the false rate is only 0.59%. By adjusting the threshold, the experimental results show that the detection efficiency can be improved.

In this [4], Systematization of Knowledge (SOK): A Systematic Review of Software Based Web Phishing Detection, Here the main purpose of Phishing is a form of cyber- attack that leverages social engineering approaches to harvest personal information from users of website. Here the HTML Phish is not using any manual feature extraction.

In this [5], Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity, This involves social networks that have become one of the most popular platforms for users to interact with each other. It has given the huge amount of sensitive data available in social network platforms, user privacy protection on social networks has become one of the most urgent research issues. Our large-scale evaluation using real-world websites shows the effectiveness of our approach. The proof of concept implementation verifies the correctness and accuracy of our approach with relatively a low performance overhead.

PROBLEM DESCRIPTION:

Now-a-days digitalization become an essential part in our daily life, it's the base of banking transactions, shopping, entertainment, resource sharing, and social networking. [1]The majority of malware is intended to steal the user's private data by using any website. In order to avoid this, detection process is done to check whether the given website is phishing or legitimate. In this digital world, most of the people are using online websites so this system is very useful and safe for online transactions and sharing information etc. This process develops our surroundings with safe and secure manner. [2][3] Detecting phishing webpages is an essential task that protects legitimate websites and their users from various malicious activities.

To classify the suspect webpage as phishing or legitimate, robust and effective features used for classification are in demand. However, recent phishing attacks usually make phishing webpages resemble the legitimate webpages in visual and functional aspects [4].Phishing is a concrete, widespread threat that combines social engineering with website spoofing. It leads to various malicious activities, including identity theft, financial gain, unauthorized account access, credit card fraud, etc. This threat causes not only tremendous financial losses to Internet users, but also long term reputation damage to the legitimate websites targeted by phishing scams.

METHODOLOGY USED:

LOGISTIC REGRESSION:

We are using a logistic regression algorithm and decision tree algorithm. By using these algorithms we analyze the accuracy for detection [4]. Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets [5] [6].

The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1. [7][8][9] The trained datasets which is preprocessed and it is stored into the database and which is compared with the given input datasets by the logistic regression algorithm and decision tree algorithm. The decision tree algorithm is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. [1][10][11]It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

SIGMOID FUNCTION:

$$f(x) = \frac{1}{1+e^{-(x)}}$$

DECISION TREE:

A decision tree is a flowchart-like tree structure where an internal node represents feature, the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. [12][13]In this, we compare these algorithms based on their accuracy level for the website detection. Based on the higher accuracy we can conclude the algorithm to find the website whether it is a phishing or legitimate. This gives a user to find the better identity to find to enable future prediction

WORKING:

1. Select the best attribute using Attribute Selection Measures (ASM) to split the records.
2. Make that attribute a decision node and breaks the dataset into smaller subsets.
3. Starts tree building by repeating this process recursively for each child until one of the condition will match:
    - All tuples belong to the same attribute value.
    - There are no more remaining attributes.
    - There are no more instances.

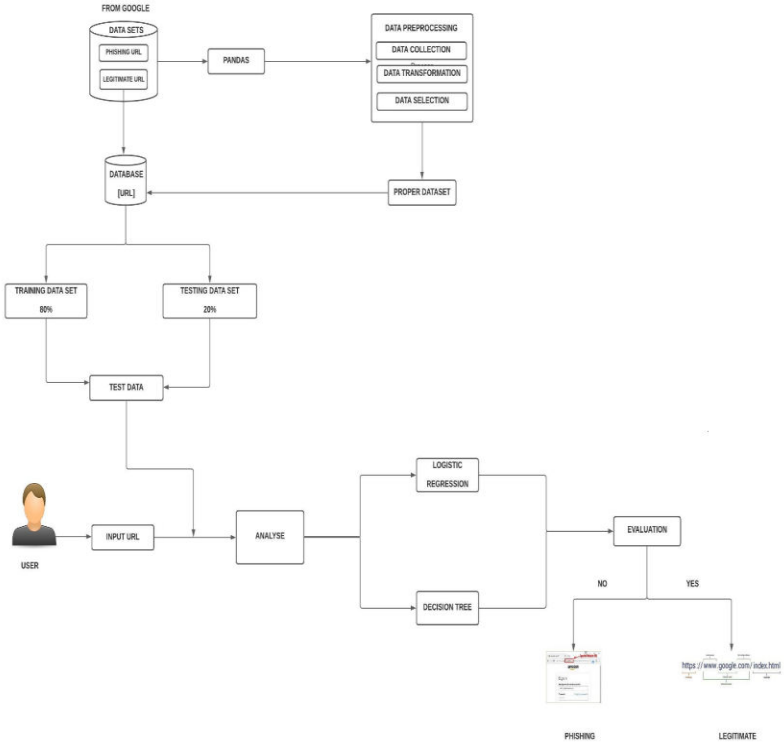


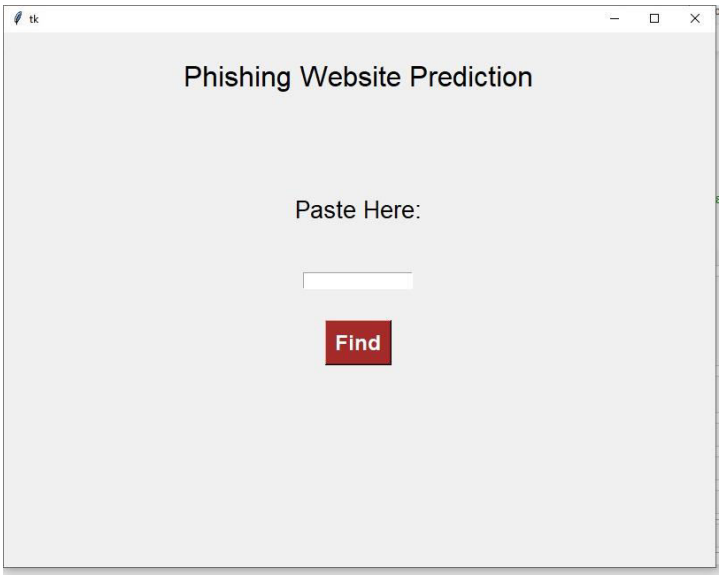
Figure 1.1

In fig 1.1, we collect data sets from google, the collected data can be phishing URL or legitimate URL. After that we import a library pandas, which is used to directly read the csv.file (datasets). Data sets are then preprocessed and it is stored in the system database itself. Next the datasets are then goes for training (80%) and testing (20%) process by using both logistic regression algorithm and decision tree algorithm. Keep on training the datasets only we can get a higher accuracy. Further the user gives an input URL, which is then analyze with the test data. Then it is evaluated with both the algorithms.

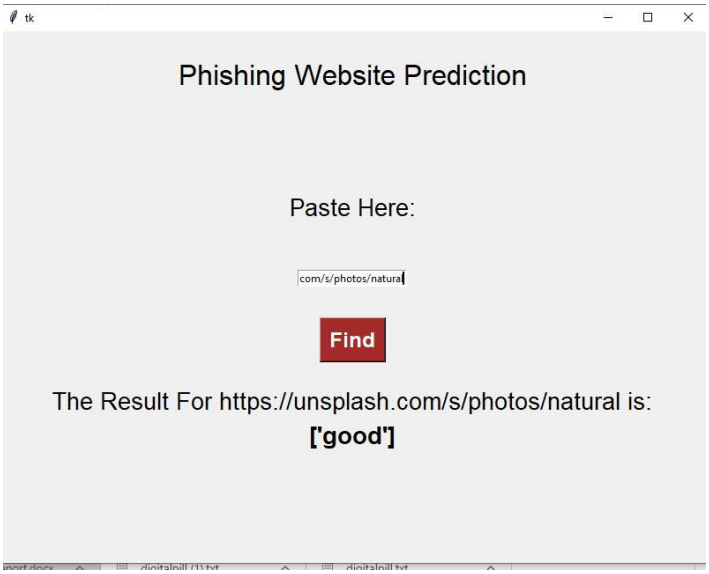
After the evaluation process, we can identify any websites whether it is a phishing website or legitimate website. If it is a phishing website the given URL website shows as good and if it is a legitimate the URL shows as bad.

IMPLEMENTATION AND RESULT:

In this phishing website prediction user can easily identify which URL is phishing or legitimate, the given URL is then divided using tokenization process. Tokenization process is to divide the special characters like @, #, $, %, & etc. Length of the URL is also very useful for find whether the given URL is phishing or legitimate.



Here the Fig 1.1.2 is good URL website so it is identified as a legitimate website.



LEGITIMATE WEBSITE:

FIGURE 1.1.2

CONCLUSION:

Phishing is a way to obtain user's private information through email or website, as the usage of internet is very fast almost all things are available in the online. So as the technology increases phishing attacks use new methods day by day, so to avoid the problem of privacy, it is enabled through online to identify the phishing websites for the detection. In this proposed system, we performed a detailed information in problem description about the detecting the phishing websites in machine learning based on the algorithm detection. Hence by using this logistic regression algorithm of about 98% of accuracy detection this is the best suitable approach than other.

Even though our proposed features present high classification power against state of the phishing practice, they could be stale and ineffective someday as a result of the evolution of phishing ecosystem. Hence constant analysis and update of the features is required to maintain the high detection power of the system.

FUTURE ENHANCEMENTS

A future work will force on collecting phishing and non-phishing websites that are currently accessible in the WWW and extract a list of features that are different from the one commonly used in phishing detection. This work of the proposed system is to evaluate these machine learning classifiers with layer dataset. We already have classifiers which gives good prediction rate of the phishing website, but after our survey that it will be better to use a hybrid approach for the prediction rate of phishing website. In future if we get structured dataset of phishing we can perform phishing detection much faster than any other technique, also we can use a combination of any other two or more classifier for getting maximum accuracy .In particular we extract features from URL'S and pass it through the various classifier.

REFERENCES:

[1] YAZAN AHMAD ALSARIERA [1], VICTOR ELIJAH ADEYEMO [2], ABDULLATEEF OLUWAGBEMIGA BALOGUN [3] ,[4], (Member, IEEE), AND AMMAR KAREEM ALAZZAWI [3] "AI Meta-Learners and Extra-Trees Algorithm for the Detection of Phishing Websites", 1Department of Computer Science, Faculty of Science, Northern Border University, Arar 73222, Saudi Arabia ,2School of Built Environment, Engineering, and Computing, Leeds Beckett University, Leeds LS6 3QS, U.K, 3Department of Computer and Information Sciences, Faculty of Science and IT, University Technology PETRONAS, Seri Iskandar 32610, Malaysia ,4Department of Computer Science, Faculty of Communication and Information Sciences, University

of Ilorin, Ilorin 1515, Nigeria, 2020.

[2] WALEED ALI AND SHARAF MALEBARY, (Member, IEEE), "Particle Swarm Optimization-Based Feature Weighting for Improving Intelligent Phishing Website Detection", the Deanship of Scientific Research (DSR), King Abdul-Aziz University, Jeddah, under Grant No. (DF-438-830-1441), Digital Object Identifier 10.1109/ACCESS.2020.3013699, 2020.

[3] PENG YANG, GUANGZHEN ZHAO, AND PENG ZENG, "Phishing Website Detection Based on
Multidimensional Features Driven by Deep Learning", the National Natural Science Foundation of China under Grant 61472080 and Grant 61672155, in part by the Consulting Project of Chinese Academy of Engineering under Grant 2018-XY-07, and in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization, 2019.

[4] ZUOCHAO DUO, Student Member, IEEE, ISSA KHALIL Member, IEEE, ABDALLAH KHREISHAH, Member, IEEE, ALA Al-FUQAHA, Senior Member, IEEE and MOHSEN GUIZANI, Fellow, IEEE, "Systematization of Knowledge (SOK): A Systematic of Software Based Web Phishing Detection", Citation information, 2017.

[5] JIAN MAO[1], (Member, IEEE), WENQIAN TIAN, PEI LI1, TAO WEI[2], (Member, IEEE), AND ZHENKAI LIANG[3], (Member, IEEE), "Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity", the National Natural Science Foundation of China under Grant 61402029, in part by the National
Natural Science Foundation of China under Grant 61370190 and Grant 61379002, and in part by the Singapore Ministry of Education, NUS, under Grant R-252-000-539-112, 2017.

[6]A. BELABED, E. AIMEUR, and A.CHIKH, "A personalized whitelist approach for phishing webpage detection," in Proc. 7th Int. Conf. Availability, Rel. Security (ARES), Aug. 2012, pp. 249–254.

[7]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.

[8] T.-C.CHEN, S.DICK, and J.MILLER, "Detecting visually similar Web pages: Application to phishing detection, "ACM Trans. Internet Technology", vol. 10, no. 2, pp. 1–38, May 2010.

[9] N.CHOU, R.LEDSHMA, Y.TERAGUCHI, D.BONEY and J. C.MITCHELL, "Client side defense against Web-based identity theft", in Proc. 11th Annu. Network Distribution System, Security Symp. (NDSS), 2004, pp. 1–16

[10] C. Inc. (Aug. 2016). Could mark Toolbar. [Online]. Available: http://www.cloudmark.com/desktop/ie-toolbar

[11] J.CORBETT, L.INVERNIZZI, C.KRUEGEL, and GVIGNA, "Eyes of a human, eyes of a program: Leveraging different views of the Web for analysis and detection", in Proceedings of Research in Attacks, Intrusions and Defenses (RAID). Gothenburg, Sweden: Springer, 2014.

[12] X.DENG, GHUANG, and A. Y.FU, "An antiphishing strategy based on visual similarity assessment", Internet Computer, vol. 10, no. 2, pp. 58–65, 2006.

[13] Z.DONG, K.KANE, and L. J.CAMP, "Phishing in smooth waters: The state of banking certificates in the US", in Proc. Res. Conf. Common., Inf. Internet Policy (TPRC), 2014, p. 16.