PHISHING DETECTION FROM URLS BY USING NEURAL NETWORKS

Ozgur Koray Sahingoz, Saide Işılay Baykal and Deniz Bulut

Department of Computer Engineering, Istanbul Kultur University,
Istanbul, Turkey

ABSTRACT

In recent years, Internet technologies are grown pervasively not only in information-based web pages but also in online social networking and online banking, which made people's lives easier. As a result of this growth, computer networks encounter with lots of different security threats from all over the world. One of these serious threats is "phishing", which aims to deceive their victims for getting their private information such as username, passwords, social security numbers, financial information, and credit card number by using fake e-mails, webpage's or both. Detection of phishing attack is a challenging problem, because it is considered as a semantics-based attack, which focuses on users' vulnerabilities, not networks' vulnerabilities. Most of the anti-phishing tools mainly use the blacklist/white list methods; however, they fail to catch new phishing attacks and results a high false-positive rate. To overcome this deficiency, we aimed to use a machine learning based algorithms, Artificial Neural Networks(ANNs) and Deep Neural Networks(DNNs), for training the system and catch abnormal request by analysing the URL of web pages. We used a dataset which contains 37,175 phishing and 36,400 legitimate web pages to train the system. According to the experimental results, the proposed approaches has the accuracy in detection of phishing websites with the rate of 92 % and 96 % by the use of ANN and DNN approaches respectively.

KEYWORDS

Phishing Detection System, Artificial Neural Networks, Deep Neural Networks, Big Data, Machine Learning, Tensor flow, Feature Extraction

1. INTRODUCTION

Due to the extensive growth in the number of internet users, lots of our daily life operations are transferred from the real world to the cyber world such as communication, coordination, commerce, banking, registrations, applications, etc. Because of this, the malicious peoples and attackers also transferred to this world and make their threats and crimes easily anonymously. To ensure the security and privacy of cyber data, technology must be used and organized carefully by using "Cyber Security" concept [1].

According to ITU-T, cyber security is the accumulation of tools such as policies, security safeguards, training, risk management approaches guarantee and technologies that can be used to protect the cyber organization and environment. [2] Another source explains this concept as follows: Cyber security is the body of technologies about processes, networks, computers programs and data. Its aim is designed for protect these components of technologies from attack, damage and unauthorized access [3]. According to Craigen et.al. cyber security is the organization and collection of resources, processes and structures used to defend cyberspace and cyberspace enabled systems from events that misrelate by default ownership rights [4]. Cyber

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 41–54, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81705

security applies precaution methods used to protect data from being stolen, concurred or attacked [5]. All the definitions of cyber security say about prevent and protect: Cyber security prevents from fraud or thief who wants to seize person/public/national information or connection.

“Identity theft” or specifically “phishing” is one of the most threatening security deficits of the users in the Internet. In this type of crimes, attackers use some malicious web pages which impersonate as legitimate web sites, to collect the victims’ critical information such as username, passwords, financial data, etc. Typically, a phishing attack starts with an electronic mail which seems to come from a reputable company as depicted in Figure 1. The content of the mail encourages the victim to click on the address, which can also be hidden as a hypertext. This address directs the victim to a fake web site, which is designed exactly similar with a valid website, such as an e-mail site social engineering site of generally financial institutions web sites.

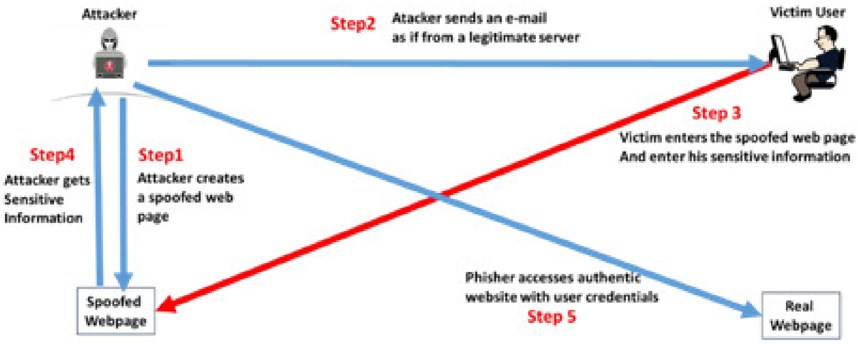


Figure 1: Life Cycle of a Phishing Attack

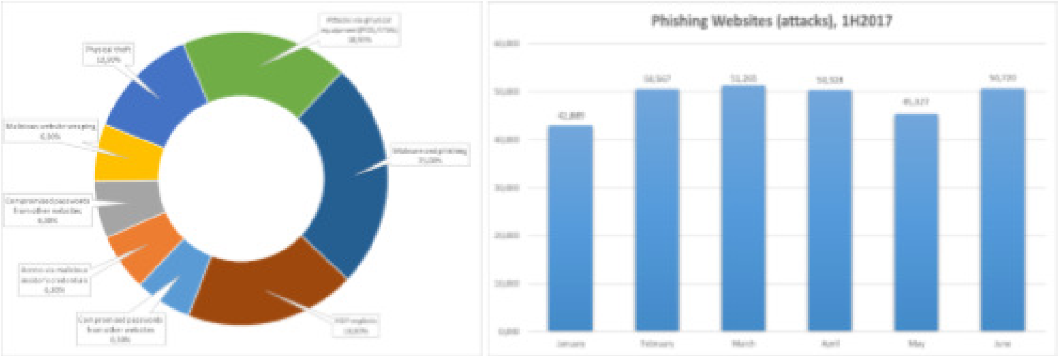
As can be seen from this life cycle, even experienced computer user can fall into the phishing attack and be a victim. Therefore, for detection of phishing attacks, a dynamic support and security mechanism is needed. As a phishing detection algorithm, generally blacklists/whitelists are used. This is an effective prevention mechanism and it quickly classify an URL as a phishing or legitimate. However, as emphasized in [18] between 47% and 83% of phishing web pages are blacklisted in 12 hours, which is enough duration for deceiving most of the people. Additionally, within the first 2 hours, about 63% of phishing campaigns are finished. Therefore, blacklists/whitelists are not effective especially for zero-day attacks.

To overcome this type of attack there is need to construct a dynamic and efficient algorithm which can learn the structure of the legitimate web pages and classifies the abnormal ones. Therefore, in this project, we aimed to set up a classification system, which can identify whether an URL is either phishing or legitimate. To train the system we have used a dataset which contains about 74,000 items in both these types. To compare the efficiency of the different algorithms and select the best one, we used both Artificial Neural Network (ANN) and Deep Neural Network(DNN) approaches for training and testing the system with the help of Tensorflow framework. And experimental results showed that the proposed approaches produce very good accuracy rates for detecting phishing URLs. Within the proposed approaches, DNN gives better accuracy rate than ANN with the related values as depicted in the results section.

The rest of the paper is organized as follows: In the next section the background knowledge is given. Section 3 depicts the design details of the proposed system. Experimental Result are shown with comparative graphic in Section 4. Finally, Conclusion and future works are listed

2. BACKGROUND

In oxford dictionary, phishing means "an effort by hackers to destroy or damage a computer system or network". It means broking the "confidentiality, integrity, and availability"-CIA triad rules. In the real world there are many attack types for broking this CIA such as Sniffing, Denial of Service (DoS), Sql Injection, Spyware, Viruses, Trojans, Social Engineering, Worm, Botnet and Phishing [8]. However, as can be seen from Figure 2.a. Phishing attacks are located at the first position.

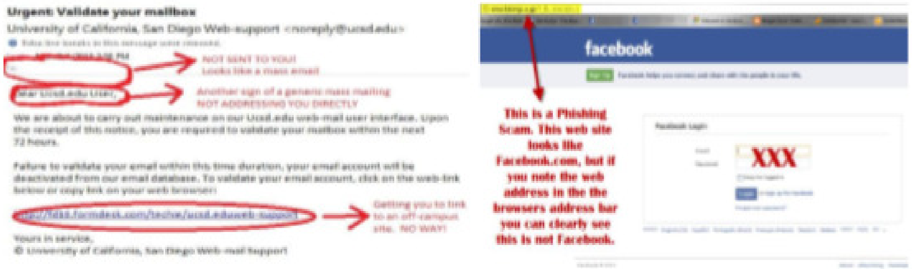


a) Phishing vs other attacks

b) # of Phishing web sites in 1st half of 2017

Figure 2. Phishing Statistics

Also, as can be seen from Figure 2.b. Phishing attack is a continuing process, in every part of the year this attack takes its place in the cyber world. Phishing is an attack type using both social engineering and technical hints to have users' personal identity information and bank account details [9]. There are many phishing attack types in the literature. The most preferred one the use of emails. Attacker prepares an email which urges the user for entering his valuable information on a malicious webpage as depicted in Figure 3.a. In this e-mail there are some hyperlink which directed the user to this malicious webpage, which is exactly similar with the original one. After the user enters the information on the webpage, attacker can access the victim's sensitive information



Urgent: Validate your mailbox

University of California, San Diego Web-support <noreply@ucsd.edu>

Dear Ucsd.edu User,

We are about to carry out maintenance on our Ucsd.edu web-mail user interface. Upon the receipt of this notice, you are required to validate your mailbox within the next 72 hours.

Failure to validate your email within this time duration, your email account will be deactivated from our email database. To validate your email account, click on the web-link below or copy link on your web browser:

http://formdesk.com/techie/ucsd.eduweb-support

Yours in service,
© University of California, San Diego Web-mail Support

NOT SENT TO YOU!
Looks like a mass email

Another sign of a generic mass mailing
NOT ADDRESSING YOU DIRECTLY

Getting you to link to an off-campus site. NO WAY!

This is a Phishing Scam. This web site looks like Facebook.com, but if you note the web address in the the browsers address bar you can clearly see this is not Facebook.

a)Phishinge-Mail

b)SpoofedWebpage

Figure3.DeceivinguserwithE-mailandspoofedwebpage

2.2. TYPES OF PHISHING ATTACKS

Phishing attacks can be divided in 2 layers: social engineering and technical subterfuge. Social engineering layer includes attackers, victim, sending fake email, which contains spoofed webpages. This process starts by sending this email, which comes from a legal and famous organizations for gathering some sensitive information such as user name, id, password, credit card information etc. Second layer is about spoofed webpage. Fake e-mail directs the victim to the spoofed webpage which appears visually very similar to the original webpage. This layer also uses cross-site scripting, session hijacking, malware phishing, DNS poisoning and key/screen loggers' techniques. These layers send the obtained information and get remote access by attackers to victim's computer or original webpage [12, 13]. According to [14], mostly attacked websites are shown in Figure 4.

.

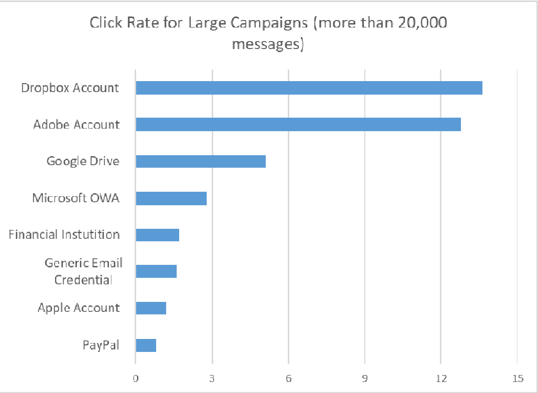


Figure 4: According to statistics of company phishing

2.2. DETECTION OF PHISHING ATTACKS

Phishing attacks can be applied by using a lot of methods. Detection ways have been found on the basis of this attack types. This section will explain the detection methods of phishing attacks. The main vulnerability of phishing is about the Human Factor. Therefore, the main prevention is about the education of the workers, how to avoid from this type of attacks. However, due to the type of new attacks, even experienced used can fall into this type of attack. Therefore, a cyber support will be helpful for the users.

The mostly preferred methods to prevent phishing is the use of Blacklists, which are periodically updated list which includes some keywords lists, URLs and IP addresses. The famous blacklist using methods are: Google Safe Browsing API, DNS-Based Blacklist, Phish Net: Predictive Blacklisting, Automated Individual White-List. However, due to its deficiency for detecting zero day attack, some security managers prefer the use of Heuristics approach, which analyses and investigates the feature of the web page and detect whether the page use this information or not [19]. The reputable heuristics anti-phishes are Spoof Guard, Collaborative Intrusion Detection, Phish Guard, Phish wish, CANTINA, and etc.

Visual Similarity method uses the visual similarity of the webpage like its source code, contained pictures, text and additionally some formatting, logo, CSS and HTML tags, etc. These features are compared with the previous form of the web page or its stored copy in the local server. However, this technique has an important deficiency that it cannot detect the phishing attacks of the newly generated web pages. Besides, its image-based operation, comparison gets too much time for detection.

The dynamic approach can be seen as the use of data mining and/or machine learning techniques. If there are sufficient number of legitimate and non-legitimate web pages and their related features, it can be easy to train the system with this dataset by the use of some machine learning techniques. Support Vector Machines, Bayesian Classifier, KNN techniques, Ad boost, Random Forest, decision tree, neural networks, etc.

2.3. NEURAL NETWORKS

Machine learning is one of the very important field of computer science, which allows software to learn and adapt to inputs and improve performance on a specific task. Machine learning is highly used to follow human behaviours and to make some predictions by using either supervised or unsupervised algorithms. Neural networks are designed influenced from biological neural networks. In real neurons, the input data are processed and transmitted by use of electrical signals. In artificial neural networks, system works with input nodes –it is called as neuron-, edges as functions, layers, and output neurons. All these components related with nodes and edges. Input neurons connecting other neurons via functions. A simple diagram of a neuron is shown in Figure 5.

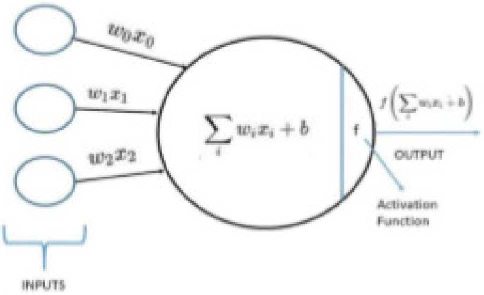


Figure 5: A simple neuron structure.

Even though given inputs are the same, weight and bias criteria can be changed to calculation. Almost all neurons calculate for the next neurons by that formula. And they are collecting activation functions such as RELU, TANH, etc. According to activation functions all these multiplication and addition process collecting fully connected layer. Than predicting output decreased by some loss functions. This output is gathering and comparing real value. At the end of the output, this result optimizing and so on. Figure 6 shows structure of neural networks. [15]

Neural networks are divided into two sub networks, which artificial neural networks and deep neural networks, which use multiple layer in its framework as depicted in Figure 6. According to the parameters and size of the problem the number of hidden layers and also the number of neurons in each layer can be changed. If you only use a single hidden layer, this is mainly called as ANN structure.

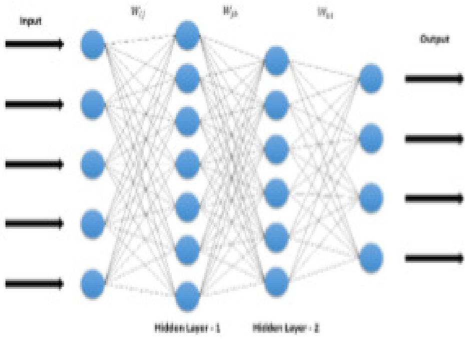


Figure 6: Two Hidden Layer Deep Neural Network

2.4. RELATED WORK

There are many works in the literature, which are focussed on phishing detection. According to study of James et. Al. attackers are swindled via e-mails to get individual information and bank account details like usernames, userid and passwords [16]. This paper contains information about machine learning methods, which is used for detecting phishing websites. In this paper two success rate are analysed which is WEKA and MATLAB. The J48 Decision Tree gave best result in WEKA. When dataset is splitted 60% for testing, detection accuracy was 93.2% in lexical features. Regression Tree was given best result with 91.08% accuracy in MATLAB when using 40% dataset for training however accuracy was decreased when using 10% of dataset for testing.

Buber et.al. suggest that, cyber-attacks affect to many people and foundation and this attack can cause financial damages in this work [17, 21, 22]. There are a lot of cyber-attack types. Purpose of phishing attack, which is one of them, is getting confidential information of users by using people's weaknesses. In this paper, a machine learning based system was developed for detecting phishing attacks. Some features were generated by using taking advantages of Natural Language Processing (NLP) in this system. For detecting URL which is used in phishing attacks, a system was developed by using these features. According to tests Random Forest Algorithm showed the highest result success rate.

3. METHODOLOGY

In the implementation phase we developed two different classifiers with: Artificial Neural Networks and Deep Neural Networks. Due to their structure we need to use some numeric values as the input of our system. Therefore, we need to select some features from the URL and then train and execute our system based on these parameter values.

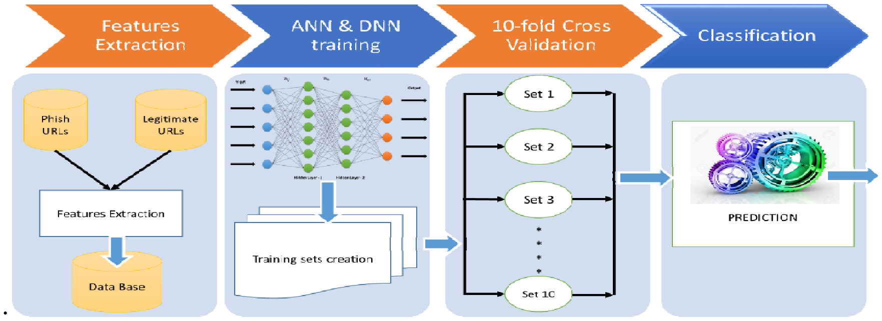


Figure 7. Execution Diagram of the Proposed System

To understand the meaning of each feature, firstly we need to identify the parts of URLs. In the next subsection, this concept is explained. After that, the selected features are detailed.

3.1. URLS

To understand the approach of phishers, firstly, the components of URLs and their aim should be understood. The basic components of a URL is depicted in Figure 8.

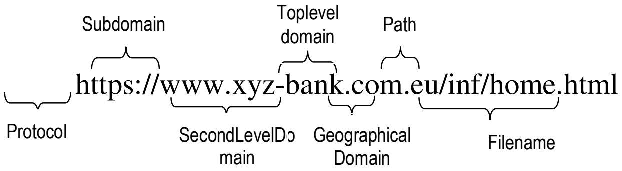


Figure 8. Components of a URL

In the standard form, a URL starts with its protocol name, such as hypertext transfer protocols, file transfer protocols, etc., which are used to access the web page. Consequently, the subdomain and the Second Level Domain (SLD) names identify the server hosting the web site. SLD name is very important for us, because this part mainly contains the name of the firm, therefore, phishers focussed on this part and try to produce different forms of name which are like original ones. The Top-Level Domain (TLD) name shows the domains in the Domain Name System root zone of the Internet such as educational, commercial government, etc. Finally, Geographical Domain name shows the geographical location of the web site such as, Germany, Turkey, France, etc.The previous four parts compose the domain name (host name) of the web page; however, the inner address is represented by the path of the page in the server and with the name of the page in the html form. The ongoing part is like a folder a file name which shows the location of the file in the server.

3.2. SELECTED FEATURE

In this subsection, we detailed the selected features that are used in the implementation of the proposed system. There are total 27 features, and they are detailed as follows.

1. Length of the URL: Phishers generally hide the address of their spoofed web page by increasing the length of the address. In this long text they also add the name of the attacked web page, but this is not the domain name part of the URL. Additionally, if this length is increased too much, then it will not fit the address bar, and the victim cannot see the domain part. Some researchers focussed on this size and they grouped the URL according the following rule [20]:
If the length of the URL<54, then it is classified as “legitimate”, If the length of the URL is between 54 and 75, then it is classified as “suspicious”, If the length of the URL>75, then it is classified as “phishing”,

However, in our study, we don’t make this type of classification. Classification is executed by the classifier, and this value is only a parameter for our classifier. Shorter URLs have the greater possibility for being “legitimate”.

2. Punctuation character count: Phishers use some meaningless characters for confusing the victim. Therefore, they can also use some punctuation characters, especially “.”, ”;”, ”!”, ”&”, ”%”, etc. Increased value has more tendency to be a phishing webpage as depicted in Figure 9.

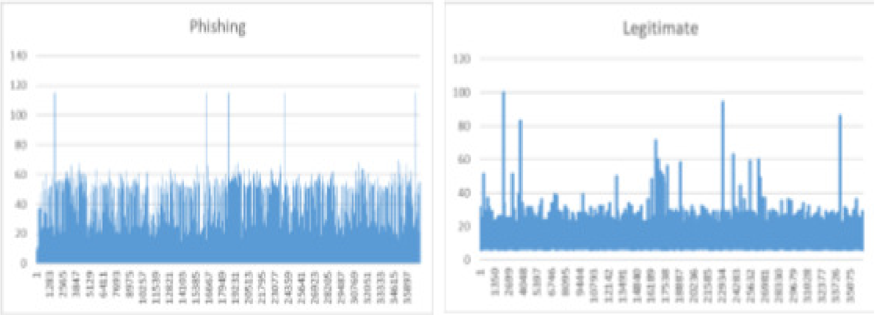


Figure 9. Number of Punctuation Characters in the URLs

3. Is it an IP address: This is a binary feature, if it is an IP address then its value is 1, else 0. For deceiving the users, phisher generally try to use some part of the original URL in the spoofed URL addresses. Therefore, they do not prefer the use of IP address in their attack.

4. Suspicious words count: Phisher prefers some specifics words such as ’confirm’, ’account’, ’secure’, ‘admin’, ’login’, ’submit’, ’update’, ‘setup’, ’secure’, etc. These words helps for the victims to think the related web page is legitimate. Therefore, we get ?? suspicious words which are selected in the study of Buber et.al. The number of these words are used as a feature in our system.

5. Alexa ranking: There are more than 1.7 billion websites all around the world. Alexa holds popular websites and ranking them. Generally, the popular websites are not preferred for phishing attacks. Most of the phishing campaigns execute their attack in the first 2 hours and after 12 hours it can enter blacklists. Therefore, these sites cannot get upper location in this list. If a website has a higher location in the list, this increase the probability of being legitimate.

Apart from the others, this is a domain-based feature. This feature is not directly derived from the URL. We need to use a third-party service to calculate the Alexa ranking. Therefore, use of this feature slow down the execution of system.

6. Number of brands: Use of brand names is generally preferred by the phishes. We collected our brand name list from the first 500 firms in the Fortune, some brands from the Alexa ranking system, some banks (international), some social networking and micro blogging sites.

7. (3 Features) Average/Longest/Shortest Word lengths: For confusing the victims mind, Phishers use different length of works in their address. The length of the words in the URL is also an important feature for us. We get three different features as average, shortest and longest words in the URL.

8. Number of keywords: Use of some special keywords can also deceive the computer users. Therefore, we identify some keywords such as “login, secure, account, server” which are mostly preferred in the malicious URLs and then construct a keyword list. This list contains about 176 words and is constructed especially from the URLs in the Phish tank and this list only contains English words.

9. (8 features) Number of special characters and words (‘.’, ‘=’, ‘_’, ‘-’, ‘\\’, ‘@’, ‘com’, ‘cmd’). While investigating the phishing URLs which are get from the Phishtank, it is seen that some special characters and words are mostly preferred. Therefore, we get the number of all of these as different features in the proposed system. For example, if we look at “paypal.com-login.com”, we

can see that “paypal” is only a subdomain and original host name is “com-login.com”. However, use of “paypal”, “.” and “com” together results the user to see the host name as “paypal.com”. A standard computer user is hardly seeing this fact, therefore a software based support is important for us. For example, the comparison of the “number of @ characters” between the legitimate and phishing URLs is depicted in Figure 10.

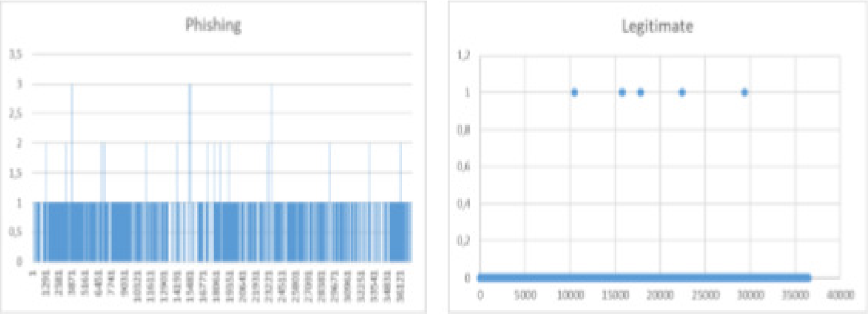


Figure 10. Number of @ Characters in the URLs

Additionally, the use of special characters can also be so deceptive. For example, “mail.google.com” is a legitimate webpage, however, phishers can change it as “mailgoogle.com” with different host name, which is hard to distinguish from the original one.

10. Subdomain number: Legitimate URLs generally have a smaller number of subdomains, however, as explained in the previous example phishers can use the subdomain names as if the domain names. Additionally, they can use several subdomains name the address similar to the original ones. Therefore, a smaller number of subdomains increase the probability of being legitimate web page.

11. Number of Digits: To pass some spamming filters, phishers use some numeric characters in their URLs. Generally, there is no occurrence of numeric characters in the domain name of the legitimate web site.

12. Standard deviation of the words’ length: In the URL (especially in long phishing URLs) there are a number of words. The standard deviation of them is get as a feature in the system.

13. Number of words: The number of words is also an important feature This feature also contains the compound words, which are two or more words that are combined to form a new word with different/similar meaning. To deceive the users, phishers also use compound words in the URL. Therefore, there is need to find each word (and compounds words) in this address. The comparison of phishing dataset and legitimate dataset is shown in Figure 11.

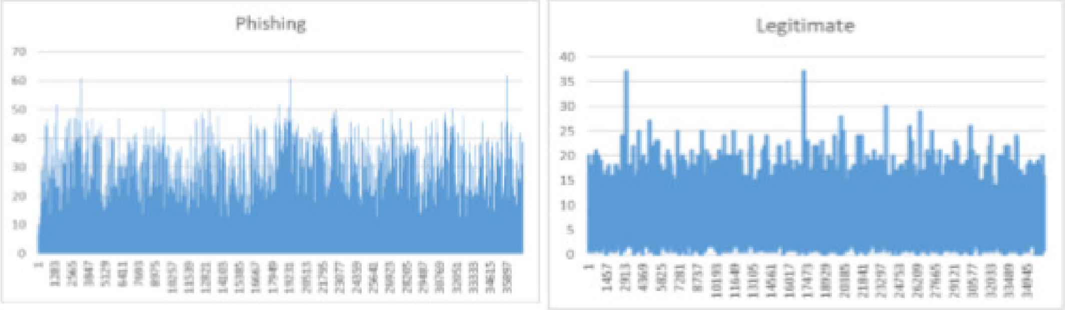


Figure 11. Number of words in the URLs

14. Average length of the compound words: In the previous feature we get the number of compound words. However, the size, especially the average size, of these words is also important to detect the phishing attacks.

15. Character Repetition: To cheat the use phisher can repeat some characters in the domain name. For example, “apple.com” can be repeated by “applle.com” or “applee.com”. This type of names can also be distinguished by the use of similarity index. Usage of some distance measures can ease the calculation of this value.

16. (2 Features) Use of “username” and “userid”: While analysing the phishing URLs from gathered from the Phishtank, it is seen that many of the URLs contains these specific words, “as “username” and “userid”, which are used for deceiving the user. Therefore, these features are defined as binary features and if these words exist then their values are 1, else 0.

3.3. TRAINING THE SYSTEM

The success of the system depends on the learning/training mechanisms used. In the proposed system we used two different learning mechanism: Artificial Neural Network and Deep Neural Network. In Artificial Neural Network(ANN) approach we used a one hidden layer framework, which contains 20 neurons in it. Due to its structure, we trained the system with only 100 epochs and we preferred the use of “adam” optimizer. As an activation function different functions can be selected: RELU, SIGMOID and TANH. Therefore, we tested all of them and found that SIGMOID function gives the best performance among them.

In the Deep Neural Networks design we increased the number hidden layers to two and at every layer, ‘RELU’ activation function is used. Each hidden layer contains 20/40 neurons is used. In the output layer, the activation function is preferred as sigmoid while the optimizer function is preferred as ‘adam’. Training is executed for 100 epochs and we can reach about 91% accuracy rate. To train and test the proposed system we used the Tensorflow, which is an open-source library for data science. It contains some learning algorithms that can be used in different application areas. As an important advantage, system can be run not only on multiple CPUs but also on Graphics Processing Units (GPUs).

3.4. CROSS VALIDATION

Cross-validation is a statistical method to evaluate a stability of the training models by splitting the original dataset into two parts: a training set and a test set. Due to its simplicity and understandability, it is a popular method, which results in a less biased or less optimistic experimental results. To reach a randomness free experimental result we used these set as 10-fold cross validation and divide original data to the ten parts and get one of them as test set while using the other nine as train set.

3.5. CLASSIFICATION

After training the system, we can easily classify any URL in the system. Before executing the classification, firstly related features must be extracted from the URL. After that according to used third party depended features, such as Alexa Ranking, there is a need to connect with this part. After collecting each features classification algorithm is executed. 4

4. EXPERIMENTAL RESULTS

In this study, we compared deep neural network approach with the artificial neural network approach by using the defined features. To train the system we need to use a dataset. Therefore, we prefer the up to date dataset of Buber et. al., which contains 36400 legitimate and 37175 phishing URLs in it.

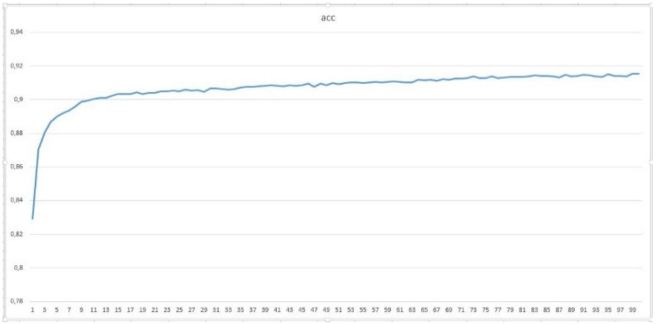


Figure 12. Accuracy Rate of ANN Approach with Sigmoid Activation Function

After, training and testing the data set, best result is reached in Deep Neural Network approach up to 96% accuracy rate with 100 epochs as depicted in Figure 13 with different number of neurons in the hidden layers. If we increase the epoch number, this rate is increasing a little bit more.

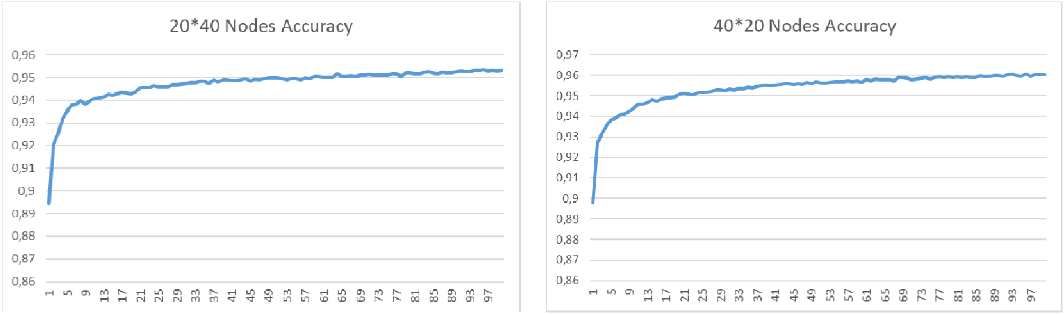


Figure 13. Accuracy Rate of Deep Neural Network with two hidden layers

The execution time of the proposed system is also an important parameter for selection of the phishing detection system. This execution time can be divided into two parts: the feature extraction time and classification time. To measure the feature extraction time, we tried to classify 100 different URLs and measure the all required time needed for calculating the related features in the system, and also total time for all features. The average time in calculated as about 0.6 sec for feature extraction of a URL. We also investigate the Feature based time need and result is depicted in Table 1.

Table 1. Some important features' calculation time

Feature	IP Address	Total Word	Standard Deviation	Number of Brands	Longest Word	Shortest Word	Alexa Rank
Average	**0.1263**	**0.0107**	**0.0036**	**0.0002**	**0.0105**	**0.0105**	**0.4080**
Max	0.7416	0.1371	0.0159	0.0010	0.0338	0.0339	2.1824
Min	0	0	0	0	0	0.00099	0.2813

As can be seen from this table the dominant factor of the feature is the Alexa Ranking part. Due to its need for connecting the third-party services it needs almost 2/3 of all calculation time. Therefore, if it is wanted to decrease the decision time this feature can be disabled. In the table some other time-consuming features are also shown. The other features are calculated less than 10-4 sec, therefore, they not listed.

5. CONCLUSIONS AND FUTURE WORKS

Due to the growing use of Internet in our daily life, cyber attackers aim their victim over this platform. One of the mostly encountered attack is named as "phishing" which creates a spoofed web page to obtain the users sensitive information such as userid and password in financial websites by using social networking facilities. The malicious web page is created as if a legitimate web page, especially copying the original web page one to one. Therefore, detection of these pages is a very trivial problem to overcome due to its semantic structure which takes the advantage of the humans' vulnerabilities.

Software tools can only be used as a support mechanism for detection and prevention this type attacks, and these tools especially use whitelist/blacklist approach to overcome this type of attacks. However, they are static algorithms and cannot identify the new type of attacks in the system. Therefore, as an efficient solution, we propose the use of Artificial Neural Network and Deep Neural Network based system for classifying the incoming URLs. The experimental results show that both these approaches result satisfactory accuracy rate and DNN with 40*20 hidden layer structure produce best solution with about 96% of accuracy.

The latency of the execution time of the algorithm is also an important metric for selection of the detection algorithms. As seen from the results use of Alexa Ranking results a great increase in the execution time, although it has a great importance for detection of phishing. Therefore, according to aim of the system this feature can be disabled for decreasing the execution time.

As the Future works, to decrease the execution time and increase the efficiency of the system, the power of the Graphics Programming Units can be used. Additionally, the other approaches of Deep Learning, such as recurrent neural networks, convolutional neural networks and LSTM can be tested for increasing the performance of the system.

REFERENCES

[1] "USOM," 2018. [Online]. Available: https://www.usom.gov.tr/dosya/1418807122-USOM-SGFF001-Siber%20Guvenlige%20Giris%20ve%20Temel%20Kavramlar.pdf. [Accessed May 2018].

[2] "ITU-," 2008. [Online]. Available: https://www.itu.int/rec/T-REC-X.1205-200804-I. [Accessed May 2018].

[3] Rouse, Margaret, "whatis," November 2016. [Online]. Available: http://whatis.techtarget.com/definition/cybersecurity. [Accessed May 2018].

[4] Diakun, Nadia – Thibault, Purse, Randy & Craigen, Dan, "Defining Cybersecurity," October 2014. [Online]. Available: http://www.timreview.ca/sites/default/files/article_PDF/Craigen_et_al_TIMReview_October2014.pdf. [Accessed May 2018].

[5] "Technopedia," [Online]. Available: https://www.techopedia.com/definition/24747/ cybersecurity. [Accessed May 2018].

[6] Stallings, William, "Introduction," in Network Security Essentials: Applications and Standards, New York, Pearson, 2011, pp. 4-5.

[7] Chia, Terry. "IT Security Community Blog," Stack Exchange, 20 August 2012. [Online]. Available: http://security.blogoverflow.com/2012/08/confidentiality-integrity-availability-thethree-components-of-the-cia-triad/. [Accessed May 2018].

[8] Arslan, Mehmet Emin, "Cyber Security and Cyber Attack Types," Gazi University, Ankara, 2016.

[9] Arachchilage, Nalin Asanka Gamagedara, Psannis, Konstantinos E. & Gupta B. B., "Defending against phishing attacks: taxonomy of methods, current issues and future directions," Springer Science Business Media, New York, 2017.

[10] Podjarny, Guy., "SNYK," SNYK, 10 May 2017. [Online]. Available: https://snyk.io/blog/owasptop-10-breaches/. [Accessed 19 May 2018].

[11] Anti Phishing Working Group, "Phishing Activity Trends Report 1st Half," Anti Phishing Working Group, San Francisco, 2017.

[12] Khonji, Mahmoud, Iraqi, Youssef, Senior Member, IEEE, & Jones, Andrew, "Phishing Detection: A Literature Survey," IEEE COMMUNICATIONS SURVEYS & TUTORIALS, vol. 15, no. 4, pp. 2091-2092, 2013.

[13] Jain, Ankit Kumar & Gupta B. B., "Phishing Detection: Analysis of Visual Similarity," Security and Communication Networks, p. 4, 10 January 2017.

[14] Crowe, Jonathan., "Blog of Barkly," Barkly Protects, July 2017. [Online]. Available: https://blog.barkly.com/phishing-statistics-2017. [Accessed 20 May 2018].

[15] Ivan Galkin, "Crash Introduction to Artificial Neural Networks," Ulcar, [Online]. Available: http://ulcar.uml.edu/~iag/CS/Intro-to-ANN.html. [Accessed 27 May 2018].

[16] Sandhya, L., Thomas Ciza & James, Joby, "Detection of phishing URLs using machine learning techniques," in Control Communication and Computing (ICCC), India, 2013.

[17] Buber, Ebubekir, Diri, Banu & Sahingoz, Ozgur Koray, “NLP based Phishing Attack Detection from URLs”, 17th International Conference on Intelligent Systems Design and Applications (ISDA), Delhi, India,

[18] Khonji, Mahmoud, Iraqi, Youssef., & Jones, Andrew, (2013). Phishing detection: a literature survey. IEEE Communications Surveys & Tutorials, 15(4), 2091-2121.

[19] Nguyen, Luong Anh Tuan, et al. "A novel approach for phishing detection using URL-based heuristic." Computing, Management and Telecommunications (ComManTel), 2014 International Conference on. IEEE, 2014.

[20] Rami M. Mohammad, Fadi Thabtah, Lee McCluskey, (2012) An assessment of features related to phishing websites using an automated technique. In: The 7th international conference for internet technology and secured transactions (ICITST-2012), London

[21] Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, Banu Diri, (2018) "Machine learning based phishing detection from URLs", Expert Systems with Application, 2018, https://doi.org/10.1016/j.eswa.2018.09.029.

[22] Ebubekir Buber, Banu Diri and Ozgur Koray Sahingoz, (2017) "Detecting phishing attacks from URL by using NLP techniques," 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, 2017, pp. 337-342.

AUTHORS

Saide Isilay Baykal is a computer engineer. She graduated from Computer Engineering Department of Istanbul Kultur University in 2018. Her research areas are Deep Learning, Machine Learning, Artificial Intelligence and Cyber Security.

Deniz Bulut is a computer engineer. She graduated from Computer Engineering Department of Istanbul Kultur University in 2018. She graduated from Kadriye Moroglu High School. Her research areas are Deep Learning, Machine Learning, Artificial Intelligence and Cyber Security

Ozgur Koray Sahingoz is currently an associate professor in the Department of Computer Engineering at Istanbul Kultur University. He graduated from the Computer Engineering Department of Bogazici University in 1993. He received his M.Sc. and Ph.D. degree from Computer Engineering Department of Istanbul Technical University, in 1998 and 2006, respectively. His research interests lie in the areas of Artificial Intelligence, Deep Learning, Parallel and Distributed Computing, Soft Computing, Information Systems, Wireless Sensor Networks, Intelligent Agents, Multi Agent Systems.