

HYBRID PHISHING SITE DETECTION

Nilesh M. Patil¹, Sunny P. Dias², Ashley A. Dcunha³, Rohit J. Dodti⁴

^{1,2,3,4} Fr. Conceicao Rodrigues College of Engineering, Bandra, Mumbai, Maharashtra, India.

Abstract

Phishing is an attempt to steal private and financial statistics from customers, along with passwords, credit card numbers, via electronic verbal exchange along with e mail and different messaging services. Attackers claim to symbolize a corporation that will manual customers to a fake internet site reminiscent of a phishing website, for you to then be used to acquire private information from users. Attackers may also trick clients into downloading malicious codes or malware after clicking on a connection within the email address. Phishing identification is a massive problem that numerous researchers with different superior approaches should contemplate on. Effective mechanisms to defend users from phishing websites are required.

Keywords— Machine Learning, Classification, Phishing sites, Legitimate sites

I. INTRODUCTION

Phishing website is a type of attack aimed at stealing from users' confidential information which is important to the attacker. Phishing is achieved by giving the victim an email to be a lure, or by "instant messaging" in chat rooms. Victims get into a false website belonging to the abuser, thinking this is the one they want. Then, the attacker steals personal information including credit card details and passwords. The reason behind users being tricked may come from good production of identical web pages to the original ones that appear legitimate (having the same website layout and design, identical domain names, etc.), poor knowledge of technology and lack of understanding of possible attack techniques. In the report released on March 2018 by the Anti-Phishing Working Group (APWG) on phishing activity, they found that the number of unique phishing websites detected was 183,555 where it was increased from 48,114 detected in October 2017. Recent developments in phishing detection have contributed to the development of several new machine-learning techniques. In machine learning based techniques, a classification algorithm is trained using certain attributes that can differentiate a phishing website from the legit one. The phishing websites are low-lived, and every day we build thousands of fake websites. Real-time, quick, and cognitive solution to phishing detection is therefore required. Recent developments in phishing detection have contributed to the development of several new machine learning techniques.

II. LITERATURE SURVEY

Ali Hadi and Dyana Rashid Ibrahim[1] have suggested a WEKA approach widely used by Machine Learning experts, where it has illustrated its strength in data mining applications such as; clustering, regression and classification, as it offers multiple algorithms for these aspects that could be used on our repositories.

M. Aburrous[2] has suggested a technique that makes use of data mining techniques for classification. The detection of e-banking phishing websites is specially implemented. Six separate classification algorithms are in this technique i.e. it implements PART, PRISM, C4.5, MCAR, PRISM and JRip.

Hadi Zamani and Muhsamad Kamal[3] explained the data collection and dataset and defined the data set division used to train and test classifiers and then explained the evaluation metrics used to evaluate each algorithm's performance.

Ankit Kumar Jain and B.B Gupta[4] measured performance in terms of correct classification rate (CCR), true positive rate (TPR), true negative rate (TNR) and geometric mean (GM) of kNN, RBFN, RF, NB, C4.5, BPNN and SVM and Fivefold cross validation using WEKA software to check the efficiency of wrapper-based machine learning classifiers.

Mohammed Nazim Feroz and Susan Mengel[5] classified the URLs automatically based on their mainly phonological and host- attributes. To get a feature vector that aptly describes the URL, classification requires careful mapping for various types of stigmas associated with the measurements of the predictors. The hashing of features, for example, is used to transform raw feature data into feature vectors. The size of the feature vector was selected after careful analysis of the dataset to be in the space

of 59,000 dimensions which limited collisions with features. Clustering is performed throughout the entire dataset and a cluster ID (or label) is extracted for each URL which is used as a predictive feature by the form method in return.

Abdulhamit Subasi, Esraa Molah[6] has provided an intelligent framework for the detection of phishing attacks. They used various data mining techniques to determine website categories: true, or phishing. Specific classifiers had been used to create accurate intelligent gadgets for website detection of phishing. ROC and F-measure is used to test the efficiency of the data mining techniques. Results showed that by achieving the highest accuracy of 97.36 percent, Random Forest has outperformed best among the classification methods.

III. PROPOSED SYSTEM

The proposed system focuses on predicting the phishing websites by checking its URL that is their IP address, domain name, etc. as well as taking screenshots of the website and identifying whether it's a fake or a genuine website

The proposed system could be very beneficial for the users who want to find out whether a website is fake or genuine one because sometimes a website asks for account details and without knowing we give them our credentials. We should always know that the websites we are giving our details is a genuine. To deal with the real life problem phishing detection can be very useful.

The system makes use of Logistic Regression Algorithm as well as tensor flow transfer algorithm which has been found to be the best and most adaptable for this model. Basically logistic regression is regression evaluation to conduct whilst the structured variable is dichotomous (binary). Logistic regression is used to characterize the facts and to illustrate the relation between one dependent binary variable and one or more nominal, ordinal, C program language time or independent ratio-level variables. Transfer learning algorithm is best classification mechanism algorithm till date for image classification with high accuracy.

In our system we use logistic regression on our URL attributes such as IP address, domain age and domain name. Logistic regression gives us two outputs, whether or not the website is phishing. We also use image processing algorithm which takes a screenshot of the websites and compares it simultaneously to the images produced by a phishing website such as excessive advertisement and pornographic content. Our system will also have a database of blacklisted URLs which will be cross checked once a user enters any URL, domain of the website. We can update the database and add new malicious URLs into the blacklisted database.

IV. SYSTEM DESIGN AND ALGORITHM USED

The framework of the proposed system to detect the phishing website is depicted in figure 1 below.

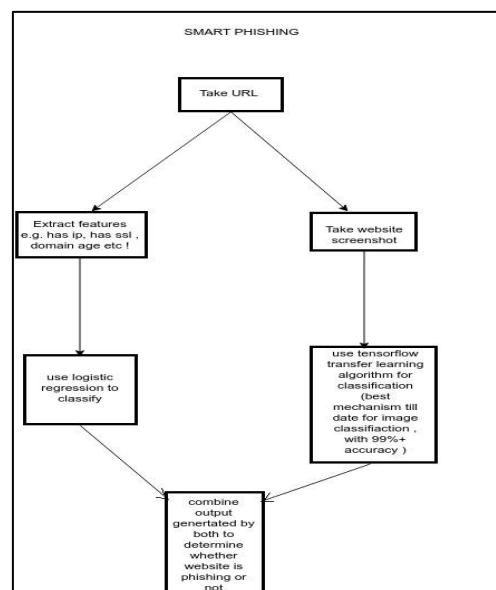


Figure 1 – Proposed System

The use case functionality is shown in figure 2 below. It provides the function performed by both user and server.

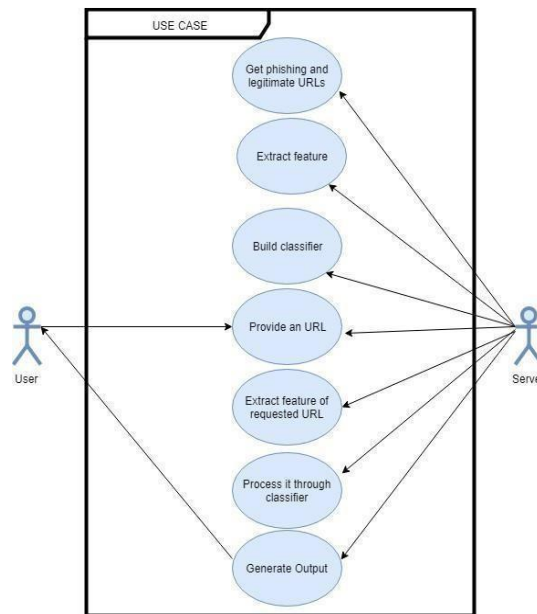


Figure 2 – Use Case Diagram

ALGORITHMS

Logistic regression

It is used to classify the URLs whether if they are phishing websites or not. Logistic regression is another approach borrowed by system studying from the sector of statistics. It is the go-to method for binary classification troubles (issues with magnificence values). Logistic regression is a statistical method for the study of a dataset in which a result is calculated by one or more independent variables. The end result is determined using a dichotomous vector (where the best two possible outcomes are). The defined variable in logistic regression is binary or dichotomous, i.e. it contains only records coded as either 1 (TRUE, progress, fictional website, etc.) or 0 (FALSE, failure, authentic website, etc.). The goal of logistic regression is to locate the excellent (but biologically reasonable) fitting model to explain the relationship between the binary function of interest (defined variable = response or outcome variable) and a set of discrete (predictor or explanatory) variables.

Derive Logistic Model from Linear Model

- Linear Model
 $y = ax + b$
 where y is dependent variable, a is constant, x is independent variable and b is the intercept
- Logit Function / Sigmoid Function

$$Y = \frac{1}{1 + e^{-y}}$$



Figure 3- Logistic Regression Example [8]

URL Feature Extraction

The heuristic-based technique and blacklisting approach are the methods typically used to detect the fake websites. The blacklist approach maintains the list of universal URLs that consist of user information. If such page is requested, the connection to that website is blocked. A URL is a web address that specifies its location on computer network and its retrieval mechanism. The URL is composed of the protocol identifier, resource name or IP address. It can be HTTPS, HTTP or FTP. Protocols are regulation policies used for communication and they can be in varied forms.

The subdomain allow user to organize and navigate to different sections of the website.

The domain given by the Domain Name System (DNS) identify the IP addresses. The primary domain like .com, .edu, .org, etc are the top-level domains and they are at the highest level of hierarchical DNS. For each part of the URL we define features; these features are used to detect the phishing sites.

A classifier is obtained using pre-collected URLs from fake websites and legit websites during the learning process. The accumulated URLs are relayed to the extractor method, which harvests model values based on the URLs through evidently defined capabilities. The extracted features are reported as qualifying and beating the generator computer algorithm, generating a classifier by using the entered features and recognizing the gizmo's algorithm. The classifier must decide in the detection process whether or not a requested site is an online phishing website.

When a page request occurs, which extracts the characteristic values in full via the predefined URL-based functions, the URL of the requested online website is transmitted to the extractor device. The classifier provides those values for the functions. Whether a brand new online website is primarily a phishing website based on the information acquired is determined by the classifier. This then warns the person who requests the article about the outcome of the classification.

Thus, there are several URL features which were used in many phishing detection studies. We integrate features used in previous studies in the present study and use best suited features along with logistic regression, recognizing known phishing sites.

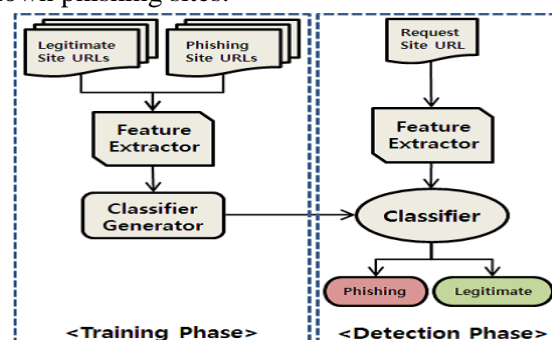


Figure 4 – Process of Detection

Part 1: Build classifier

Step 1) Get data set from uci.com

Step 2) Split the dataset into (train-test split) 70 percent (training data)-30 (testing data)

Step 3) Use Logistic Regression algorithm to train the model and build the classifier

Step 4) After training the classifier, test the classifier to check the accuracy of the classifier

Step 5) Repeat the training and testing till we get minimum error.

Part 2: Prediction

Step 1) Get the URL submitted by the user

Step 2) Extract Features from the URL using following steps:

2.1. Check its SSL state (whether it has SSL or not).

2.2. Get DNS info of the URL.

2.3. Check if the site content has iframes and alert used.

2.4. Check if the URL is running on port default 80 port or not.

2.5. Similar to above mentioned all other features will be extracted such as Google page rank, DNS record, google index, traffic, etc.

2.6. Feature extracted will be stored in a .csv file and later used for prediction

Step 3) Pass the extracted features and the URL to the classifier, which is prior trained and tested.

Step 4) Send the output from the classifier to the Frontend system.

We considered 30 different features that would help as deciding variables such as Using Non-Standard Port, Re-quest URL, Anchor URL, The Existence of HTTPS Token in the Domain Part of the URL, Links in Meta tag, Link tags and Script tag, Server Form Handler (SFH), Disabling Right Click, Abnormal URL, Website Forwarding, Status Bar Customization, Submitting Information to Email, Using Pop-up Window, Iframe Redirection, Domain Age, DNS Record, Website Traffic, PageRank, Google Index, Hyperlinks, Feature Based Statistical Reports, etc.

TensorFlow

TensorFlow is an open source software library for the use of data-float graphs for numerical computation. It was originally created by the Google Brain Team within Google's Artificial Intelligence research enterprise to get to know system and study deep neural networks, but the software is generally adequate to be applicable in a wide range of other domains. TensorFlow is transversal. It runs on nearly everything: GPUs and CPUs including wireless and integrated platforms and even Tensor Processing Units (TPUs), which are sophisticated hardware for tensoring math. The execution engine allotted to TensorFlow abstracts the various supported gadgets and provides the TensorFlow framework with an unnecessary performance-core implemented in C++. The Python and C++ frontends (with extra to come) rest on the pinnacle of that. In deep learning models, the Layers API offers a less challenging framework for commonly used layers.

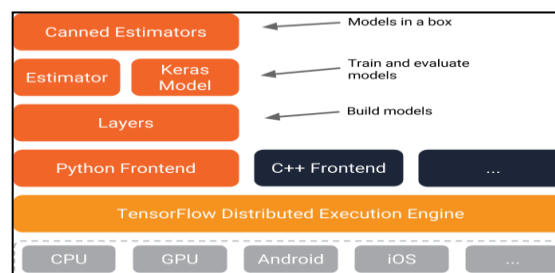


Figure 5 – TensorFlow Architecture [10]

In addition to this, there are higher-degree APIs, which include Keras (extra on the Keras.io site) and the Estimator API, which enables the learning and evaluation of dispensed models. And eventually, a range of commonly used versions are fitted out of the box to submit, with more to follow.

V. RESULT ANALYSIS

URL Classification

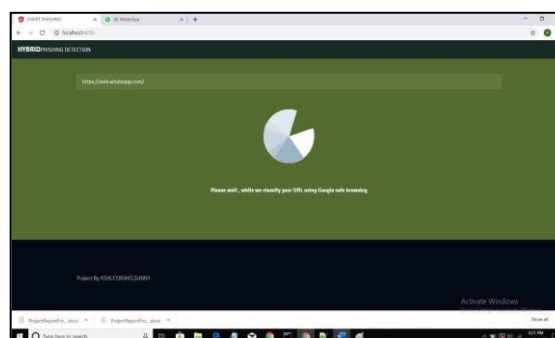
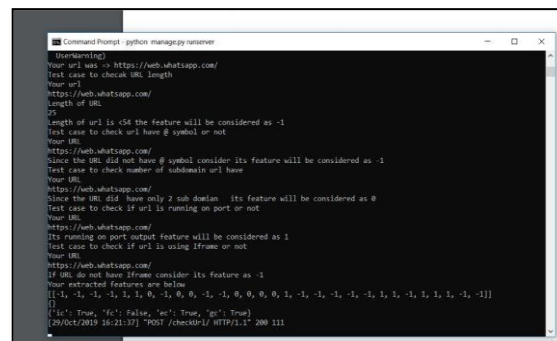


Figure 6 – URL Classification

Feature Extraction



```
Command Prompt - python manage.py runserver
UserWarning:
Your url is https://web.whatsapp.com/
Test case to check url length
Your url
https://web.whatsapp.com/
Length of url
25
Length of url is <54 the feature will be considered as -1
Test case to check url have @ symbol or not
Your url
https://web.whatsapp.com/
Since the url did not have @ symbol consider its feature will be considered as -1
Test case to check number of subdomain url have
Your url
https://web.whatsapp.com/
Since the url did have only 2 sub domain its feature will be considered as 0
Test case to check if url is running on port or not
Your url
https://web.whatsapp.com/
Its running on port output feature will be considered as 1
Test case to check if url is using iframe or not
Your url
https://web.whatsapp.com/
If url do not have iframe consider its feature as -1
Your extracted features are below
[[-1, -1, -1, -1, 1, 1, 0, -1, 0, 0, 0, 0, 1, -1, -1, -1, -1, -1, 1, 1, 1, 1, -1, -1]]
[{'ic': True, 'fc': False, 'ec': True, 'gc': True}]
[29/Oct/2019 10:11:37] "POST /checkUrl/ HTTP/1.1" 200 111
```

Figure 7 – Feature Extraction

Home page consists of a search bar and a check button. In google safe browsing the URL is checked against a blacklist that is provided by Google which lists URL's that are malicious. Every feature has been assigned a numerical value (1, 0, -1) corresponding to their safety level.

Image Classification

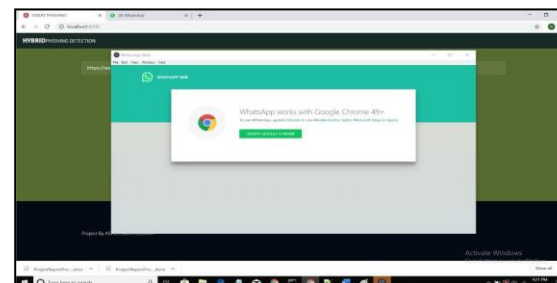


Figure 8 – Image Classification

The snapshot of the webpage is taken and used for classification.

Result for safe website

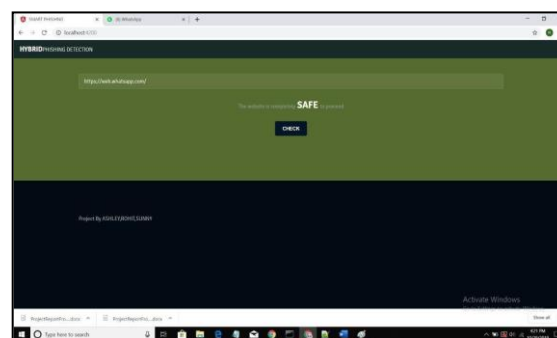


Figure 9 - Result for Safe Website

From clean and unclean URL database we get images of all the websites. We use Keras CNN machine learning algorithm and train the model with two classes and then create a model and save it. Take the URL entered by the user, get screenshot of the website and then classify that image from trained model. Get output and display the result on the frontend.

For unsafe website

The same process is carried out. URL extraction and Image Classification for unsafe website is done. The outputs are shown in figure 10, 11 and 12 respectively.

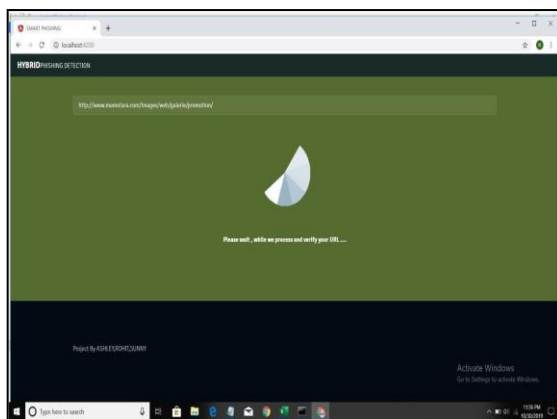


Figure 10 – Unsafe Website

VI. CONCLUSION AND DISCUSSION

We have recognized several new features for identifying phishing websites in this post. These functions are focused upon the information given in the URL of the website. We used these features to educate classifier logistic regression which achieved high precision in phishing detection and legitimate websites. Our suggested solution to phishing detection depends entirely upon the website's URL. Adding some more features may improve the accuracy of the classification.

Phishing detection strategy that utilizes URL-based technology has been introduced in the past, using several technologies that render the system complex. With a large number of features, creating classifiers and performing classification is time-consuming for the heuristic-based method. But the suggested plan will provide protection of personal information, and raising the harm done by phishing scams. As it can monitor new and temporary spammy links circumventing existing techniques of phishing detection, such as blacklist-based techniques. The experimental results showed that the proposed method is very good in the detection of phishing websites, because it has a true positive rate of 98.39 percent and a total accuracy of 98.42 percent. By adding a few more features we will improve the accuracy of our approach.

We have used algorithms for image processing such as the TensorFlow transfer algorithm which is used to process website screenshots. We've improved our accuracy through both of these algorithms. In addition, attempts may be made to boost the classifier accuracies and provide additional parameters.

VII. FUTURE WORK

Our phishing site can detect spam and tricky websites as well. The Image classification will help to detect those websites that appear clean but are actually fake websites. Various advanced technologies can help detect fake sites which are coded in such a way that they appear to be legitimate. One limitation of our Project is that it cannot identify if a phishing site is embedded within a legitimate site. This issue should be solved/worked upon in the future to improve efficiency.

Image Classification for Unsafe URL

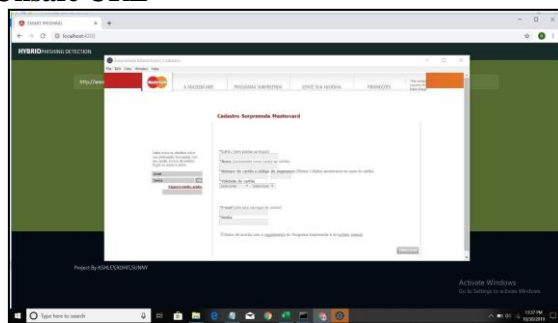


Figure 11 - Image Classification for Unsafe URL

Result of Unsafe Website

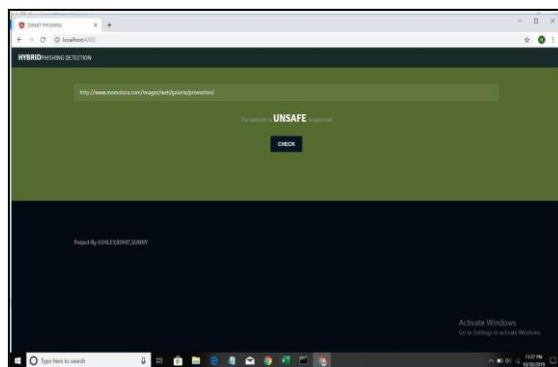


Figure 12 – Unsafe Website Result

REFERENCES

1. Ali Hussein Hadi, Dyana Rashid Ibrahim, “Phishing Websites Prediction Using Classification Techniques”, IEEE International Conference on New Trends in Computing Sciences (ICTCS), May 2017, pp. 133-137.
2. Maher Aburrous, M. A. Hossain, Keshav Dahal, Fadi Thabtah, “Predicting Phishing Websites using Classification Mining Techniques with Experimental Case Studies,” IEEE Seventh International Conference on Information Technology: New Generations, April 2010, pp. 176-181.
3. Hadi Zamani, Muhamad Kamal bin Mohammed Amin, “Classification of phishing websites using machine learning techniques”, Journal of Advanced Research in Applied Sciences and Engineering Technology 5, Issue 2 (2016) 12-19
4. Ankit Kumar Jain, B. B. Gupta, “A machine learning based approach for phishing detection using hyperlinks information”, Journal of Ambient Intelligence and Humanized Computing, Issue 10, April 2018, pp. 2015-2028.
5. Mohammed Nazim Feroz, Susan Mengel, “Phishing URL detection using URL Ranking”, IEEE International Congress on Big Data, 2015, pp. 635-638.
6. Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, Touseef J. Chaudhery, “Intelligent Phishing Website Detection using Random Forest Classifier”, IEEE International Conference on Electrical and Computing Technologies and Applications (ICECTA), Nov. 2017, pp. 1-5.
7. <https://us.norton.com/internetsecurity-online-scams-how-to-protect-against-phishing-scams.html>
8. <https://medium.com/@YvesMulkers/what-is-the-tensorflow-machine-intelligence-platform-33f0166a05d>
9. <https://support.eset.com/enEN/kb3100/?locale=en EN&viewlocale=en US>
10. <https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc>