International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 01 | Jan 2019 www.irjet.net p-ISSN: 2395-0072

Identification of Phishing Website using Deep Learning Algorithm

Navin R T[1], Dr Yuvaraj N[2]

[1]Student, Dept. of Computer Science and Engineering , KPR institute of Engineering and Technology, Tamil Nadu, India
[2]Associate Professor, Dept. of Computer Science and Engineering , KPR institute of Engineering and Technology, Tamil Nadu, India

---

Abstract - *Phishing is a crime, the personal information like username, password, bank account details are obtained by the phisher by acting like a legitimate entity through email. The phishing makes the users to enter their personal information at a fake website, which will be similar to the genuine site. The cost of phishing attacks in 2015 averaged more than $1.5 million per incident, and only 3% of companies are unharmed during the attack, while other companies suffered losses in the millions of dollars. To find whether the website is a phishing site or a genuine site is a challenging task. The existing phishing website detection method uses blacklists/whitelists approach to find the phishing site. The existing detection methods are not able to make accurate prediction to find whether the site is a phishing site or not. The pro-posed system uses the machine learning algorithm and deep learning algorithm to train the system and to find the phishing site. These two algorithms are used to classify the URL based upon the training dataset and predict whether the site is a phishing site or a genuine site. The deep learning algorithm is used for increasing the accuracy of the prediction. In the study made between the machine learning algorithm and deep learning algorithm, we have found that deep learning algorithm gives high accuracy.*

***Key Words*: Deep learning; Machine learning; Phishing website; Random forest; URL.**

1. INTRODUCTION

Internet technology has grown so extensively over the last few decades from online social networking to online e - commerce and banking technologies to make people's lives more comfortable. This uncontrollable growth has resulted in many security threats to network systems: the most frequently encountered is “phishing”. Phishing is a web - based attack in which attackers attempt to reveal sensitive information such as user id / passwords or account information by sending an email from a reputable person or entity. Phishing attacks can occur in many different forms of communication such as SMS, VOIP and e - mail. Every internet users have many accounts in social networks, banks and lot more. These users are considered as a target for the phishing attack. Still most of the web users are unaware of the phishing attack. Phishing attack typically takes advantage of social engineering to attract the victim by sending a spoofed link to a fake web page. The spoofed link are sent to the victim through e-mail or sms. Once the user opens the link then a fake webpage will be opened similar to the genuine webpage, so when the user enters the personal information the information will be sent to the attacker. The cost of spear- phishing attacks in 2015 was an average of more than $ 1.5 million per incident, only about 3 percent of companies suffered losses in tens of millions of dollars during an attack (or in one case in 2015, $100,000,000, breaking the $61,000,000 record of 2014).

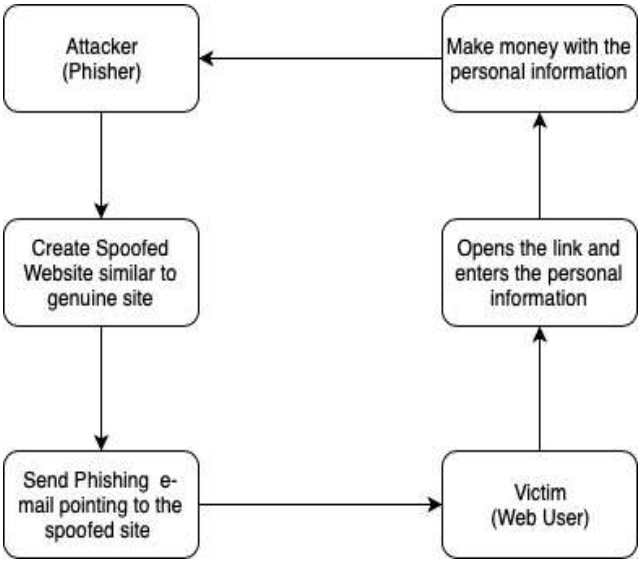

Fig. 1: Flow diagram of phishing attack using e-mail

Keep in mind that we refer to how much money the thieves got away with in these latter cases. The data breach spread over two years to approximately 70 million customers and cost a total of $ 162 million. According to the Ponemon Institute, successful phishing attacks (against large companies with > 10,000 employees) now amount to $ 3.7 million per attack in a report from the first quarter of 2016. And this was despite the use of security solutions, some of which were actually intended to prevent such fraud [1]. To overcome the phishing attack we need to create awareness among the web user about the phishing attacks and need to develop an efficient detection method to find whether a URL is a phishing link or not. Without visiting the website we need to find whether the link is a genuine link to ensure security to the web users.

2. Types of phishing attacks

2.1 Deceptive phishing

Deceptive phishing [2] is the most common kind of phishing. An attacker is trying to obtain confidential information from the victims in this case..

2.2 Spear phishing

Spear phishing targets individuals rather than a large group of people. Attackers frequently investigate their victims on social media and other sites.

2.3 Whaling

When an attacker follows a "big fish" like a CEO, then it is called as whaling. Often, these attackers spend considerable time profiling the target to find the right time and means to steal login credentials.

3.4 Pharming

Pharming sends users to a fraudulent website, which seems legitimate, similar to phishing. In this case, however, victims don't even have to click on a malicious link to the wrong site.

3. Common phishing scams

3.1 Credit card phishing scams

It is easy to monitor [3] your credit card accounts online in our digital age. Many people are so busy and time- pressed that they assume that every email they receive from their credit card company is legitimate.

3.2 Bank phishing scams

Banking customers are popular targets for phishing attackers. If you have a bank account, you will most likely have online access from time to time. Every user have a username and password linked to your online account.

3.3 Email phishing scams

A spoofed email message is often the cornerstone of a phishing scam that is well done. Since the earliest days of phishing, fraudulent emails have been used to catch unaware Internet users.

3.4 Website phishing scams

It's never a good idea to trust a website blindly. If you assume a site is legitimate, you may fall prey to phishing attacks. If this happens, you can disclose sensitive information inadvertently to people who can use it to identify theft and other malicious things.

4. RELATED WORKS

4.1 Blacklist based phishing detection

The proposed system [4] uses an improved blacklist method that uses key distinguishing features extracted from the website's source code to detect phishing sites. Each phishing website has a unique fingerprint that is generated by the set of features proposed. The Simhash algorithm is used for each website to generate fingerprints and calculated the fingerprints. The accuracy obtained from the experiment is 84.36%.

4.2 Heuristic based phishing detection

A heuristic phishing detection [5] technique using uniform resource locator (URL) features was developed to find the phishing URLs. The features of the phishing URLs are identified and used these phishing detection features. A data set of 3,000 phishing site URLs and 3,000 legitimate site URLs was evaluated.

A new approach [9] for the detection of phishing sites using the URL features. Various components in the URL are identified and calculated a metric for each component. Page ranking is then combined with the metrics achieved to determine if the websites are phishing websites. The phishing detection technique has been evaluated with 9,661 phishing websites and 1,000 legitimate websites in the dataset. The accuracy obtained from the system is 97 %.

Identifies fraudulent websites [6] by submitting incorrect credentials and analyzing the response. The server responses were also analyzed in order to determine the legitimacy of a particular website. All pages with Alexa 500 and Phishtank login forms were analyzed. The accuracy obtained from the system is 96%.

4.3 Content based phishing detection

Based upon the experiment [7] shows that CANTINA is good at detecting phishing sites and that approximately 95% of phishing sites are marked correctly.

A framework of file matching algorithms is implemented [8] to detect phishing websites based on their content using a custom data set of 17,684 phishing attacks aimed at 159 different brands. The results of experiments on various algorithms show that some phishing detection approaches can achieve a detection rate of more than 90%.

4.4 Machine learning based phishing detection

The phishing websites can be identified using a combined approach by constructing resource description framework (RDF) models and using ensemble learning algorithms for the classifications of websites. This approach uses

supervised learning techniques to train the system. This approach has a promising true positive rate of 98.8% which is definitely appreciable. As this method have used random forest classifier that can handle missing values in dataset it can able to reduce the false positive rate of the system to an extent of 1.5%.

A detection system [11] with a wide scope of protection using URLs only, which depends on the fact that users deal directly with URLs for surfing the Internet and offers a good approach to detect malicious URLs. The simulation results from the system give accuracy of phishing URLs of 93%.

4.5 Deep learning based phishing detection

Google PageRank, Google Position, Alexa rank and other URL-based features were considered in the sys-tem [12], and its accuracy and performance are improved by using neural networks where optimum weight is calculated using the firefly algorithm. The experimental results show that the proposed technique works more efficiently in terms of accuracy, true positive rate, true negative rate, false positive rate and false negative rate than the existing technique. The proposed technique has been shown to be 99.52% accurate.

Oh, Nguyen et autres [13], proposed a dynamic approach to detect phishing sites through the use of the artificial neural single - layer network. In this paper, the first step of the technique calculates the value of six heuristic. The Neural network is trained with a data set of 11,660 sites and 2 test data sets are available to verify accuracy. The best result is that this heuristic technique detects 98.43% of fake sites. This technique won't result better for a large dataset.

Zhang et.al [14], proposed multi - layer perception of phishing email neural networks and calculated the effectiveness and efficiency of this proposed approach. He compared many classification algorithms such as NN, SVM & decision tree, Naive Bayes, but Neural Network gave 95% accuracy to the highest recall value and shows that the neural network detects phishing emails best.

An efficient approach [15] to detect phishing websites based on a single- layer neural network. In particular, the proposed technique objectively calculates the value of heuristics. The heuristic weights are then generated by a single- layer neural network. A data set of 11,660 phishing sites and 10,000 legitimate sites is evaluated for the proposed technique. The results show that more than 98% of phishing sites can be detected by the technique.

Table -1: Accuracy

Detection methods	Approximate accuracy
Blacklist based	84%
Heuristic based	96%
Content based	92%
Machine learning based	98%
Deep learning based	99%

From the study of related works it is clearly proved that machine learning based random forest algorithm and deep learning algorithm gives higher accuracy in the detection of phishing website URL.

5. Proposed model

This section describes the proposed detection model for phishing attacks. The phishing attack occurs when the victim clicks the URL that is sent to the victim through e-mail and the victim is directed to a fake website similar to the genuine site. In this detection method we only focus on URL to differentiate between the legitimated and fake websites. The URL is processed using the random forest machine learning algorithm and deep learning algorithm. The URL are classified upon several criteria like IP address, URL having “@“ symbol, “//“ symbol in between the URL, long URL address, URL with prefix or suffix. These criteria are checked and classified using the random forest algorithm and deep learning algorithm in order to detect the URLs of phishing websites from the legitimate websites.

5.1 IP address

If an IP address is used in the URL as an alternative to the domain name, such as " http:/130.210.6.195/phish.html, " users can confirm that someone is trying to steal their sensitive personal data. In this case the machine learning algorithm identifies the IP address and conforms that the URL is a phishing link.

RULE: IF domain parts of URL == IP address THEN phishing site ELSE legitimate site.

5.2 Long URL

Long URLs are used by the phisher to hide the suspicious data inside the URL. According to the authors in [4] reported that average length of the legitimate URL is 40 and the average length of the phishing URL is greater than 75.

RULE: IF URL length > 75 THEN phishing site ELSE legitimate site.

5.3 Prefix or suffix

The phisher adds the prefixes or suffixes to the do-main name of the URL separated by (-) symbol to make the user to trust that the given URL is a link legitimate site.

RULE: IF symbol (-) present in domain name THEN phishing site ELSE legitimate site.

5.4 Additional address

The phisher adds additional address in-front of the real address this addition of address is usually done by adding "//" before the real URL. In this the position of the "//" is checked based upon the position the phishing URL are identified. For HTTP the position of the symbol "//" is six and fro HTTPS the position of the symbol "//" is seven.

RULE: IF symbol "//" position in the URL > 7 THEN phishing site ELSE legitimate site.

5.5 "@" symbol

The phisher uses the "@" symbol in the URL because the browser ignores everything before the "@" symbol. The real address are often placed after the "@" symbol. The "@" symbol in the URL indicates that the URL is a phishing URL.

RULE: IF symbol "@" present in the URL THEN phishing site ELSE legitimate site.

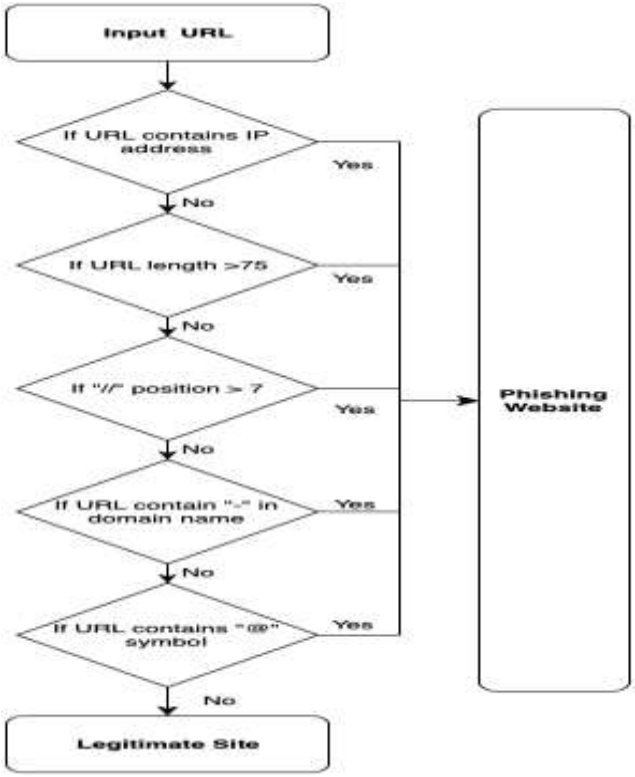

Fig. 2: Processing of input URL

5.6 Random Forest algorithm

Random Forest [15] is a supervised machine learning algorithm that can be used to perform both regression and classification task in data mining. It is an ensemble based technique that can be used to perform classification. It makes use of a number of classification trees (like decision trees) and then gives the final result. This algorithm works by creating a number of classification trees randomly. These trees are created by making use of different samples from the same dataset and also they may use different types of features each time to create the trees. Thus, all the trees are created randomly by making use of different sub sets of the same dataset, and also the features are taken randomly for the creation of any tree. By doing so, Random Forest ensures that it does not over fit the data, as in the case of the decision trees. Once the trees have been formed, we can do the classification by finding the results of each tree and then assigning it to the class that has been determined by the most number of trees.

5.7 Deep learning algorithm

Deep Learning [16] is a machine learning subfield that involves algorithms inspired by the structure and function of the brain called artificial neural networks. The field of artificial intelligence [17] is mainly when machines can perform tasks that typically require intelligence from people. It involves learning machines where machines can learn from experience and acquire skills without involvement of people. Deep learning is a subset of machine learning in which artificial neural networks, human brain- inspired algorithms, learn from large amounts of information. Similar to how we learn from experience, the profound learning algorithm repeatedly performs a task, changing it a little each time to improve the result. We are talking about deep learning because the neural net-works have different (deep) levels that allow learning. A deep learning problem can learn to solve just about any problem that requires " thought" to figure out. Deep learning enables machines to solve complex problems even if they use a varied, unstructured and interconnected data set. The deeper they learn the algorithms, the better they perform.

5.7 Training and Prediction

The detection system is trained with the dataset obtained from UCI machine repository [18], the random forest algorithm and deep learning algorithm are used to train the system with the given input dataset and the trained system is tested with a new dataset if the system correctly predicts the phishing website then the detection system can stop the training or else the detection system is further trained, only after a good evaluation result the training is stopped and the detection system are ready to detect the phishing sites. The detection system learns about the phishing website or the URL by using machine

learning and deep learning algorithm. After completion of the training process the detection system is ready to detect the phishing website by identifying the URL. The detection system gets the input from the user and compares the given input with the trained detection system and predicts the given input URL whether it is a phishing site or a legitimate site.

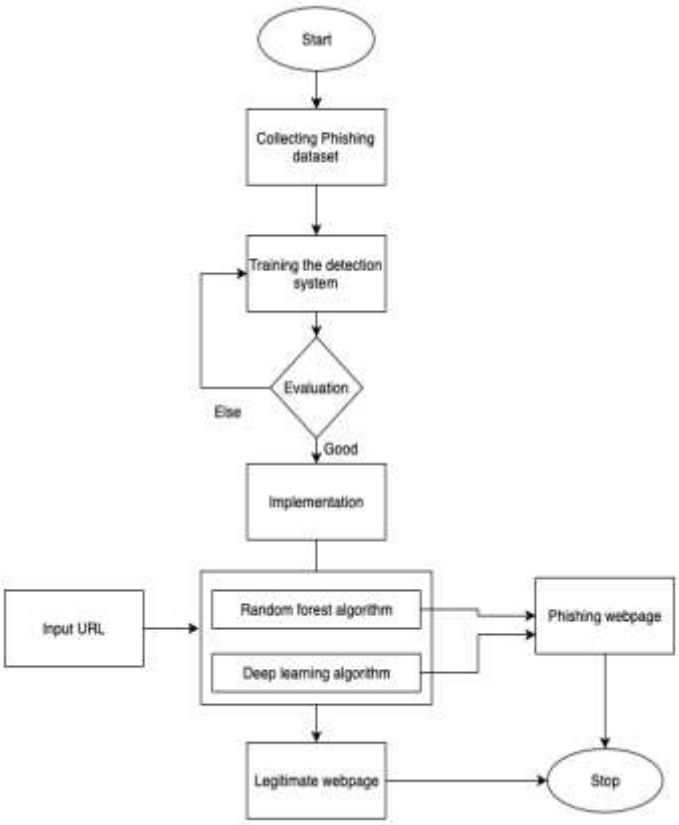

Fig. 3: Overall architecture

6. Conclusion and future work

Due to the lack of awareness about the phishing attacks makes the attack successful. The users should not blindly open the link received through e-mail and enter their personal information. The phishing attack are not easily identified. The web spoofing attacks occur even after inventing new prevention methods. The main reason for this study is to educate users and help them to identify the phishing website from the legitimate site by using the URL.

The most important way to protect the user from phishing attacks are by educating the user about the possible ways of phishing attacks. The user need to check URL before blindly opening the URL. There are some limitations in this detection method. The deep learning algorithm gives higher accuracy than the random forest algorithm. The deep learning algorithms But the detection system doesn't give a 100% accuracy still the system lacks in finding the phishing website with high accuracy.

The future work of this project to is obtain higher accuracy and an application is created for the mobile phone for the mobile users to find the phishing website through mobile phone. As most of the people uses their mobile phone more than the laptop or a computer so a mobile application for the detection of phishing site will protect the users.

REFERENCES

1. Infosec Institute https://resources.infosecinstitute.com/category/enterprise/phishing/phishing-as-a-risk-damages-from-phishing/financial-losses/#gref

2. Cisco https://www.cisco.com/c/en/us/products/security/email-security/what-is-phishing.html

3. Phishing.org http://www.phishing.org/common-phishing-scams

4. Routhu Srinivasa Rao, Alwyn Roshan Pais "An Enhanced Blacklist Method to Detect Phishing Web-sites", 2017.

5. Jin-Lee Lee, Dong-Hyun Kim, Chang-Hoon, Lee "Heuristic-based Approach for Phishing Site Detec-tion Using URL Features" Proc. of the Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology - CEET 2015.

6. Ahmed Nafies Okasha Mohamed "A New Heuristic Based Phishing Detection Ap- proach Utilizing Selenium Web-driver" Master's Thesis (30 ECTS).

7. Anjali Gupta1, Juili Joshi2, Khyati Thakker3, Chitra bhole4 "CONTENT BASED APPROACH FOR DETECTION OF PHISHING SITES" Apr-2015.

8. Brad Wardman, Tommy Stallings, Gary Warner, Anthony Skjellum "High-Performance Content-Based Phishing Attack Detection".

9. Luong Anh Tuan Nguyen, Ba Lam To, Huu Khuong Nguyen "Detecting Phishing Web sites: A Heuristic URL-Based Approach " in the 2013 International Conference on Advanced Technologies for Communications (ATC'13).

10. Akansha Priya, Er. Meenakshi, "Detection of Phishing Websites Using C4.5 Data Mining Algorithm", 2017 IEEE International Conference On Recent Trends in Electronics Information & Communication Technology (RTEICT).

11. Ammar Yahya Daeef , R. Badlishah Ahmad , Yasmin Yacob , Ng Yen Phing "Wide Scope and Fast Websites Phishing Detection Using URLs

Lexical Features ", 2016 3rd International Conference on Electronic Design (ICED), August 11-12, 2016, Phu-ket, Thailand.

12. Swetha Babu K.P1, Dr.Radha Damodaram2 "Phishing Detection in Websites Using Neural Networks and Firefly" in 2016.

13. . L. A. T. Nguyen, B. L. To, H. K. Nguyen and M. H. Nguyen "An efficient approach for phishing detection using single-layer neural network."In Advanced Technologies for Communications (ATC), pp.435-440, 2014.

14. N. Zhang and Y. Yuan, "Phishing detection using neural network,"CS229 lecture notes.

15. S. Jagadeesan, AnchitChaturvedi, Shashank Kumar (2018) "URL Phishing Analysis using Random Forest".

16. Machine Learning Mastery https://machinelearningmastery.com/what-is-deep-learning/

17. Forbes https://www.forbes.com/sites/bernardmarr/2018/10/01/what-is-deep-learning-ai-a-simple-guide-with-8-practical-examples/#3d028d1a8d4b

18. UCI Machine Learning Repository https://archive.ics.uci.edu/ml