Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 5, May -2015

An Efficient Approaches for Website Phishing Detection using Supervised Machine Learning Technique

[1]Riddhi J. Kotak, [2]Sagar H. Virani

[1]*Research Scholar, Master in Computer Engineering (M.E.-C.E.),V.V.P. Engineering College, Rajkot-360001*
[2]*Assistant Professor, Computer Engineering Department,V.V.P. Engineering College, Rajkot-360001*

Abstract —*Internet has become a useful component of our regular social and financial activities. Internet users may get harm due to different types of web threats which may cause loss of private information, financial damage , damage brand reputation due to which it loses customers confidence in E-commerce and online Transaction. Phishing is a form of web threats that is defined as the art of mimicking a website to illegally acquire and use someone else's data on behalf of legitimate website for own benefit (e.g. Steal of user's password and credit card details during online communication). So far, there is no single solution that can capture every phishing attack.This paper employs Machine-learning technique for modelling the prediction task and supervised learning algorithms namely Multi-layer perceptron, Decision tree induction and Naïve Bayes classification are used for exploring the results.*

Keywords- *Classification; Machine learning; Phishing; Prediction; Supervised learning.*

I. INTRODUCTION

Internet facilities are day by day reaching to customer to all over the world without any physical market place and without any restriction with effective use of E-commerce. Due to which customers who use Internet to purchase product are increasing more. Every day millions of money transactions are done through Internet. Phishing is a form of web threats that is defined as the art of mimicking a website of an authentic enterprise aiming to acquire private information on behalf of legitimate website for own benefit (e.g. Steal of user's password and credit card details during online communication). Social engineering and technical tricks are commonly combined together in order to start a phishing attack.

In general, phishing attacks are performed with the following four steps:

1) A fake web site which looks exactly like the legitimate Web site is set up by phishers

2) Phishers then send link to the fake web site in large amount of spoofed e-mails to target users in the name of legitimate companies and organizations, trying to convince the potential victims to visit their web sites.

3) Victims visit the fake web site by clicking on the link and input its useful information there.

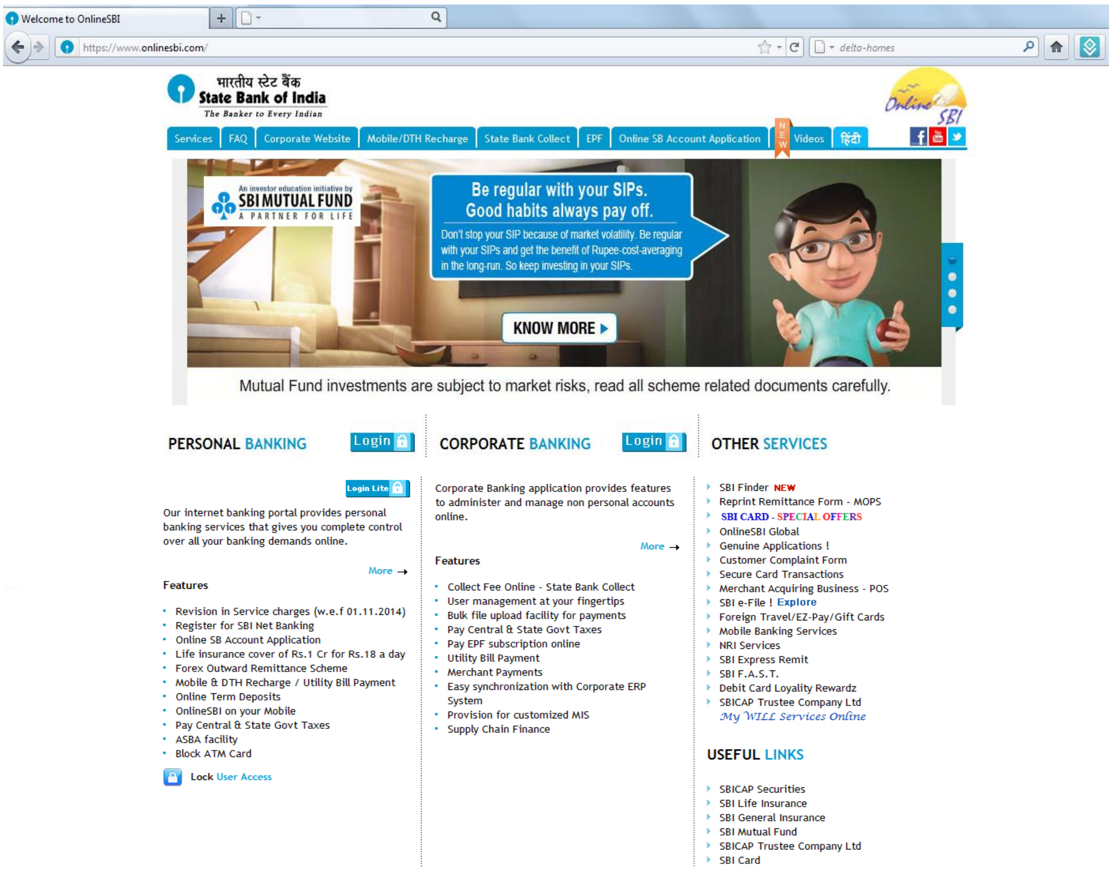
4) Phishers then steal the personal information and perform their fraud such as transferring money from the victims' account.

II. PROBLEM DEFINITION

Detecting fraudulent websites is a critical step towards protecting online transactions. Phishing is considered a binary classification problem since the target class has two possible values phishy or legitimate. Several approaches were proposed to discover and prevent these attacks. Consider the original website and the phished website of a bank namely, the State Bank of India (SBI) which is involved in e-banking. Unless the user is a known visitor of the site it is not possible for him/her to identify the authentication of the site based on its look and feel.

When we take a close look at the two sites some differences can be observed, (1) URL is different - The URL of the original site is **www.onlinesbi.com** and the URL of the phished website is **www.sbionline.com** and (2) Validation of the EV SSL certificate - Extended Validation Secure Sockets Layer (SSL) Certificates are special SSL Certificates that work with high security Web browsers to clearly identify a Web site's Organizational identity. Extended Validation (EV) helps you make sure a Web site is genuine and verified. In original websites, the address bar turns green indicating that the site is secured by an EV certificate.

- **The following figure 1 is the original website of SBI,**



Be regular with your SIPs.
Good habits always pay off.

Don't stop your SIP because of market volatility. Be regular with your SIPs and get the benefit of Rupee-cost-averaging in the long-run. So keep investing in your SIPs.

Mutual Fund investments are subject to market risks, read all scheme related documents carefully.

PERSONAL BANKING

Our internet banking portal provides personal banking services that gives you complete control over all your banking demands online.

Features

- Revision in Service charges (w.e.f 01.11.2014)
- Register for SBI Net Banking
- Online SB Account Application
- Life insurance cover of Rs.1 Cr for Rs.18 a day
- Forex Outward Remittance Scheme
- Mobile & DTH Recharge / Utility Bill Payment
- Online Term Deposits
- OnlineSBI on your Mobile
- Pay Central & State Govt Taxes
- ASBA facility
- Block ATM Card

Lock User Access

CORPORATE BANKING

Corporate Banking application provides features to administer and manage non personal accounts online.

Features

- Collect Fee Online - State Bank Collect
- User management at your fingertips
- Bulk file upload facility for payments
- Pay Central & State Govt Taxes
- Pay EPF subscription online
- Utility Bill Payment
- Merchant Payments
- Easy synchronization with Corporate ERP System
- Provision for customized MIS
- Supply Chain Finance

OTHER SERVICES

- SBI Finder NEW
- Reprint Remittance Form - MOPS
- SBI CARD - SPECIAL OFFERS
- OnlineSBI Global
- Genuine Applications !
- Customer Complaint Form
- Secure Card Transactions
- Merchant Acquiring Business - POS
- SBI e-File ! Explore
- Foreign Travel/EZ-Pay/Gift Cards
- Mobile Banking Services
- NRI Services
- SBI Express Remit
- SBI F.A.S.T.
- Debit Card Loyality Rewardz
- SBICAP Trustee Company Ltd

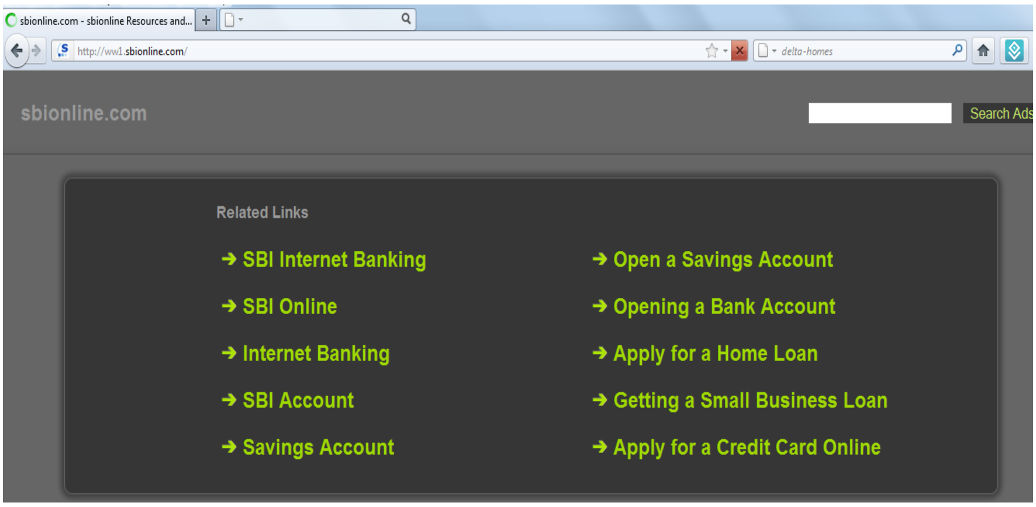
My WILL Services Online

USEFUL LINKS

- SBICAP Securities
- SBI Life Insurance
- SBI General Insurance
- SBI Mutual Fund
- SBICAP Trustee Company Ltd
- SBI Card

Figure.1.4The original website of SBI

- **The figure 2 shows the phished website of SBI**


sbionline.com - sbionline Resources and...
http://ww1.sbionline.com/
delta-homes
sbionline.com
Search Ads

Related Links

→ SBI Internet Banking
→ SBI Online
→ Internet Banking
→ SBI Account
→ Savings Account
→ Open a Savings Account
→ Opening a Bank Account
→ Apply for a Home Loan
→ Getting a Small Business Loan
→ Apply for a Credit Card Online

Figure: 2 – Phished website of SBI

III. MOTIVATION

Phishing website is a huge effect on the financial and online commerce, detecting this attack is an important step towards protecting against website phishing attacks. The motivation behind developing this system is that people are now days rely heavily on E-commerce website for buying anything to help them for protecting their personal information like username, password, credit card details, and social security number. To detect whether websites are Legitimate or Phishing this system will be developed. Hence there is a need for efficient mechanism for the Detection of phishing website. Phishing website is a very complicated and complex issue to understand and to analyses, since it is a combination of technical and social dynamics for which there is no known Single silver bullet to solve it entirely. Despite the great quantity of applications available for phishing website detection, there are only a few solutions that utilize machine learning mining techniques in detecting phishing websites.

IV. SUPERVISED LEARNING ALGORITHMS

Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object and a desired output value called the supervisory signal. A supervised learning algorithm analyses the training data and produces an inferred function, which is called a classifier. The classifier is then used for predicting the accurate output value for any valid unseen input object. The three classification algorithms used for learning the website data namely Multilayer perceptron, Decision tree induction, Naive Bayes are briefed below.

4.1 Multi-Layer Perceptron

Multilayer Perceptron network is the most widely used neural network classifier. MLP networks are general purpose, nonlinear models consisting of a number of units organized into multiple layers. The complexity of the MLP network can be changed by varying the number of layers and the number of units in each layer. Given enough hidden units and enough data, it has been shown that MLPs can approximate virtually any function to any desired accuracy.

4.2 Decision Tree Induction

Decision Tree Classification generates the output as a binary tree like structure called a decision tree. A Decision Tree model contains rules to predict the target variable. This algorithm scales well, even where there are varying numbers of training examples and considerable numbers of attributes in large databases. J48 algorithm is an implementation of the C4.5 decision tree learner. The algorithm uses the greedy technique to induce decision trees for classification [12]. A decision-tree model is built by analysing training data and the model is used to classify unseen data.

4.3 Naïve Bayes

The Naive Bayes classifier is designed for use when features are independent of one another within each class, but it appears to work well in practice even when that independence assumption is not valid. It classifies data in two steps (a) Using the training samples, the method estimates the parameters of a probability distribution, assuming features are conditionally independent given the class.(b) For any unseen test sample, the method computes the posterior probability of that sample belonging to each class. The method then classifies the test sample according the largest posterior probability.

4.4 Blacklist approach

Where the requested URL is compared with a predefined phishing URLs. The downside of this approach is that the blacklist usually cannot cover all phishing websites since a newly created fraudulent website takes considerable time before it is being added to the list. This gap in time between launching and adding the suspicious website to the list may be enough for the phishers to achieve their goals. Hence, the detection process should be extremely quick, usually once the phishing website uploaded and before the user starts submitting his credentials.

4.5 Heuristic approach

The second technique is known as heuristic-based approaches, where several features are collected from the website to classify it as either phishy or legitimate. In contrast to the blacklist method, a heuristic based solution can recognize freshly created phishing websites in real time. The effectiveness of the heuristic based methods, sometimes called features-based methods. Depends on picking a set of discriminative features that could help in distinguishing the type of website.

V. PHISHING WEBSITES FEATURES

There are several features that distinguish phishing websites from legitimate ones. In our study, we used 18 features descried briefly hereunder:

1. IP address: Using IP address in the hostname part of the URL address means user can almost be sure someone is trying to steal his personal information.

2. Long URL: Phishers resort to hide the suspicious part of the URL, which may redirect the information submitted by the users or redirect the uploaded page to a suspicious domain.

3. URLs having "@" symbol: The "@" symbol leads the browser to ignore everything prior it and redirects the user to The link typed after it.

4. Prefix and Suffix in URLs: Phishers deceive users by reshaping the URL to look like legitimate ones. A technique Used to do so is by adding prefix or suffix to the legitimate URL so users might not notice any difference.

5. Sub-domain(s) in URL: Another technique used by the phishers to deceive the users is by adding sub-domain(s) to the URL thus the users may believe that they are dealing with a credited website.

6. Misuse of HTTPs protocol: The existence of the HTTPs protocol every time sensitive information is being transferred reveals that the user certainly connected with an honest website. However, phishers may use a fake HTTPs protocol so that users might be deceived. In a recommendation to check whether the HTTPs protocol is offered by a trusted issuer such as "GeoTrust, Go Daddy".

7. Request URL: A webpage usually consists of a text and some objects such as images and videos. Typically, these objects are loaded to the webpage from the same domain where the webpage exists. If the objects are loaded from a domain different from the domain typed in the URL address then the webpage is potentially suspicious.

8. URL of Anchor: Similar to "Request URL" but for this feature the links within the webpage might refer to a domain Different from the domain typed on the URL address bar. This feature is treated exactly as "Request URL".

9. Server Form Handler "SFH": Once the user submits his information, that information will be transferred to a server to be processed. Normally, the information is processed from the same domain where the webpage is being loaded. Phishers resort to make the server form handler either empty or the submitted information are transferred to different domains.

10. Abnormal URL: If the website identity does not match its record shown in the WHOIS database (http://who.is/) the website is classified as "Phishy". This feature is a binary feature.

11. Redirect Page: This feature is commonly used by phishers by hiding the real link which asking users to submit Their information to a suspicious website.

12. Using Pop-up Window: It is unusual to find a legitimate website that asks users to submit their credentials through a popup window.

13. Hiding the Suspicious Links: Phishers resort to hide the suspicious link by showing a fake link on the status bar of the browser or by hiding the status bar itself.

14. DNS Record: If the DNS record is empty or not found the website is classified as "Phishy", otherwise it is classified as "Legitimate".

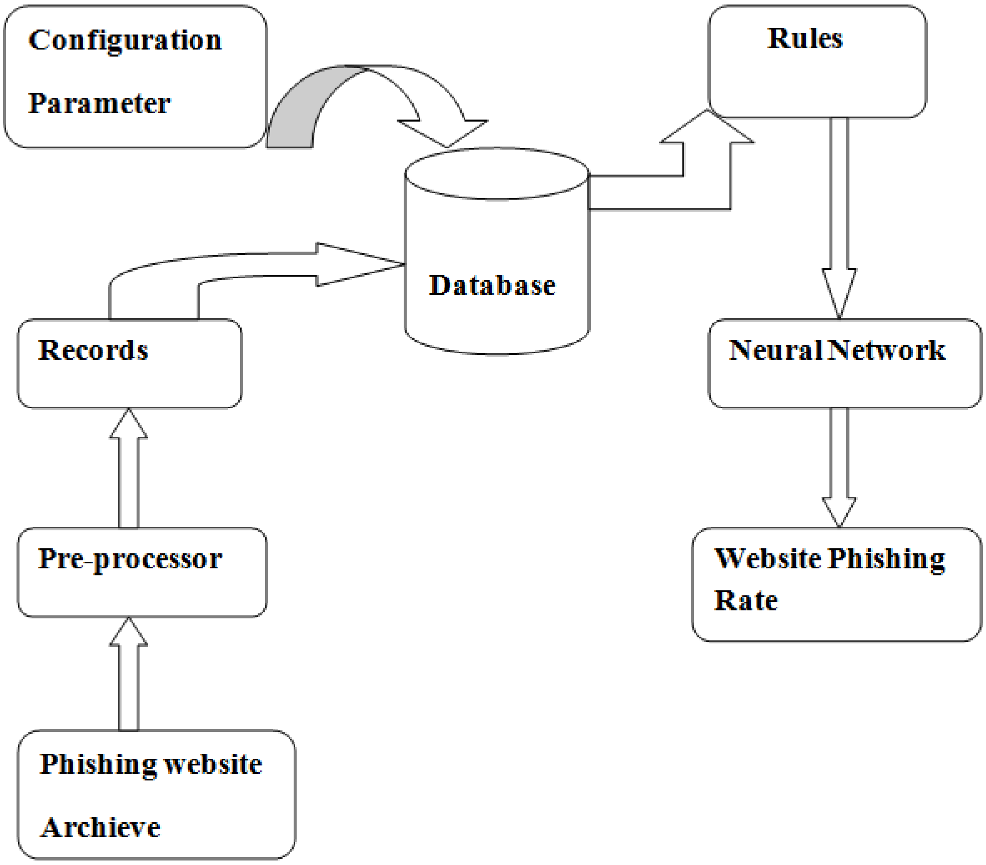
15. Website Traffic: Legitimate websites are of high traffic since they are visited regularly. Phishing websites often a Short life thus their web traffic is either not exists ranked is below the limit that gives it the legitimate status.

16. Age of Domain: the website is considered "Legitimate" if the domain aged more than 2 years [11].

17. Disabling Right Click: Phishers use JavaScript to disable the right click function, so that users cannot view and save the source code.

18. Port number: We examine if there is a port number in the URL and check if the port belongs to the list of well-known HTTP ports such as 80, 8080, 21, 443, 70, and 1080. If the port number does not belong to the list, we flag it as a possibly phishing URL.

VI. PROPOSED WORK



1) Phishing website Archived:

Initially all the phishing website details are collected and stored in the phishing website archive. Two publicly available phishing datasets were used to test our implementation: The “PhishTank” from the phishtank.com and yahoo Directory.

PhishTank is a collaborative clearing house for data and information about phishing on the Internet. Also, PhishTank provides an open API for developers and researchers to integrate anti-phishing data into their applications at no charge. It’s a free community site where anyone can submit, verify, track and share phishing data. The PhishTank database records the URL for the suspected website that has been reported, the time of that report, and sometimes further detail such as the screenshots of the website and is publicly available.

2) Pre-processor:

Then it is sent to a pre-processor to convert into machine understandable format. Initially Data are in text form URL , information about website and also screenshots of websites but which is not understand by machine so have to convert it in to machine under stable form.

3) Record:

The data comes from pre-processor which is in machine under stable form are store in Record and giving it to Database.

4) Data Base:

The result is then stored as records in the database. Blacklist Approach and white list Approach will be used.

In Blacklist approach the requested URL is compared with a predefined phishing URLs.

5) Configuration parameter:

The database also stores configuration parameters (the 27 phishing indicators that are being extracted from the code).

The six criteria of website phishing attack, there are several characteristics and factors that can distinguish the forged faked phishing website from original legitimate website like long URL address and abnormal DNS record, spelling errors etc. Phishing detection rate of a website is obtained based on six criteria and there are different components for each criterion, the criteria are as follows:

1. Encryption & Security.

- Certification authority.
- Using SSL certificate.
- Distinguished Names Certificate (DNC).
- Abnormal cookie.

2. Identity of URL & Domain

- Using the IP address.
- Abnormal URL.
- Abnormal request URL.
- Abnormal URL of anchor.
- Abnormal DNS record.

3. Contents &Page Style.

- Copying website.
- Spelling errors.
- Using forms with “Submit” button.
- Disabling right click.
- Using Popup windows.

4. Java script &Source Code.

- Redirect pages.
- Straddling attack.
- Server Form Handler (SFH).
- Using onMouseOver to hide the Link.

5. Social Human Factor.

- Public generic salutation.
- Much emphasis on security and response.
- Buying Time to Access Accounts.

6. Web Address Bar.

- Adding a prefix or suffix.
- Long URL address.
- Using hexadecimal character codes.
- Replacing similar characters for URL.
- Using @ symbol to confuse.

The aim of identity extraction is to extract the identity of a web page. Identity of a web page is a set of words that uniquely identifies the ownership of the website. Even though phishing artist can create and design replica of website, there are some identity relevant features which cannot be exploited. The change in these features affects the similarity of the website. Therefore these features are useful to find the identity of the web page. Features extracted in identity extraction phase include META Title, META
Description, META Keyword, HREF of <a> tag.

6) Rules:

Using the data collected in the database, rules are generated to detect the website phishing rate using the neural network

7) Neural network:

Once the neural network has been created it needs to be trained with the existing data in the archive. One way of doing this is initialize the neural net with random weights and then feed it a series of inputs.

This includes the types of connections within the network, the order of the connections and the values of the weights. One class of neural network architectures is the feed-forward networks. For this class, the data always propagate in unidirectional form starting from the input layer down to the output layer.

The other class of neural network architecture is the recurrent neural network, which contains feedback connections from units in the subsequent layers to units in the preceding layers. Recurrent networks have feedback connections between neurons of different layers and loop type self-connections. This implies that the output of the network not only depends on the external inputs, but also on the state of the network in the previous training iteration. Determining the network architecture is one of the difficult tasks in constructing any model but one of the most essential steps to be taken.

The advantage of multilayered perceptron is that the number of neurons in the hidden layer can be changed to adapt to the complication of the relationships between input and output variables. Although neural network construction has been widely researched, there is no known procedure or algorithm for the general case.

We will use MATLAB to train our model. MATLAB is a numerical computing environment and a programming language as well. The NN Toolbox is used to design, implement, visualize and simulate our NNs. MATLAB provides wide-ranging support for several NN paradigms, and graphical user interfaces (GUIs) supported by MATLAB enables the user to design NN in a very simple way.

8) Website phishing rate:

We then check to see what its output is and adjust the weights accordingly so that whenever it sees something looking like the existing data it outputs the same result as that data.

K-FOLD CROSS VALIDATION_is use to testing data and Training data are divided in different sets randomly.NN is trained using training data and tested on testing data. This process is continued until desired accuracy of NN is achieved.

VII. CONCLUSION

The prediction of phishing websites is essential and this can be done using neural networks and Fuzzy Logic. For the prediction of phishing websites, earlier works were done using various data mining classification algorithms were used but the error rate of those algorithms were very high. When an element of the neural networks fails, it can continue without any problem because of its parallel nature. Thus performance can be made better by considering neural networks as it reduces the error and gives better classification.

REFERENCES

[1] A.Martin, Na.Ba.Anutthamaa, M.Sathyavathy, Marie Manjari Saint Francois, Dr. Prasanna Venkatesan, a Framework for Predicting Phishing Websites Using Neural Networks. IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011.

[2] Maher Aburrous , M.A. Hossain , Keshav Dahal , Fadi Thabtah Intelligent phishing detection system for e-banking using fuzzy data mining, Science Direct , 2010 Elsevier Ltd. All rights reserved.

[3] Santhana Lakshmi Va, Vijaya MSb, a*, Efficient prediction of phishing websites using supervised learning algorithms, science direct, 2011 Published by Elsevier Ltd.

[4] Radheshyam Panda, Rajesh Tiwari, Protection from Phishing Attacks by Exploiting Page Rank, Reputation and Source Code of the Webpage, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 3, March 2014.

[5] Rami M. Mohammad • Fadi Thabtah • Lee McCluskey, Predicting phishing websites based on self-structuring neural Network, Received: 17 April 2013 / Accepted: 10 September 2013 / Published online: 21 November 2013_ Springer-Verlag London 2013

[6] Ms. Kranti Wanawe 1, Ms. Supriya Awasare 2, Mrs. N. V. Puri , An Efficient Approach to Detecting Phishing A Web Using K-Means and Naïve-Bayes Algorithms, International Journal of Research in Advent Technology, *Vol.2, No.3, March 2014.*

[7] Rami M. Mohammad1, Fadi Thabtah2, and Lee McCluskey , Predicting Phishing Websites using Neural Network trained with Back-Propagation , World Congress in Computer Science, Computer Engineering, and Applied Computing , Las Vegas, Nevada, USA, pp. 682-686. ISBN 1601322461.

[8] Mona Ghotaish Alkhozae, Omar Abdullah Batarfi , Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code, International Journal of Information and Communication Technology Research, ©2010-11 IJICT Journal. All rights reserved, Volume 1 No. 6, October 2011

Website:
[10] http://www.phishtank.com
[11] http://www.sbionline.com
[12] http:// www.onlinesbi.com