JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

RESEARCH PAPER ON DETECTION OF URL BASED PHISHING WEBSITES USING MACHINE LEARNING IN DJANGO FRAMEWORK

Khushboo Kumari
MCA Scholar
Department of Master of Computer Applications
School of CS & IT
JAIN(Deemed-to-be-University)

Feon Jaison
Assistant Professor
Department of Master of Computer Applications
School of CS & IT
JAIN(Deemed-to-be-University)

Abstract- In this modern world, Phishing website is one of the most dangerous things in the world. In the recent times, a lot of people have suffered phishing attack due to phishing website as it is a simplest way to obtain sensitive information from innocent users. Machine Learning plays an important role in prediction to detect the phishing website in the website. The proposed method predicts the URL based phishing websites based on features and also gives maximum accuracy to give the result. The proposed method will use uniform resource locator (URL) features to detect the phishing website. The proposed method takes those features to detect the phishing website. The Security for phishing detection website is a major concern which is solved by providing administration who can manage the phishing detection website and user who can check the phishing website.

Keywords- Phishing site, Machine learning, Security, Legitimate, Prediction, Logistic regression, MultinomialNB

I. INTRODUCTION

The work proposed in this model focus on detection of those phishing website which is based on URL in machine learning. These phishing websites are made to look like an original organization website. The attacker uses their skills to manipulate user to fill up the personal information by giving urgent messages or need to update or loose money etc so that they fill the required information which can be used by them to misuse it and gain access to there account. They make the circumstances such that the user is not able to think twice and think that they have not any other option but to visit their fake website. Machine learning has been widely used in many areas to create automated solutions to predict the result. The phishing attacks can be carried out in many ways such as email, website, sending message. Similarly, which are labelled as legitimate will be detected as legitimate URL and those who contain any phishing activity then detected as Alter. The dataset with bad phishing website is to be used for machine learning must consist these features. There are many machine learning algorithms and each algorithm has its own working mechanism.

The algorithm used are logistic regression, multinomial NB, super vector machine to predict the phishing site. For security of the phishing detection website the administration and password

are used to protect the detection website. The dataset is also stored in the cloud so that it can be integrated.

II. LITERATURE REVIEW

Rishikesh Mahajan et.al [1] proposed paper that are based on machine learning technology for detection of phishing URLs by extracting and analysing various features of legitimate and phishing URLs. To detect phishing websites Decision Tree, random forest and Support vector machine algorithms are used. To improve the extracting, we used some more algorithm.

Jain A.K et.al [2] introduced a URL-based anti-phishing machine learning method. They took 14 features of the URL to detect the website. They differentiate between phishing website and valid URLs by using Support Vector Machine (SVM) and Naïve Bayes (NB) classifiers. They extracted 14 different features, which were used to differentiate between phishing websites and legitimate websites. The result of his experiment is 90% of precision websites when it is done with SVM Classification.

Purbay M., and Kumar D [3] proposed multiple ML methods to detect URLs by analysing various URL components using machine learning. Authors introduced various methods of supervised learning for the identification of phishing URLs based on PageRank, traffic rank information and page importance properties. They studied how the volume of different training data influences the accuracy of classifiers. The research includes Support Vector Machine (SVM), K-NN, random forest classification (RFC) techniques for the classification.

Y. Sönmez et.al [4] proposes a classification mode in order to classify the phishing attacks. This model contains of feature extraction from sites and classification of website. In feature extraction, 30 features have been taken from UCI Irvine machine learning repository data set and phishing feature extraction rules has been clearly defined. In order to classification of these features, they used Support Vector Machine (SVM), Naïve Bayes (NB). In Extreme Learning Machine (ELM), six activation functions were used and achieved 95.34%accuracy than SVM and NB. The results were obtained with the help of MATLAB.

D. Akila et.al [5] focuses on detecting phishing website URLs with domain name features. Web spoofing attack categories and content-based, heuristic-based approaches are explained and the proposed model Phish Checker is developed with the help of Microsoft Visual Studio Express 2013 and C# language. Dataset used from Phish tank and Yahoo directory set and obtained an accuracy of 96%. This paper checks only the validity of URLs.

By analysing all the papers, we proposed the feature with the security in the phishing detection website as an administration with password and users with password feature for admin to login and admin will manage the number of users which is signed in and manage to read the feedback from the user if there provide any feedback relate to login or website detection.

In the user page, they can login with username and password and can check phishing website and provide feedback based to phishing website.

The prediction of phishing website has found that system provides us with 95 % of accuracy for Multinomial NB and 97 % of accuracy for Logical Regression. The prediction is done between legitimate and phishing website and training accuracy is 97% and testing accuracy is 95%.

Author	Models Used	Accuracy Rate %	Contribution	Limitation
Rishikesh Mahajan et.al [1]	Decision tree	96.71	Dataset is divided into training set and testing set in 50:50, 70:30 and 90:10 ratios	hybrid technology does not implement to detect phishing websites more accurately.
	Random tree	96.72		
	Support vector machine	96.40		
Jain A.K. et.al [2]	Naive Bayes	95	Employed both NB and SVM algorithms to identify the malicious websites.	Both SVM and NB are slow learners and does not store the previous results in the memory. Thus, the efficiency of the URL detector may be reduced.
	Support Vector Machine	90		
	Support vector machine	95.35		
	Random tree	95		
Purbay M., and Kumar D [3]	Random tree	95.78	Utilized multiple ML methods for classifying URLs.	They compared the performance of different types of ML methods. However, there were no discussions about the retrieval capacity of the algorithms.
	Support Vector Machine	94.22		
	K-NN	93.57		
Y. Sönmez et.al [4]	Support Vector Machine (SVM)	95.34	a classification mode in order to classify the phishing attacks.	This model comprises of feature extraction from sites and classification of website.
	Naive Bayes (NB)	93.45		
	Extreme Learning Machine (ELM)	92.34		
D. Akila et.al [5]	Phish Checker	96.0	on detecting phishing website URLs with domain name features. Web spoofing attack categories and content-based, heuristic-based approaches	It is written in C# language and This paper checks only the validity of URLs

III. PROPOSED METHODOLOGIES

To examine a model accurately predict phishing website. We experimented with different classification algorithms and ensembles to predict phishing website. Next, we briefly analyse each phase.

A. Dataset Description - The data is collected from the Kaggle dataset with the name of phishing_site_urls. The security to the dataset is given by saving dataset in the cloud and retrieving by using access key and secret key.

The record has 2 attributes:

a. URL
b. Label (Good/bad)

B. Data Pre-processing - Now that we have the data, we have to vectorize our URLs. We used Count Vectorizer and gather words using tokenizer, since there are words in URLs that are more important than other words e.g., 'virus', '.exe','.dat' etc.

For this we used Regexp Tokenizer - A tokenizer that splits a string using a regular expression, which matches either the tokens or the separators between tokens.

C. Using Machine learning – Now we classifying models which are used to predict phishing URL are Decision Tree, Random Forest, Support vector machine, Logistic Regression and multinomial NB.

i. Decision Tree Algorithm

One of the mostly used algorithm in machine learning technology is Decision tree as it is easy to understand and easy to implement. Decision tree is a Supervised learning technique but mostly it is used for solving Classification problems. The decision tree has a model based on a tree-like structure that describes the classification process based on the input function. Input variables can be of any type, such as B. graphic, text, discrete, continuous, etc

ii. Random Forest Algorithm

A random forest is a machine learning technique which is used to solve regression and classification problems. Random Forest is developed by Leo Bremen. It is done by building a large number of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual tree. It gives information gain methods to find the best splitter. This process will get continue until random forest creates n number of trees. Every tree in the forest predicts the target value and then calculating the algorithm to vote for each predicted target. At last random forest algorithm considers high voted predicted target as a final prediction and give result.

iii. Support Vector Machine Algorithm

Support vector machine is another powerful algorithm in machine learning technology to create the best line or decision boundary that can segregate n-dimensional space into classes so that it will be easy to put the new data point in the correct category in the future. It seeks for the closest points which is known as support vectors and once it finds out the closest point it constructs a line connecting to them. This machine then draws separating line which bisects and perpendicular to the connecting line. Aiming to classify the data perfectly with the margin should be at maximum. Now the margin is defined as a distance between hyperplane and support vectors. In real scenario it is not possible to separate complex and non-linear data, to solve this problem support vector machine uses kernel trick which transforms lower dimensional space to higher dimensional space.

iv. Logistic regression

Logistic regression is a supervised learning classification algorithm which is used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes to find the result.

It means, the dependent variable should be binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, data prediction etc.

v. MultinomialNB

Multinomial Naive Bayes algorithm is used for classification with many features. The algorithm is centered on Bayes theorem and predicts with help of the tag of a text such as a piece of email or article. It is calculating the probability of every tag which is given in the sample and then provides the tag with the highest probability as output.

Naive Bayes classifier is a collection of many algorithms in which all the algorithms share one common principle which is every feature being classified is not related to any different feature.

Bayes theorem, formulated by Thomas Bayes, calculates the probability of an event occurring based on the prior knowledge of conditions related to an event. It is based on the following formula:

P(A|B) = P(A) * P(B|A)/P(B)

the probability of class A when predictor B are:

P(B) = prior probability of class B

P(A) = prior probability of class A

P(B|A) = occurrence of predictor B given class A probability.

D. After having result by using ML, we have provided a GUI with administration and user so that user can use detection website securely. Admin will manage all the users and detection site and manage dataset.

IV. SCHEME DESIGN

The Design for the phishing website detection Diagrams are typically brought into the research process to know the procedure how the project work. It is a common place to see a diagram illustrating how concepts or themes relate to each other or to explain how the research data relates to an user.

The system architecture is used to see how the system will look and to describes the process and relations of a number of elements that together create a defined output.

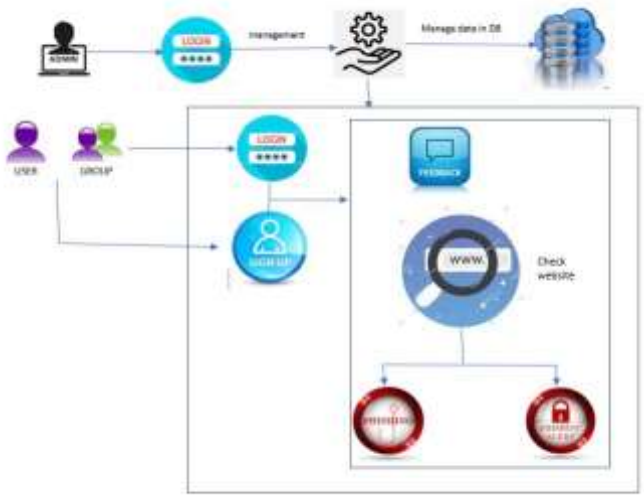

Fig 1: System Architecture

A data flow model is a diagram representation of the data and exchange of information within a system. Data flow models are used to graphically represent the flow of data in an information system by describing the procedure involved in the system from data input to predicting result process.

One of the dataflow diagrams is the graphical user interface to understand the process of the system looks and another is to understand working of the phishing detection procedure, how the ML works in the prediction.

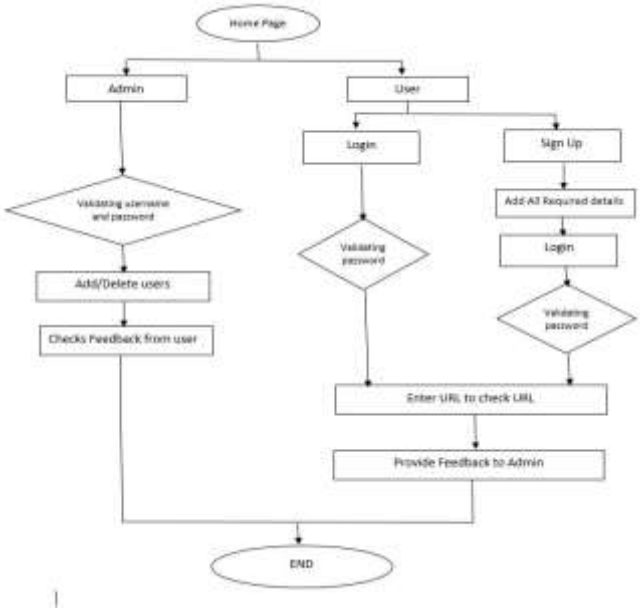

Fig 2: Dataflow of GUI system

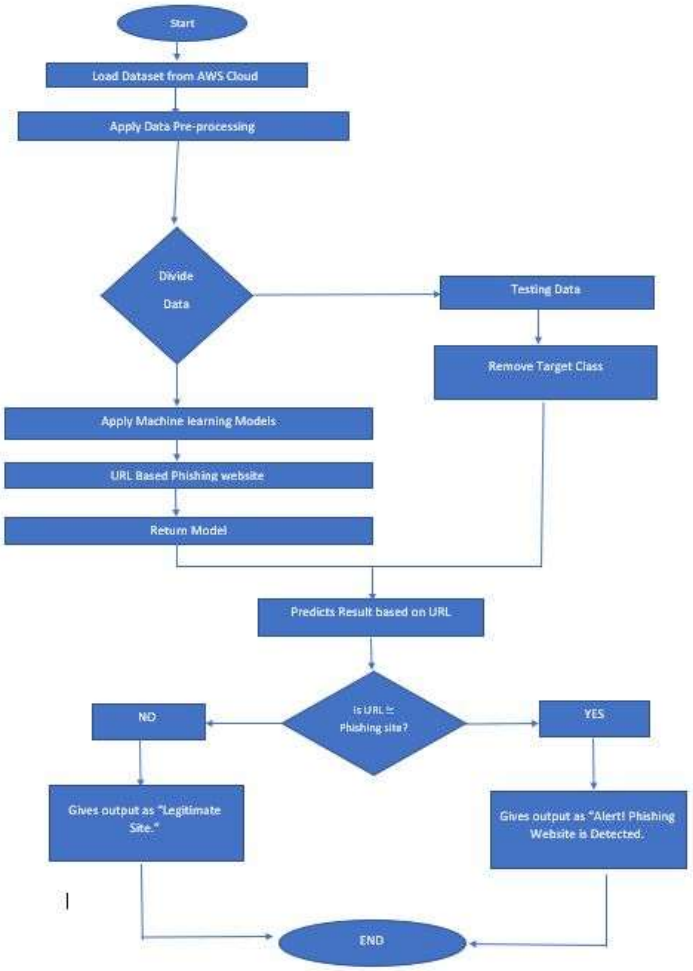

Fig 3: Dataflow of Phishing Detection

V. IMPLEMENTATION AND RESULT

Since the subsequent sections in this survey will compare a number of detection technique, we found matrix with comparison of techniques to use in phishing detection. Based on the review of the literature, the following are the mostly commonly used evaluation matrices.

•True Positive (TP) rate — measures the rate of correctly of phishing attack

•False Positive (F P) rate — measures the rate of legitimate website.

•True Negative (T N) rate — measures the rate of correctly legitimate website

•False Negative (FN) rate — measures the rate of phishing attack

• Precision (P) — measures the rate of correctly detected phishing attack in relation to all instance that were detected as phishing.

•Accuracy- measures the overall rate of correctly detected phishing site.

It has been used to import Machine learning algorithms. Dataset are divided into training set and testing set.

The prediction of phishing website has found that system provides us with 95 % of accuracy for MultinomialNB and 96 % of accuracy for Logical Regression.

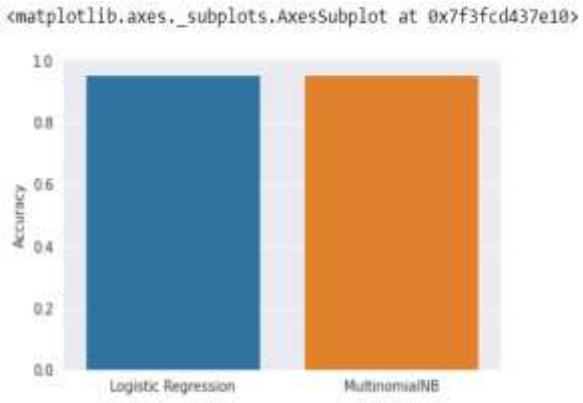

Fig 4: Graph for the accuracy both ML are providing

The classification report gives the training set and testing set report with accuracy percentage of the ML, which we used to do prediction of the phishing detection.

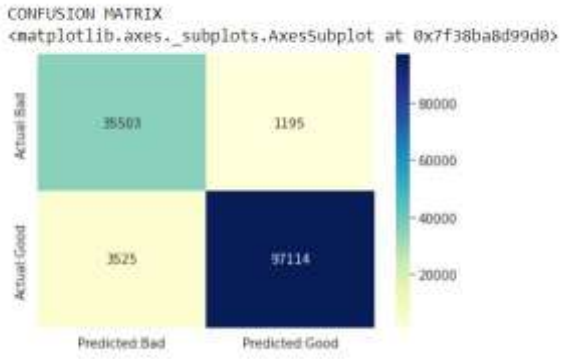

Fig 5: Confusion matrix that defined the accuracy of result

Accuracy by the logistic regression:

Training Accuracy: 0.9784859068612579

Testing Accuracy: 0.964284934140108

Accuracy by the MultinomialNB:

Training Accuracy: 0.9740175578688816

Testing Accuracy: 0.9585326605357624

The security which needs to give to the admin for the phishing website detection and admin knows the number of users whom have registered and can read the feedback which was given by user.

User can register themselves by providing username and password. They can check phishing website by providing URL and based on the result they can provide feedback to the admin.

VI. CONCLUSION

In the modern world, the phishing detection is a critical task and it is important for us to get a method to detect it. It can be solved by using any of the machine learning algorithm with the classifiers. Classifiers which provide good prediction rate of the phishing website, but after our analyses it is good to use a hybrid approach such as Logistic regression and Multinomial NB for the prediction and it improve the accuracy prediction rate of phishing websites. Existing methodologies gives less secure as it only provides area to check the website so we proposed a new security concept with administration for URL based features and machine learning algorithms so that it can be manage. The user will use the website to detect phishing website by signing up and all these is written in Django framework.

VII. FUTURE SCOPE

In future if we get more structured dataset of phishing URLs where we can perform phishing detection much faster than any other technique. We can merge any other two or more classifier to get maximum accuracy. In particular, we abstract features that predicts URLs and predicts through the various classifiers. The GUI can also be enhanced by providing more features in the admin and user page.

VIII. REFERENCE

1. Rishikesh Mahajan, Irfan Siddavatam, “Phishing website detection using machine learning” International Journal of Computer Applications (0975 – 8887) Volume 181 – No. 23, October 2018
2. Jain A.K., Gupta B.B. “PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning”, Cyber Security. Advances in Intelligent Systems and Computing, vol. 729, 2018, Doi: 10.1007/978-981-10-8536-9_44
3. Purbay M., Kumar D, “Split Behaviour of Supervised Machine Learning Algorithms for Phishing URL Detection”, Lecture Notes in Electrical Engineering, vol. 683, 2021, Doi: 10.1007/978-981-15-6840-4_40 [CrossRef] [Google Scholar]
4. Y. Sönmez, T. Tuncer, H. Gökal, and E. Avci, “Phishing web sites features classification based on extreme learning machine,” 6th Int. Symp. Digit. Forensic Secure. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.
5. R. Kiruthiga, D. Akila,” Phishing Websites Detection Using Machine Learning” International Journal of Recent Technology and Engineering (IJRTE) ISSN:2277-3878, Volume-8, Issue -2S11, September 2019
6. Hodžić, A., Kevrić, J., & Karadag, A. (2016). Comparison of machine learning techniques in phishing website classification. In International Conference on Economic and Social Studies (ICESoS'16) (pp. 249-256). PMCID: PMC8504731 PMID: 34634081
7. Hung Le, Quang Pham, Doyen Sahoo, and Steven C.H. Hoi, “URL Net: Learning a URL Representation with Deep Learning for Malicious URL Detection”, Conference’17, Washington, DC, USA, arXiv:1802.03162, July 2017.
8. J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran and B. S. Bindhumadhava, “Phishing Website Classification and Detection Using Machine Learning,” 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1–6, 10.1109/ICCCI48352.2020.9104161.
9. Wu CY, Kuo CC, Yang CS,” A phishing detection system based on machine learning” In: 2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA), pp 28–32, 2019.
10. Rao RS, Pais AR. Jail-Phish: An improved search engine-based phishing detection system. Computers & Security. 2019. Jun

1;83:246–67. [Google Scholar]

11. Aljofey A, Jiang Q, Qu Q, Huang M, Niyigena JP. An effective phishing detection model based on character level convolutional neural network from URL. Electronics. 2020. Sep;9(9):1514. [Google Scholar]
12. Praisy Evangelin A1, Jeenath Laila N: PRIVACY PROTECTION OF USER BROWSING DETAILS AND UNSAFE URL DETECTION. International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 08 Issue: 03 | Mar 2021
13. Ashit Kumar Dutta," Detecting phishing Website using machine learning technique" Published online 2021 Oct 11. Doi: 10.1371/journal.pone.0258361
14. Neda Abdelhamid, Fadi Thabtah and Hussein Abdel-jaber, Phishing Detection: A Recent Intelligent Machine Learning Comparison based on Models Content and Features, IEEE Int. Conf. on Intelligence and Security Informatics (ISI), pages 7277, 2017
15. Ashritha Jain R, Mrs. Mangala Kini, Chaithra Kulal, Deekshitha S," A Review Paper on Detection of Phishing Websites using Machine Learning", Published Online 2019 June 13. Paper ID: IJERTCONV7IS08034 IJERT