Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018)
IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2

A new method for Detection of Phishing Websites: URL Detection

[1]Shraddha Parekh
(*Student*)
Department of Computer Engineering
K J Somaiya College of Engineering
Mumbai, India
shraddha.np@somaiya.edu

[2]Dhwanil Parikh
(*Student*)
Department of Computer Engineering
K J Somaiya College of Engineering
Mumbai, India
dhwanil.p@somaiya.edu

[3]Srushti Kotak
(*Student*)
Department of Computer Engineering
K J Somaiya College of Engineering
Mumbai, India
srushti.kotak@somaiya.edu

[4]Prof. Smita Sankhe
(*Assistant Professor*)
Department of Computer Engineering
K J Somaiya College of Engineering
Mumbai, India
smitasankhe@somaiya.edu

Abstract- **Phishing is an unlawful activity wherein people are misled into the wrong sites by using various fraudulent methods. The aim of these phishing websites is to confiscate personal information or other financial details for personal benefits or misuse. As technology advances, the phishing approaches used need to get progressed and there is a dire need for better security and better mechanisms to prevent as well as detect these phishing approaches. The primary focus of this paper is to put forth a model as a solution to detect phishing websites by using the URL detection method using Random Forest algorithm. There are 3 major phases such as Parsing, Heuristic Classification of data, Performance Analysis in this model and each phase makes use of a different technique or algorithm for processing of data to give better results.**

Keywords— URL detection, phishing, random forests classification, ROC curve, detection using Rstudio.

I. INTRODUCTION

Some years ago, since there wasn't any access or exposure to online procedures, online dealings or transactions, there was very minimal threat to the then already existing online systems. But, in the last 5 years, the world has seen a great boom in the IT sector which has led to most of the daily routine come online; right from shopping to bank transactions. The term 'Phishing' was coined in the 1996 timeframe by hackers who were stealing America On-Line accounts by scamming passwords from unsuspecting AOL users. [1] The word phishing comes from the phrase "website phishing" is a variation on the word "fishing". The idea is that bait is thrown out with the hopes that a user will grab it and bite into it just like the fish. In most cases, bait is either an e-mail or an instant messaging site, which will take the user to hostile phishing websites. Over the years, phishing attacks grew in number and intensity too. Phishing attacks now target users of online banking, payment services such as PayPal, and online e-commerce sites. There are different modes through which phishing can be carried out and hence there are various types of phishing like vishing (voice over phishing), smishing (Phishing via SMS), whaling, Mishing (mobile phishing), social engineering, spear phishing, etc. Usually, there are four phases in a typical phishing attack like preparation, mass broadcast, mature and account hijack.[2] For most of the phishing attacks, whether carried out by emails or any other medium, the objective is to get the victim to follow a link that appears to go to a legitimate web resource but actually redirects the victim to a malicious web page. The simplest approach to link manipulation is to create a malicious URL and directing the user to the desired malicious page that the attacker wants.

In this paper, we are focusing on detection of phishing websites by URL detection. In the earlier times, methods such as k-nearest neighbour, list-based approach, fuzzy logic, Phishzoo and other mining and classification approaches were used for detection but as the intensity of the attack grew over the times more sophisticated algorithms and techniques were introduced to detect and prevent the attack.

II. PREVIOUS WORK

Now, different journals, conferences have different studies and research for detection of phishing websites and one of the approaches that had been proposed was multi-tier classification for phishing URL filtering. In this, the authors put forward an innovative method for extracting of phishing URL features based on the weightage of message content [3]. A multiple classification algorithm is used which includes SVM, AdaBoost, and Naive Bayes. These algorithms are divided into three tiers using 21 fixed yet different features [3]. Then a two-step procedure takes place with the help of

another classification algorithm but the problem here is the time consumed and the complexity involved, overheads involved and the performance issues and hence this wasn't an optimal method.

Yet another method that was adopted by the authors of one of the IEEE 2017 papers was the pattern recognition i.e. different features are extracted from emails to obtain a model that helps to discriminate phishing messages from non-phishing ones [4]. One of the primary methods used in this context is to recognize the attacks and then make use of feature extraction and then classification. The main limitation of this proposal is that there are too many features that are evaluated without considering whether they really are essential to identify phishing. Therefore, it could lead to unnecessary computational cost.

According to Institute of Research Engineers and Doctors, USA[1], phishing detection techniques are divided into blacklist-based and heuristic-based approaches [5]. The blacklist-based approach maintains a database list of addresses (URLs) of those sites that are classified as malicious. If a user requests a site that is included in this list, the connection is blocked. The blacklist-based approach has the advantages of easy implementation and a low false- positive rate [5]; however, it has a flaw that it cannot detect phishing sites which are not listed in the database, including temporarily sites.

According to International Journal of Advanced research and innovative ideas in education(IJARIIE) journal paper, the Multi-Label Classifier based Associative Classification (MCAC) data mining approach is also one of the methods that is used for detecting phishing websites. The associative classification algorithm detects phishing websites with mediocre accuracy.[6] MCAC consists of three main steps which are Rule discovery, classifier building and class assignment. In the first step of this algorithm, rules are found and extracted by iterating over the training data set (historical websites features or data collected from various sources) .[6] In this step, merging of any of the resulting rules that have the same antecedent (left hand side) takes place and are linked with different classes to produce the multi-label rules. Along with this, redundant rules are eliminated. The outcome of the second step is the classifier which contains single and multi-label rules. The last step involves testing the classifier on testing data set to measure its performance. In the prediction process, the rule in the classifier group which matches the test data features often fired to guess its type (class). The MCAC algorithm generate rules further that rules are sorted by using sorting algorithm. [6]. The main problem that MCAC faced was difficulty in determining minimum confidence and minimum support when there is a large amount of data and later on there came more sophisticated algorithms to replace this which were more accurate and had lesser time complexity.

a. PROPOSED WORK

Our proposal for the above-mentioned topic is to improve the efficiency in detection of phishing websites. A lot of research work and survey was done to compare various classification algorithms to best fit our model[7]. Along with using WEKA to determine the accuracy and performance of each of the algorithms, lots of journals and papers had been read and surveyed to decide upon the classification algorithm. The idea that we are putting forward here is to improve the efficiency by using Random forests as our classification algorithm with the help of Rstudio tool that helps us in better analysis. Below is the flow diagram of our proposal:

Figure 1: Proposed Model

The dataset needed for the entire procedure was gathered from Phish tank[8]and since there was a large amount of data to process, parsing was performed primarily. Parsing is done to analyze feature set. We restrict our feature set to 8 out of the 31 features that are considered by parsing and by rigorous analysis and they are mentioned in Table 1. Here, parsing is done using Attribute Subset Selector which includes two parts 1) Attribute Subset Evaluator algorithm using Information Gain 2) Search Method algorithm using Ranker Method. Parsing is implemented using a java code which imports WEKA tool libraries for IG and Attribute selector. Those attributes that contribute more information will have a higher information gain value and can be selected, whereas those that do not add much information will have a

lower score and can be removed. Table 1 shows the features that we have taken into consideration in our model which help in classification.

TABLE 1: FEATURES

1	IP Address
2	Redirection of page using "//"
3	Adding Prefix or Suffix Separated by (-) to the Domain
4	Sub domain and multi-sub domain
5	URLs having @ symbol
6	Using different functions in the URL to submit information
7	Page Rank
8	Google Index

In weka, given the entropy is a criterion of impurity in a training set S, we can define a measure reflecting additional information about Y provided by X that represents the amount by which the entropy of Y decreases [9]. IG is given by the

formula: $IG = H(Y) - H(Y \backslash X) = H(X) - H(X \backslash Y)$ (1)

IG is a symmetrical measure. measure. The information gained about Y after observing X is equal to the information gained about X after observing Y.

After this, the second step is the classification for which Random forest algorithm is implemented using Rstudio. The parsed dataset undergoes heuristic classification where the dataset is spilt into 70% and 30%. The 70% data is considered for training and 30% for testing. Using the libraries of random forest and inbuilt R functions, the classification model is constructed, and this model is tested using testing data. Using this model, other URLs of different websites that are input by the user are predicted. The last phase in the model to be performed is Performance Analysis which was done using ROC Curve. Along with the ROC Curve, other factors such as sensitivity, confusion matrix, Fp Rate, etc. also form a part of performance analysis which help in better understanding. Fig 2(a), 2(b) and 2(c) show the metrics considered in performance analysis phase and displays the ROC Curve which helps in the better understanding of the accuracy or fallacy levels. Also, the three figures help us understand the outcome of the model and give a clear view about the results.

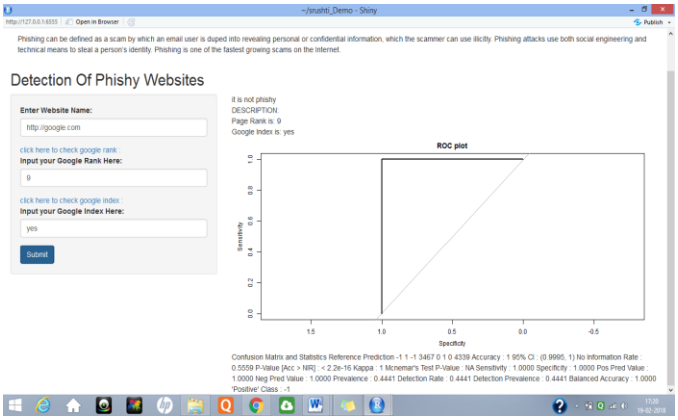

Fig 2(a)

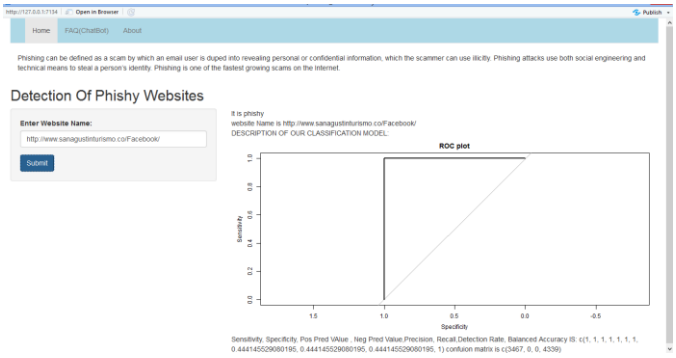

Fig 2(b)

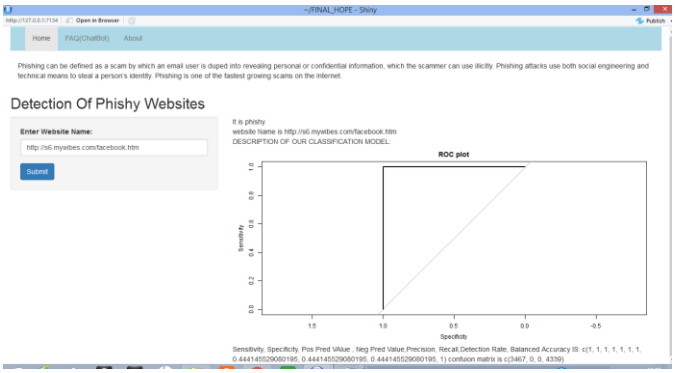

Fig 2(c)

III. CONCLUSION

In this paper, a different methodology has been proposed to detect phishing websites by using random forests

as the classification algorithm with the help of Rstudio. Here, we empirically demonstrated that the proposed features out of 31 of them are the most suitable for detection of phishing websites. The performance metrics along with our literature survey also proved the accuracy level of random forest to be the highest around 95% and thus Random Forests were chosen for classification [10]. This model has used a wide range of metrics, including true positives, true negatives, false negatives, the F-measure, ROC, precision, and sensitivity for analysis purposes thus giving a clear view on the performance and accuracy each time the detection takes place. There is no single solution to phishing till now and with the upcoming technology, the type and number of phishing attacks are expected to increase. For these, the browsers have to be made capable enough to setup methods that detect and warn of potential phishing attacks. Future work will aim to develop a system that can learn by itself about new types of phishing attacks by adding a more enhanced feature to the detection process.

REFERENCES:

[1] APWG-Unifying Global response to cybercrimehttp://docs.apwg.org/word_phish.html

[2] *International Journal of Advanced Research in Computer(IJARCET)* – "An Efficient Approach To Detecting Phishing A Web Using K-Means And Naïve-Bayes Algorithms"

[3] International Journal of Advanced Computer Technology (IJACT), "A Review of Various Techniques for Detection and Prevention for Phishing Attack".

[4] IEEE 2017 - Feature selection for machine learning based detection of phsihng websites "http://ieeexplore.ieee.org/abstract/document/8090317/?reload=true"

[5] Heuristic-based Approach for Phishing Site Detection Using URL Features- Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology - CEET 2015 Copyright © "Institute of Research Engineers and Doctors, USA. All rights reserved."

[6] Prof.T.BhaskarAher Sonali, Bawake Nikita , Gosavi Akshada ,Gunjal Swati' 'Detection of Website Phishing Using MCAC Technique Implementation', "http://ijariie.com/AdminUploadPdf/Detection_of_Website_Phishing_Using_MCAC_Technique_Implementation_ijariie1807.pdf"

[7] 2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA) , "Detection of Phishing Emails using Data Mining Algorithms"

[8] Phishtank- "https://www.phishtank.com"

[9] "Performance Comparison of Feature Selection Methods" -Thu Zar Phyu , Nyein Nyein Oo, MATEC Web of Conferences.

[10] A Novel Multi-Layer Heuristic Model for Anti-Phishing- "https://dl.acm.org/citation.cfm?id=3078580", ACM", paper 2017.