IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 5, Issue 4, November 2016)

A Novel Approach for Phishing URLs Detection using Naïve Bayes, Neural Network and Random Forest Classifiers

Gangeshwari Sharma
gangeshwarisharma@gmail.com

Abhishek Tiwari
abhi.tiwari23@gmail.com

Abstract – **Seeking sensitive user data in the form of online banking user-id and passwords or credit card information, which may then be used by 'phishers' for their own personal gain is the primary objective of the phishing e-mails. With the increase in the online trading activities, there has been a phenomenal increase in the phishing scams which have now started achieving monstrous proportions. This paper gives strategies for distinguishing phishing sites by dissecting different components of phishing URLs by Machine learning systems. It talks about the systems utilized for identification of phishing sites in view of lexical features, host properties and page significance properties. We consider different machine learning algorithms for assessment of the features to show signs of improvement comprehension of the structure of URLs that spread phishing. We use Naïve Bayes, Neural Network and Random Forest Classifiers.**

Keywords – **Machine Learning System, Phishing URL, Naïve Bayes, Neural Network and Random Forest.**

I. INTRODUCTION

A Phishing is an attempt to steal personal confidential information such as passwords, credit card information from innocent victims for financial gain, identity theft and other fraudulent activities by an individual or a group. The current scenario, when the user desires to access his confidential information online (like payment gateway or money transfer) by logging into his secure mail account or bank account, the individual enters information like credit card no., username, password etc. on the login page. But quite often, this information can be taken by intruders using phishing techniques (for example, when a user provides login information on a phishing website his data is stolen and then he is redirected to the genuine site). There is no such information that cannot be directly obtained from the user at the time of his login input.

Whittaker et al. [5] define a phishing web page as "any web page that, without permission, alleges to act on behalf of a third party with the intention of confusing viewers into performing an action with which the viewers would only trust a true agent of a the third party."

Phishing is a generally a web criminality in relationship with various structures, for example, virus attacks and hacking. In recent times, an expansive number of phishing web pages have been discovered. Its effect is data security rupture through the cooperation of classified information and the objectives may at long last endure loss of cash or different types.

A phishing site as appeared in Figure 1 is a generally determined social engineering attack that endeavours to cheat people of their profound data containing bank account information, credit and debit card number, social security number, and their personal credentials in order to use these details fraudulently against them. Phishing has a tremendous negative effect on associations' revenues, client connections, advertising endeavours, and general corporate picture. Phishing can cost firms a huge amount of money per attack in personnel time and fraud-associated losses. Terribly, expenses connected with the degradation in brand value and customer enthusiasm can keep running into a huge number amount.

Figure 1: Screenshot of a phishing website

Phishing web pages are fake web pages that are made by malicious individuals to mimic Web pages of genuine web sites. Most of these types of web pages have great visual similarities to trick their victims. Some of these types of web pages look

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 5, Issue 4, November 2016)

exactly like the genuine ones. Victims of phishing web pages may expose their credit card number, password, bank account or other vital information to the phishing web page owners. It includes techniques such as deceiving customers through URL, screen captures, spam messages, emails and installation of key loggers.

This study focuses on a lexical analysis of URLs because they are:

- Less expensive to process than external information or content-data.
- URLs are more likely to be stored and obtainable as they use up less resources, such as disk space than external information and content-data. Content-based analysis more costly to obtain.
- There is no point to using complex operations if we have not evaluated simpler operations. If simpler operations are not productive enough, then we may want to use more complex operations. However, as covered in subsequent sections, our study shows that a lot can be achieved with just lexically analyzing URL.

II. Proposed Method

The work comprises of lexical feature extraction of collected URLs and investigation. The primary step is the gathering of phishing and benign URLs. The lexical based feature extractions is used to shape a database of feature values. The database is learning mined utilizing various machine learning strategies. Subsequent to assessing the classifiers, a specific classifier is chosen and is executed in MATLAB. Figure 2 shows the proposed flow diagram.

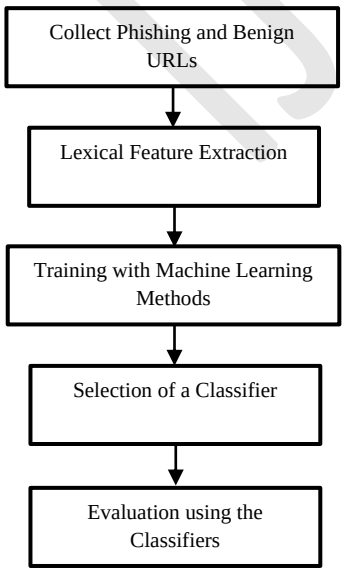

Figure 2: Flow diagram for the proposed work

Collection of URLs

In this paper, we have taken URLs of benign websites from www.alexa.com [15] www.dmoz.org [16] and personal web browser history. The phishing URLs were collected from www.phishtak.com [17]. Following properties are recognized:

Protocol

The <protocol> portion of the URL demonstrates which network protocol ought to be utilized to fetch the requested resource. The most widely used protocols are Hypertext Transport Protocol or (http), HTTP with Transport Layer Security (https), and File Transfer Protocol (ftp).

Number of Dots

There are a number of ways for attackers to construct Legitimate-looking URLs. One such method uses subdomains, like http://www.my-bank.update.data.com. Another method is to use a redirection script, such as http://www.google.com/url?q=http://www.badsite.com. To the user, this may appear to be a site hosted at google.com, but in reality will redirect the browser to badsite.com. In both of these examples, either by the inclusion of a URL into an open redirect script or by the use of a number of subdomains, there are a large number of dots in the URL. Of course, legitimate URLs also can contain a number of dots, and this does not make it a phishing URL, however there is still information conveyed by this feature, as its inclusion increases the accuracy in our empirical evaluations. This feature is simply the maximum number of dots (‘.’) contained in any of the links present in the URL, and is a continuous feature.

Number of Special Characters

Some recent browser vulnerabilities have helped in misleading the users too. One such example was the Internet Explorer URL spoofing vulnerability. This vulnerability can allow an attacker to modify the address displayed on the address bar of the browser, while a fake web site is opened. For example consider the URL given below:

http://www.genuinesite.com%01%00@fakesite.com/

If this URL is visited, the address bar in the browser only displays http://www.genuinesite.com, whereas the user is actually visiting a page on fakesite.com. This vulnerability was caused due to incorrect interpretation of URLs that contained special characters such as %01 and %00.

Number of Slashes and @

Phishers tend to use more dots in their URLs to impersonate a legitimate look of URL because there is no restrictions on the number of dots can be used in sub domains. Checking URL against special

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 5, Issue 4, November 2016)

symbols such as "@", is another feature because many of phishing URLs modified using these symbols which makes it possible to write URLs that appear legitimate but actually lead to different pages. URLs corresponding to legal websites usually do not have a large number of slashes. As a result, URL that contains a large number of slashes is considered to be a phishing.

Other URL based Properties

There are a lot of variety of URL based properties which can be used in a phishing URL. In this research work, we find such properties to detect phishing URL. Coding part of this research contains following properties: "update", "click", "user", "termination", "double com", "confirm", "account", "banking", "secure", "ebayisapi", "webscr", "login", "paypal", "free", "lucky", "bonus" and "signin".

Lexical Feature Extraction

Lexical features are the textual properties of the URL itself, not the substance of the page it indicates. URLs are human-readable text strings that are parsed in a standard manner by customer projects. Through a multistep determination process, programs make an interpretation of each URL into guidelines that find the server facilitating the site and indicate where the site or asset is set on that host.

Machine Learning Algorithms

The program flow for the classifier performance is shown in Figure 3.

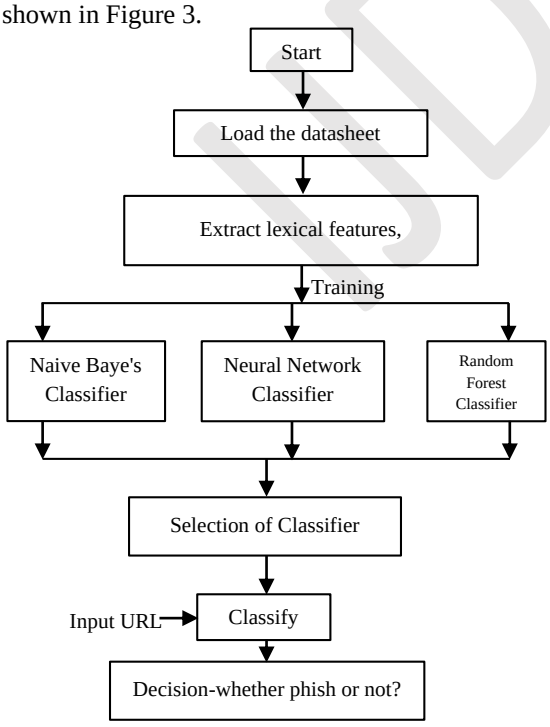

Figure 3: Flow diagram of proposed research

The three machine learning algorithms considered for processing the feature set are:

Neural Network: Back propagation neural network is a type of multi-layer feed forward network in which each layer is connected by transfer functions and can fulfil arbitrary nonlinear mapping. It is widely applied in stock price, petroleum price, economic time sequence, network flow and other nonlinear areas and attained satisfactory performance.

The basic learning process of the back propagation neural network algorithm is as follows:

1. Initialize the connection weights w_{ij}, v_{jt} and threshold θ_j in the back propagation neural network.
2. Input the first learning sample couples to the back propagation neural network.
3. Compute the input u_j of each neural unit and the output h_j in the hidden layer. The equation is:
$$u_j = \sum_{i=1}^{n} w_{ij} x_i - \theta_j \quad (1)$$
$$h_j = f(u_j) = \frac{1}{1+\exp(-u_j)} \quad (2)$$
4. Compute the input l_t of each neural unit and the output y_t in the output layer. The equation is:
$$l_t = \sum v_{jt} h_j - \gamma_t \quad (3)$$
$$y_t = \frac{1}{1+\exp(-l_t)} \quad (4)$$
5. Compute the weights error $\delta_t$which is connected to the neural unit t in the output layer.
$$\delta_t = (c_t - y_t) y_t (1 - y_t) \quad (5)$$
In the equation (5), c_t represents the expectation of the sample.
6. Compute the weights error δ_j which is connected to the neural unit j in the hidden layer.
$$\delta_j = \sum_{t=1}^{q} \delta_t v_{jt} h_j (1 - h_j) \quad (6)$$
7. Update the connection weights v_{jt} and threshold γ_t in the back propagation neural network.
$$v_{jt}(N+1) = v_{jt}(N) + \alpha \delta_t h_j \quad (7)$$
$$\gamma_t(N+1) = \gamma_t(N) + \beta \delta_t \quad (8)$$
8. Update the connection weights w_{jt} and threshold θ_j in the back propagation neural network.
$$w_{jt}(N+1) = w_{jt}(N) + \alpha \delta_j x_i \quad (9)$$
$$\theta_j(N+1) = \theta_j(N) + \beta \delta_j \quad (10)$$
9. Input the next learning sample and go to the step 3 until all of the samples are trained.

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 5, Issue 4, November 2016)

10. Back propagation neural network go to a new round of learning. If it meets the equation (11), the training of the back propagation network can be ended.

$$|\sum_{k=1}^{z} E_k| \leq \varepsilon \qquad (11)$$

In the equation (11), ε represents the accuracy requirement of back propagation neural network, E_k represents the mean square error and the definition are as follows:

$$E_k = \frac{1}{2}\sum_{t=1}^{q}(c_t - y_t)^2 \qquad (12)$$

Before training the back propagation neural network, proper connection weights w_{ij} and v_{jt} of the back propagation neural network should be chosen. Normally the initialization is randomly which can cause the convergence is slow and the defect of local optimal solutions.

Naive Bayes: Naive Bayes is a simple probabilistic classifier based on applying Bayes' theorem (or Bayes's rule) with strong independence (naive) assumptions. Parameter estimation for Naïve Bayes models uses the maximum likelihood estimation. It takes only one pass over the training set and is computationally very fast.

- Bayes Rule

A conditional probability is the likelihood of some conclusion, C, given some evidence/observation, D, where a dependence relationship exists between C and D.

This probability is denoted as $P(C|D)$ where,

$$P(D/C) = [P(D/C)P(C)]\ /[P(D)] \qquad (13)$$

- NB Classifier

Naïve Bayes classifier is one of the high detection approach for learning classification of text documents. Given a set of classified training samples, an application can learn from these samples, so as to predict the class of an unmet samples.

The features (n_1, n_2, n_3, n_4) which are present in URL are independent from each other. Every feature $n_i (1 \leq i \leq 4)$ text binary value showing whether the particular property comes in URL. The probability is calculated that the given web belongs to a class r (r_1: Non-phishing and r_2: Phishing) as follows:

$$P(r_1/N) = \ (P(r_1) * P(N/r_i))/P(N) \qquad (14)$$

Where all of $P(N)$ are constant meanwhile $P(n_i|r_1)$ and $P(r_i)$ can be easily calculated from training. The proportional to $P(r_1|N)$, $P(r_2|N)$ is calculated and the results are as follows:

$P(r_1|N)P(r_2|N) > b (b > 1)$; Non-phishing Web.
$P(r_2|N)P(r_1|N) > b$; Phishing Web
(15)

Random Forest Classifier: Random forests are recently proposed statistical inference tools, deriving their predictive accuracy from the nonlinear nature of their component decision tree members and the power of groups. Random forest committees provide more than just predictions; model information on data proximities can be exploited to provide random forest features. Variable importance measures show which variables are closely associated with a chosen response variable, while partial dependencies indicate the relation of important variables to said response variable.

Accuracy of Random Forest:

The Generalization error (PE^*) of Random Forest is given as:

$$PE^* = P_{x,y}\big(mg(X,Y)\big) < 0 \qquad (16)$$

Where, $mg(X,Y)$ is Margin function. The Margin function measures the extent to which the average number of votes at (X,Y) for the right class exceeds the average vote for any other class. Here X is the predictor vector and Y is the classification.

III. Simulation And Results

The performance of proposed algorithms has been studied by means of MATLAB simulation.

Figure 4: Confusion matrix for Neural Network

Figure 5: Confusion matrix for Random Forest

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 5, Issue 4, November 2016)

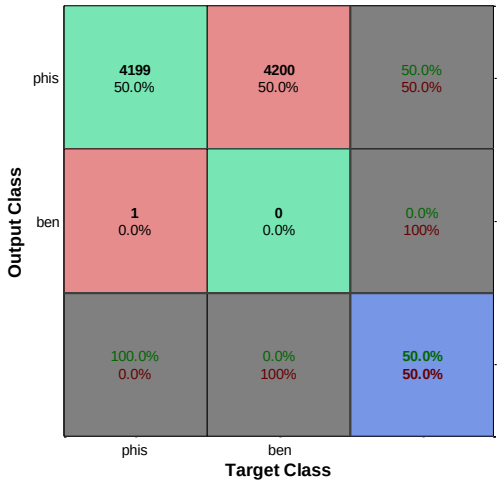

Figure 6: Confusion matrix for Naïve Bayes

Table 1: Result compared to previous work

Classifiers	Previous Work [18] (Accuracy)	Proposed Approach (Accuracy)
Naïve Bayes	74.20%	**50%**
Regression Tree	91.08%	-
KNN	79.55	-
SVM	87.65	-
Random Forest	-	**94.5%**
Neural Network	-	**94.6%**

IV. Conclusion

The performance of three different machine learning methods used is comparable, we found that Neural Network achieved consistently the best results. Random forest classifier provide slightly same results while Naïve Bayes has the lowest accuracy among all three methods. On observing Table 1, it was found that the proposed research work outperforms the previous work [18]. Previous research work [18] obtained maximum accuracy of 91.08% with Regression tree classifier while the proposed Random Forest and Neural Network classifiers achieve 94.5% and 94.6% respectively. Therefore the proposed approach is 3.5% more efficient.

As our future work, we plan to develop a framework using this approach and deploy it for a large-scale real-world test.

Reference

[1] Ollmann, G., "The Phishing Guide, Understanding and Preventing Phishing Attacks", Online Available: http://www.nextgenss.com/papers/NISR-WP-Phishing.pdf 2004.

[2] Watson, D., Holz, T., and Mueller, S. "Know your enemy: Phishing, behind the scenes of Phishing attacks", The Honeynet Project & Research Alliance, 2005.

[3] S. Garera, N. Provos, M. Chew, A.D. Rubin, "A framework for detection and measurement of phishing attacks", In: Proc. 5th ACM Workshop on Recurring Malcode, WORM"07, ACM, New York, NY, USA, 2007, pp. 1-8.

[4] J. Ma, L.K. Saul, S. Savage, G.M. Voelker, "Beyond blacklists: Learning to detect malicious web sites from suspicious URLs", In: Proc. 15th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, Paris, France, 2009, pp. 1245-1254.

[5] C. Whittaker, B. Ryner, M. Nazif, "Large-scale automatic classification of phishing pages", In: Proc. 17th Annual Network and Distributed System Security Symposium, NDSS"10, San Diego, CA, USA, 2010.

[6] Y. Zhang, J. Hong, L. Cranor, "CANTINA: a content based approach to detecting phishing web sites", In Proc. 16th Int. Conf. World Wide Web, WWW"07, Banff, Alberta, Canada, 2007, pp. 639-648.

[7] Google Safe Browsing API - Google Code, http://code.google.com/apis/safebrowsing/

[8] SmartScreen Filter – Microsoft Windows, Online available at: http://windows.microsoft.com/en-US/internetexplorer/products/ie-9/features/smartscreen-filter, 2011.

[9] N. Chou, R. Ledesma, Y. Teraguchi, D. Boneh, J. Mitchell, "Client-side defense against web-based identity theft", In: Proc. 11th Network and Distributed System Security Symposium, NDSS"04, San Diego, CA, USA, 2004.

[10] R. B. Basnet, A.H. Sung, Q. Liu, Rule-based phishing attack detection, In: Proc. Int. Conf. Security and Management, SAM"11, Las Vegas, NV, USA, 2011.

[11] SpoofStick Home, Online available at: http://www.spoofstick.com

[12] McAfee Site Advisor Software – Website Safety Ratings and Secure Search, Online available at: http://www.siteadvisor.com

[13] Netcraft Anti-Phishing Toolbar, Online available at: http://toolbar.netcraft.com

[14] AVG Security Toolbar, Online available at: http://www.avg.com/product-avg-toolbar-tlbrc#tba2

[15] The Web Information Company. Online available at: www.alexa.com

[16] DMOZ Open Directory Project. Online available at: http://www.dmoz.org

[17] PhishTank. Online available at: https://www.phishtank.com/

[18] James, Joby, L. Sandhya, and Ciza Thomas. "Detection of phishing URLs using machine learning techniques." In Control Communication and Computing (ICCC), 2013 International Conference on, pp. 304-309. IEEE, 2013.