Fraud Webpage Detection using Machine Learning

Sejal Gurkhe, Ishita More
Department of Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
{sejal.gurkhe, ishita.more}@spit.ac.in

Guide Name: Assistant Prof. Kiran Gawande
Department of Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
kiran_gawande@spit.ac.in

***Abstract*—Phishing is a frequent attack used to cheat people to get their personal information using duplex web pages created to resemble real websites. With the growing machinery, phishing attacks are on the hike. Web phishing intends to steal confidential data, such as usernames, keys, etc, by imitating a legitimate entity. It will lead to information exposure and property damage. Machine Learning is a very suitable way to identify phishing websites. This paper analyzes the existing machine learning methods that are used to identify phishing websites. The paper explains the Logistic Regression method, and Multinomial naive Bayes methods which have been implemented with accuracy of 96.38%, and 95.744% respectively.**

***Index Terms*—Phishing, Phishing attacks, Machine Learning, Logistic Regression, Multinomial naive Bayes**

I. Introduction

Phishing is a fraudulent practice in which an attacker tries to obtain sensitive information by impersonating someone else to benefit himself/herself in a malicious way. Today, most of the users are accessing the services online, so it has become very easy for phishers to obtain user's confidential information. The website contents of phishing websites look very similar to that of legitimate websites and hence prompts people to provide their sensitive information. In a conventional phishing attack, a user's sensitive information like usernames and passwords can be revealed by tricking the user to click on some link to a phishing website. The attacker may duplicate the authorized website and may have some click tactics on this website that can be used to trick a user so as to reveal their sensitive credentials. These credentials can further be made use of by the phisher for digital identity thefts or some financial profits as well.Phishing attacks can be prevented by making users distinguish between phishing and legitimate websites. s. Most of the phishers use images rather than text which are difficult to detect. Various tools and mechanisms have been developed to detect phishing websites and to prevent attacker from obtaining sensitive information. Blacklisting is one the easy way to detect phishing websites but can't be used to find new phishing websites. It is also a time consuming process. The remaining sections in this paper are classified as follows: the next section talks about the related work and Literature survey in this field. Section 4 includes our proposed approach. We have discussed the classification algorithms that are used for the classification of phishing and non-phishing URLs. In section 5, we discuss the results that were obtained on the datasets from the training models.

II. Related Work

In emerging technology industry which deeply influences today's security problems has given a non-ease of mind to some employers and home users. In the dimension of the new era, there are many security systems being developed to ensure security is given the utmost priority and prevention be taken from being hacked by those who are involved in cyber-criminal and essential prevention is also taken as high consideration in the organization to ensure network security is not being breached. Cyber security employees are currently searching for trustworthy and steady detection techniques for phishing websites detection. Due to wide usage of the internet to perform various activities such as online bill payment, banking transactions, online shopping, and, etc. Customers face numerous security threats like cybercrime. There are many cybercrimes that are extensively executed for example spam, fraud, cyber terrorism, and phishing. This phishing is known as a popular cybercrime today. Phishing has become one of the highest 3 most current forms of law-breaking in line with recent reports, and both frequencies of events and user

susceptibleness have enlarged in recent years, more combining the danger of economic damage.

III. Literature Survey

In the paper "Phishing Detection using Random Forest, SVM and Neural Network with Backpropagation" [1], the Random Forest classification method, SVM classification algorithm, and Neural Network classification algorithm were improved. Logistic regression classifier, SVM, naive Bayes, and Decision tree algorithm were discussed. The data file used is from the UCI ML repository. It consists of 11,055 URLs with 6157 fraud and 4898 fair detail. Each detail contains 30 features. Initially, lexical feature extraction is performed first on the dataset and then passed to these algorithms. A Random Forest classifier involves combining the results of various stump trees (with a single hierarchy) to reach a conclusion. This algorithm is implemented using sklearn's RandomForestClassifier module. The use of backpropagation is to reduce the error in the final result as the error gets back propagated and the weight given to each hidden layer of neurons changes with each iteration. SVM classifier has been implemented using sklearns's SVM.SVC() using the kernel mode. Initially, the classifiers when used produced accuracy rates of 87.34%, 89.63%, and 89.84% for neural networks, random forest, and SVM classifiers. Using the above results, which state that SVM is the best classifier, is used to implement it.

In the paper "Phishing Website Detection Based on Machine Learning Algorithm" [2], Logistic regression classifier, SVM, Naive Bayes, Decision tree algorithm were discussed. The data collection process of a single URL, including URL, crawling, various detection and filtering of results, etc., finally put the URL back into the URL pool. First, a collection thread selects a URL from the URL pool and then crawls the webpage corresponding to the URL. 12 well-performed features are selected as the features of our machine learning model for phishing website detection. In the Logistic regression classifier, the accuracy rate is 95.12%. The accuracy obtained with the SVM classification algorithm is 95.15%. The accuracy obtained with the Naive Bayes algorithm is 70.05%. The accuracy obtained with the Decision tree algorithm is 90.04%. Therefore, a Logistic regression classifier is selected.

In the paper "Phishing Detection Using Machine Learning Technique" [3], SVM, and Random Forest classification methods were discussed. The data set contains phishing and legitimate data or eliminates web page failures that have failed on the "error 404" page. The Random Forest algorithm is used twice. In the first test, 30 features were considered. It gives 94.27% accuracy. While in the second test, 5 features were considered, which gave 94.22% accuracy. Accuracy in the second case decreases. While SVM gives 95.66% accuracy by using 5 features. The proposed technique can detect new temporary phishing sites and reduce the damage caused by phishing attacks.

In the paper "Detection and Prevention of Phishing Websites using Machine Learning Approach" [4], a Logistic regression classifier, Decision tree, and Random Forest algorithm were discussed. The basic idea in this paper is the hybrid solution which uses all the three approaches – blacklist and whitelist, heuristics, and visual similarity. Studying the domain of each URL with the white-list of trusted domains and also the blacklist of legal domains. If found then return to the client. Otherwise using algorithms. In the Logistic regression classifier, the accuracy rate is 96.23%. The accuracy obtained with the Decision tree algorithm is 96.23%. The accuracy obtained with the Random Forest algorithm is 96.58%. Therefore, a Random Forest algorithm is selected.

In the paper "URL Phishing Analysis using Random Forest" [5], a Linear SVM, Non-Linear SVM, and Random Forest algorithm were discussed. The two distinct datasets to select the apt model. Both of the datasets are acquired from the UCI Machine Learning Repository. One dataset consists of 1 target feature and 30 features. It contains 2456 entries of fraud as well as legal URLs. The second dataset contains 1353 URLs with 10 features, and these URLs are classified into 3 categories: Phishing, non-phishing and suspicious. It consists of 11,055 URLs with 6157 fraud and 4898 fair details. In the Linear SVM, the accuracy rate is 93.07%. The accuracy obtained with the Non-Linear SVM algorithm is 94.97%. The accuracy obtained with the Random Forest algorithm is 95.11%. Therefore, a Random Forest algorithm is selected.

IV. Proposed Approach

Two methods to detect phishing websites have been implemented. The dataset is taken from the UCI Machine Learning repository. It consists of 2 parameters, namely URL and Target Label. The target label contains Good and Bad, where Good implies a legitimate website and Bad implies fraud. Some python libraries used for implementation are sklearn, seaborn, numpy, and pandas. The library "sklearn" was used to implement Logistic Regression and Multinomial Naive Bayes algorithms. The library "sklearn" is also used for splitting the data into train and test, to generate classification reports, confusion matrix, etc. The library "nltk" was used for preprocessing. Using nltk, the URL from the dataset is tokenized using RegexpTokenizer, stemming using snowball stemmer. The dataset is divided into 3:1 ratio i.e. 75% is train data and 25% is test data using train_test_split(). The features of URL are extracted using NLTK and Logistic Regression and Multinomial Naive Bayes algorithm is run to classify the websites as phishing websites or legitimate websites. The classifier algorithm giving the best accuracy score is selected as the final classifier algorithm.

Initially, the classifiers when used produced accuracy rates of 96.23% and 70.05% for the Logistic Regression and Multinomial Naive Bayes algorithm. Using the above results, which state that Logistic Regression is the best classifier. It is loaded to state whether the loaded site is prone to phishing or not.

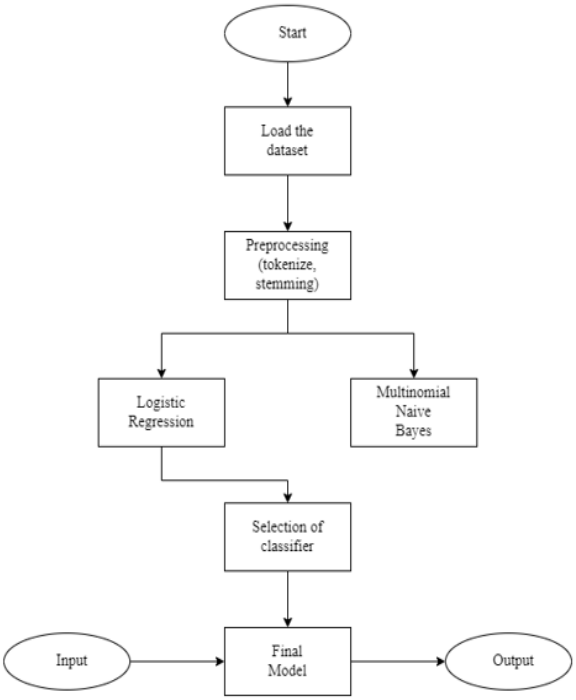

Fig. 1: System Architecture

Training Accuracy : 0.9783572688946115
Testing Accuracy : 0.9638334898825517

CLASSIFICATION REPORT

	precision	recall	f1-score	support
Bad	0.91	0.96	0.93	36442
Good	0.99	0.96	0.98	100895
accuracy			0.96	137337
macro avg	0.95	0.96	0.95	137337
weighted avg	0.97	0.96	0.96	137337

Fig. 2: Logistic Regression Classification report

V. Analysis Result

Scikit-learn tool has been used to import the two Machine learning algorithms(Logistic Regression and multinomial Naive Bayes). Dataset is divided into training set and testing set in , 75:25 ratios . Each classifier is trained using training set and testing set is used to calculate performance of classifiers. Performance of classifiers has been evaluated by calculating classifier's accuracy score, false negative rate and false positive rate. The accuracies achieved using Logistic Regression and multinomial Naive Bayes algorithms are 96.369 % and 95.74 %respectively. Logistic Regression is found to be the better algorithm among the two as it gives better than Multinomial naive Bayes.

Training Accuracy : 0.9739204726110352
Testing Accuracy : 0.9574477380458288

CLASSIFICATION REPORT

	precision	recall	f1-score	support
Bad	0.91	0.93	0.92	38035
Good	0.97	0.97	0.97	99302
accuracy			0.96	137337
macro avg	0.94	0.95	0.95	137337
weighted avg	0.96	0.96	0.96	137337

Fig. 3: Multinomial Naive Bayes Classification report

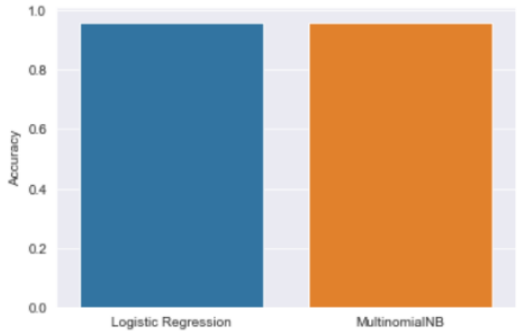

Fig. 4: Accuracy

VI. Future Scope

For future enrichment, we aim to form the phishing detection system as an expandable web service that will integrate with online learning so that new phishing attack patterns can easily be learned and enhance the accuracy of our models with better feature extraction.There is need to find out the best parameter which improve the accuracy of various classification technique and give best result in detection

VII. Conclusion

This paper aims to strengthen detection method to detect phishing websites using machine learning technology.URL phishing analysis is very useful in determining whether a certain URL is a legitimate URL or not and whether it should be visited or not. This helps the users a lot in knowing which of the websites should be avoided. Thus, it prevents them from revealing their sensitive information to unknown or illegitimate sources.

Since, Logistic Regression algorithm gave better accuracy as compared to that of Multinomial Naive Bayes algorithm, Logistic Regression is chosen as final classifier algorithm for classification of websites as phishing or legitimate.

REFERENCES

[1] A. S. F. R. Smita Sindhu, Sunil Parameshwar Patil, "Phishing detection using random forest, svm and neural network with backpropagation," *International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, 2020.

[2] W. Bai, "Phishing website detection based on machine learning algorithm," *International Conference on Computing and Data Science (CDS)*, 2020.

[3] M. W. N. T. N. Junaid Rashid, Toqeer Mahmood, "Phishing detection using machine learning technique," *First International Conference of Smart Systems and Emerging Technologies (SMART TECH)*, 2020.

[4] Z. S. X. L. Feng Vue, Jianmin Pang, "Detection and prevention of phishing websites using machine learning approach," *16th IEEE International Colloquium on Signal Processing Its Applications (CSPA)*, 2020.

[5] U. P. A. using Random Forest, "Url phishing analysis using random forest," *International Journal of Pure and Applied Mathematics*, 2018.