Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Phishing Site Detection Analysis Using Artificial Neural Network

To cite this article: M E Pratiwi *et al* 2018 *J. Phys.: Conf. Ser.* **1140** 012048

Phishing Site Detection Analysis Using Artificial Neural Network

M E Pratiwi[1], T A Lorosae[2] and F W Wibowo[3]

[1,2,3] Magister of Informatics Engineering, Universitas Amikom Yogyakarta, Indonesia.

E-mail: martina.endah05@gmail.com

Abstract. Phishing site is a fake website that is designed specifically to provide false information or instructions. One of the sites most targeted for phishing scams by cybercriminals is an e-commerce. However, there is also a social media-based website that is a target for cybercriminals. Phishing sites are also used to commit fraud action on behalf of the original site and as a medium for distributing malware/virus computer by cybercriminals. Based on these problems, the researchers will conduct an analysis of phishing sites using neural network perceptron algorithm to determine the value of accuracy, precision and recall value.

1. Introduction

The number of phishing sites has been detected in the fourth quarter was 180.577 sites based on the APWG (Anti-Phishing Working Group) report. At the end of 2016, phishing sites were found on https infrastructure less than 5%. In 2017, one-third of total phishing attacks discovered attacks on websites that had https and SSL certificates, but there were still many attacks on the domain as much as 20% [1].

A phishing is a fake website designed similar to the original site. Display or phishing website interface will make unsuspecting victims. In addition, the phishing website can also be designed to provide information or further instructions from the original website by asking the victim to fill certain data required by phisher via the form available on the website of the phishing [2][3]. Phishing also has an impact as a victim of data loss and the data used by the phisher and the impact that could have fatal namely to harm the victim of financial terms [2]. The objective of these phishing websites is as varied as social media, telecommunications, hosting, financial institutions, webmail, payment, e-commerce and much more.

Previous research revealed that human resources are also an important factor for the prevention of phishing sites, the study also discusses phishing on social networking sites and explain the details of the risk of phishing in companies [4]. Other studies are testing whether conceptual knowledge or procedural knowledge has a positive effect on the user to thwart phishing threats and the study found that the effects of the interaction of conceptual and procedural knowledge have a positive impact on users' computers with the security education [5]. While the research conducted by [6] developed a resistance against phishing attacks. The aspect of the security in both side of server and client is also very important to do in overcoming of the vulnerability of confidentiality and integrity of data [7].

This research will conduct an analysis to detect phishing sites using neural network perceptron algorithm. Based on previous research, neural network perceptron algorithm widely implemented in the study because the neural network perceptron has the ability to model the nonlinear system with conditions very complex relationship between variables [8].

2. Problem of Phising

2.1 Phising

The threats on the Internet that uses the host program, virus, phishing and more. Phishing is a new attack that could threaten the security of user data, which the user or the victim will be redirected to another website. At first, the phisher will create a fake website and send hundreds of emails containing links phony websites, and the victim would visit a link that is believed to be the link is legitimate, then the victim will include the identity or data, the offender will utilize the identity or the data to commit a crime which benefits the phisher as account sales, even falsification of data [9]. The flow diagram of the phising works is shown in figure 1.

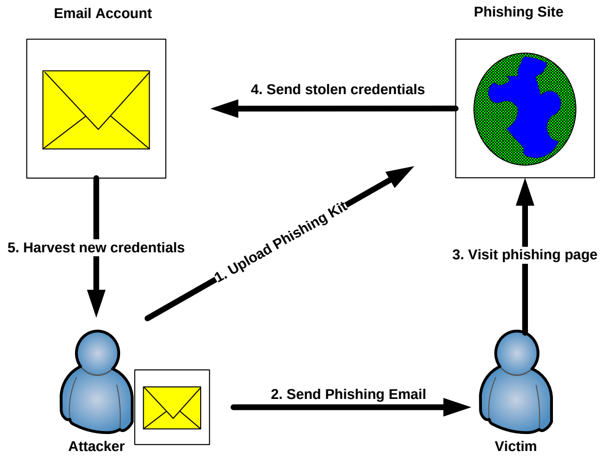


Figure 1. How phising works

2.2 Neural Network

The mimic human brain function is a concept of Neural Network. Neurons are thousands and even millions of small processing contained in the human brain are connected to one another through the connection of neurons. In the human brain, there is a set of neurons that can be used as input and output is processed and passed through each set of neurons [10]. This architecture could be shown in Figure 2.

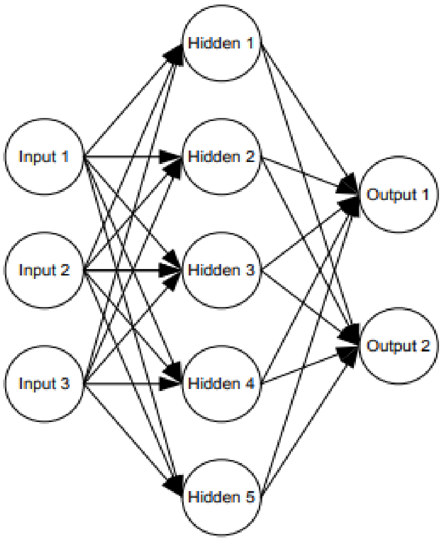


Figure 2. Neural Network Perceptron Architecture

Feed Forward Neural (FNN) network perceptron and Back Propagation (BPP) Neural network perceptron does not have a significant difference, the difference between the two is the process of calculating the weight for each link.

3. Data and Experimental Result

3.1 Research Method

3.1.1 Research Design. This research is an experimental research investigation using parameters or variables that have been defined and the use of the test or tests with the research method as follows:

1. Data collection
   Finding data is available, obtain additional data required and determine the required variables.
2. Data processing
   Perform data selection, do a cleaning of data for subsequent modeling.
3. The method used
   Perform data analysis, grouping variable, then made application of the model that corresponds to the type of data and sharing of data into two parts, namely the training data (training data) and test data (testing data).
4. Experiment
   Tests on a model that has been determined to make decisions or find out the results.
5. Evaluation and validation
   To evaluate the model has been determined to know the accuracy of the model.

3.1.2 Data collection. The uci repository provides secondary data used for the study. Data obtained by 2,455 phishing sites, which will be divided into two as training data as much as 80% or 1964 data and test data as much as 20% or 491 of data [11].

Data phishing site has 18 features [11]. Such features are as follows: having an IP address, URL length, shortening service, having @ symbol, the double slash redirecting, prefix suffix, having subdomains, SSL final state, domain registration length, favicon, port, HTTPS token, request a URL, the URL of anchor, links in tags, SFH, submitting to email and abnormal URL. This feature will be categorized as shown in Table 1.

Table 1. Features

Parameter	Features	Value Features
1 = valid 0 = suspicious -1 = phishing	Having IP Address	-1,1
	URL Length	1,0, -1
	Shortening Service	1, -1
	Having @ symbol	1, -1
	Double Slash Redirecting	-1,1
	Prefix Suffix	-1,1
	Having Sub Domains	-1,0,1
	SSL final State	-1,0,1
	Domain Registration Length	-1,1
	Favicon	1, -1
	Port	1, -1
	HTTPS Token	-1,1
	request URL	1, -1
	URL of Anchor	-1,0,1
	Links in Tags	1, -1,0
	SFH	-1,1,0
	Submitting to Email	-1,1
	abnormal URL	-1,1

Source: uci repositoy

From the data in Table 1, to determine phishing sites with non phishing sites using 18 features with the following explanation:

1. Having an internet protocol (IP) Address
   If the IP address is used as the domain of the URL, such as http://125.98.2.142/contoh.html, it can be suspected that attempts to steal information. The IP address can also be converted into hexadecimal code.

$$rule\text{: IF}\begin{cases}\text{If The Domain Part has an IP Address } \rightarrow \text{ Phishing} \\ \text{Otherwise} \rightarrow \text{ Legitimate}\end{cases}$$

2. URL Length
Long URLs can also be suspected of being a phishing site like this one:

http://mencobasitus.com.nr/4c/rea/4b53e4i6f913e51234hfyg46f363r734/?cmd=_home& dispatch=1212325vdvtyvwtew5wtetuyuijba5672uh2bi2822gy267gehh74y7@phishing. website.html. If the URL length greater than or equal to 54 characters then the URL included as a phishing site.

$$\textit{Rule: IF}\begin{cases} URL\ length < 54 \ \rightarrow\ feature = \text{Legitimate} \\ else\ if\ URL\ length \geq 54\ and\ \leq 75\ \rightarrow\ feature = \\ Suspicious \\ otherwise \rightarrow\ feature = \text{Phishing}\end{cases}$$

3. Shortening Service
URL shortening is a method in which a URL is made to be shorter, which this domain will connect to the web page that has a URL that is longer such, http://sekolah.ini.ac.id/ URLs can be shortened to "bit.ly/21FXWl5",

$$rule\text{: IF}\begin{cases}\text{TinyURL } \rightarrow \text{ Phishing} \\ \text{Otherwise} \rightarrow \text{ Legitimate}\end{cases}$$

4. Having @ Symbol
URLs using the symbol @ will lead to the browser to ignore everything that precedes the @ symbol.

$$\text{Rules: IF}\begin{cases}\text{Url Having @ Symbol} \rightarrow \text{ Phishing} \\ \text{Otherwise} \rightarrow \text{ Legitimate}\end{cases}$$

5. Double Slash Redirecting
Double slash or "//" indicates that the user or the user will be redirected to another site. The position of the use of double slash usually appears at the sixth position as written at this link http://amikom.ac.id. However, if the double slash appears in the seventh position as https://amikom.ac.id it can be suspected as a phishing site.

$$\text{Rules: IF}\begin{cases}\text{The Position of the Last Occurrence of "//"} \\ \text{in the URL } > \ 7 \rightarrow Phishing \\ \text{Otherwise} \rightarrow \text{ Legitimate}\end{cases}$$

6. Prefix Suffix
Rarely a legitimate URL using symbols dashboard, but phisher will add a prefix or suffix to be separated by (-) in the domain name, so the user will think to have a legitimate access to sites such as http://www.amikom-keren.com.

Rules: IF $\begin{cases} \text{Domain Name Part Includes } (-) \\ \text{Symbol} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

7. Having Sub Domains
The domain name may have a code for each country (cc TLD) such as "id", or for an academic educational institution "ac" and combined "ac.id" or also called two-level domain (SLD). Stages for extracting the feature of the first to do is remove the "www" in the URL and remove cc DTL if any. Then calculate the remaining dots, if the number of points is greater than one, then the URL can be classified as "suspect" because of only the subdomain. However, if the number of points greater than the two it will be categorized as a phishing because it has several subdomains, and sites categorized as legitimate if it does not have a subdomain.

Rules: IF $\begin{cases} \text{Dots In Domain Part} = 1 \rightarrow \text{Legitimate} \\ \text{Dots In Domain Part} = 2 \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases}$

8. HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)
HTTPS is an essential component in a site that looks at legality.

Rules: IF $\begin{cases} \text{Use https and Issuer Is Trusted } and \\ Age\ of\ Certificate \geq \text{ 1 Years} \rightarrow \text{Legitimate} \\ \text{Using https and Issuer Is Not Trusted} \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases}$

9. Domain Registration Length
Categorized as phishing sites are valid only in a short time and are used for a single year.

Rules: IF $\begin{cases} \text{Domains Expires on} \leq \text{ 1 year} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

10. Favicon
Favicon is an image used as an icon on a website, favicon also indicates the identity of the website. However, if the favicon is displayed apart in the address bar, it can be suspected that the website is a phishing website.

Rules: IF $\begin{cases} \text{Favicon Loaded From External Domain} \rightarrow \\ \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

11. Port
Port used to validate certain services such as HTTP. The use of a firewall, proxy, and Network Address Translation or NAT can perform automatic blocking and can be opened in accordance with the wishes. But if all the ports are opened, then the phisher will find loopholes and enable any desired services such as stealing information.

Rules: IF $\begin{cases} \text{Port \# is of the Preferred Status} \rightarrow \text{ Phishing} \\ \qquad\qquad \text{Otherwise} \rightarrow \text{ Legitimate} \end{cases}$

12. HTTPS Token
    In general, the https token can be added by phisher on the domain URL and has the objective to distract the user like at http://https-www-amikom-coolest-college.com/

Rules: IF $\begin{cases} \text{Using HTTP Token in Domain Part of The URL} \\ \qquad\qquad \rightarrow \text{ Phishing} \\ \qquad \text{Otherwise} \rightarrow \text{ Legitimate} \end{cases}$

13. Request URL
    On the website is legal, website addresses, pictures, videos, and sounds contained on the web page with the same domain and does not take away from another domain.

Rules: IF $\begin{cases} \text{\% of Request URL } < 22\% \rightarrow Legitimate \\ \text{\%of Request URL} \geq 22\% \text{ and } 61\% \rightarrow \text{ Suspicious} \\ \text{Otherwise} \rightarrow \text{ feature} = \text{Phishing} \end{cases}$

14. URL of Anchor
    If the tag is <a> and the website has a different domain name, it can be suspected as phishing.

rule: IF $\begin{cases} \text{\% of URL Of Anchor } < 31\% \rightarrow Legitimate \\ \text{\% of URL Of Anchor } \geq 31\% \text{ And} \leq 67\% \\ \qquad \rightarrow \text{ Suspicious} \\ \text{Otherwise} \rightarrow \text{ Phishing} \end{cases}$

15. Links in Tags
    In general, legal sites also use <meta> tag, the <script> tag and <link> tag. And the tag comes from the same domain.

Rules: IF $\begin{cases} \text{\% of Links in <Meta>, <Script> } and \text{ " < Link>"} \\ \qquad < 17\% \rightarrow Legitimate \\ \text{\% of Links in } < Meta > , < Script > \text{ and} \\ < \text{Link>" } \geq 17\% \text{ And} \leq 81\% \rightarrow \text{ Suspicious} \\ \qquad \text{Otherwise} \rightarrow \text{ Phishing} \end{cases}$

16. Server Form Handler (SFH)
    SFH is an empty string or the website terms "about: blank", on the website of suspected phishing, domain name on SFH will vary with the domain name of the website pages.

Rules: IF $\begin{cases} \text{SFH is "about: blank" Or Is Empty} \rightarrow \text{ Phishing} \\ \text{SFH Refers To A Different Domain} \rightarrow \text{ Suspicious} \\ \qquad \text{Otherwise} \rightarrow \text{ Legitimate} \end{cases}$

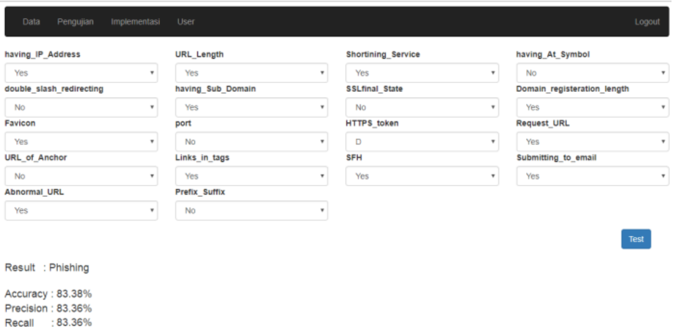
17. Submitting to Email
    The official website will generally send personal information to the server for processing. While the phisher will be sending the information to his personal email, it can be suspected by the use of scripts on the server side functions such as "mail ()" and on the client side will use the mailto.

    Rules: IF $\begin{cases} \text{Using "mail()" or "mailto:" Function to Submit} \\ \quad\quad \text{User Information} \rightarrow \text{Phishing} \\ \quad\quad \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

18. Abnormal URL
    On the website is legitimate, then the identity of the website will be contained in the URL.

    Rules: IF $\begin{cases} \text{The Host Name Is Not Included In} \\ \quad\quad \text{URL} \rightarrow \text{Phishing} \\ \quad \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

3.1.3 Stage Analysis. Stages of analysis in this study are as follows:
1. Data collection.
2. Preprocessing data.
3. Analyzing the characteristics of phishing sites and non-phishing.
4. Determining the neural network perceptron parameters and enter the parameter value.
5. Classification of data using a neural network.
6. Tests on the 2,455 data by methods that have been determined to get the classification of phishing sites as shown in Figure 3.



having_IP_Address	URL_Length	Shortining_Service	having_At_Symbol
Yes	Yes	Yes	No
double_slash_redirecting	having_Sub_Domain	SSLfinal_State	Domain_registeration_length
No	Yes	No	Yes
Favicon	port	HTTPS_token	Request_URL
Yes	No	D	Yes
URL_of_Anchor	Links_in_tags	SFH	Submitting_to_email
No	Yes	Yes	Yes
Abnormal_URL	Prefix_Suffix		
Yes	No		

Result : Phishing

Accuracy : 83.38%
Precision : 83.36%
Recall : 83.36%

Figure 3. Result from the system

The website is judged based on 18 features for the phishing website that has been determined with the value "Yes" or "No", then the percentage of accuracy of the algorithm used in the phishing site detection system will appear. Based on testing using neural network perceptron analysis then showed the value of accuracy, precision, and recall as shown in Table 2.

Table 2. Result of testing neural network

Accuracy	Precision	Recall
83,38%	83,36%	83,36%

From the results of the study to detect phishing sites using a neural network, the result 83,38% accuracy, precision 83,36% and recall 3,36%.

4. Conclusion

From the test results in studies using neural network perceptron algorithm to detect phishing sites, has obtained results 83,38% accuracy, precision 83,36% and recall 83,36%. With these result can conclude that the neural network perceptron algorithm can be further used to design a system to detect phishing sites.

5. References

[1] https://www.antiphishing.org/resources/apwg-reports/, accessed on July 10, 2018 at 19:52 hours GMT

[2] Purwiantono and A Tjahyanto, Classification Model to Detect Phishing Sites in Indonesia

[3] D Rachmawati 2014 Phishing Form As One In The World Cyber Threats *Saintikom Journal*, Vol 13 No 3 ISSN: 1978-6603

[4] M SILIC and A Back 2016 The dark side of social networking sites: Understanding phishing risks *Computers in Human Behavior* Vol 60 p 35-43, ISSN 0747-5632

[5] Arachchilage N G and Steve L 2014 Security Awarness of Computer Users: A phishing Threat Avoidance Perspective, *Computers in Human Behavior* Vol 38 p 304-312

[6] M S Narendra, C Shah, M Mahajan, S Rachh 2015 An Ideal Approach for Detection and Prevention of Phishing Attacks, *Procedia Computer Science* Vol 49 p 82-91

[7] H Utama, F W Wibowo 2015 Security Specification of WS-SecureConversation, 2015 *IEEE Student Conference on Research and Development (SCOReD 2015*) p 690 – 695

[8] S Widodo 2017 Classification of Phishing Sites by Using Neural network perceptron and K- Nearest Neighbor *Information Management for Educators and Professionals* Vol 1 No 2 p145-154 E-ISSN: 2548-3331

[9] K R Sahu and J Dubey 2014 A Survey on Phishing Attack *International Journal of Computer Applications* Vol 88 No 10

[10] Hartatik, F W Wibowo, O Antoro 2018 Implementation of Recognizing Batik Motif Pattern Based-on Wavelet Transforms Comparison and Neural Network *Journal of Advanced Manufacturing Technology*

[11] https://archive.ics.uci.edu/ml/machine-learning-databases/00327/Training%20Dataset.arff, accessed on July 1, 2018 at 7:39 o'clock pm.