A survey on phishing website detection in Machine learning and Neural networks

A S S V Lakshmi Pooja[1], M.Sridhar[2]

A S S V Lakshmi Pooja CSE Department & GRIET

M. Sridhar CSE Department & GRIET

ABSTRACT:

Web Phishing appeals to the user to connect with the fake site. The main goal of this attack is to rob the user of sensitive information. The intruder builds websites similar to those that look like the original website. It allows attackers to access Sensitive data such as username, password, details of debit cards etc.

This paper aims to review many of the phishing detection strategies recently suggested for the website. This will also provide a high-level description of various forms of phishing detection techniques.

Keywords: Phishing website, URL, Machine learning, Neural networks

I.Introduction

In present days, As there are such huge numbers of individuals are monitoring utilizing web to perform different exercises like web based shopping, online bill installment ,online mobile recharge, banking transactions .Due to wide utilization of this client face different security challenges like cybercrime .There are numerous cybercrime that are generally performed for instance spam , fraud ,cyber terrorisms and phishing. Among this phishing is new cybercrime and well known in these days. Phishing is fraud attempt, which performed to acquire delicate data of client. The site which is designed by phisher it looks like a same as any genuine site and spoof client for getting private data of client, for example, username, secret word, banking details for various reasons.

Once the innocent user hits the URL unknowingly it is cause for the attacks. For example, phishers sends an email to the user it appears to be from a trusted sender with a URL. Upon clicking the URL, the innocent user landed on the insecure Website.

As stated by Anti Phishing Working Group (APWG) total number of phishing sites detected in 4th quarter of 2019 is 162,155. It is down from the 266,387 seen in quarter 3 and the 182,465 in quarter 2 and up from the 138,328 in 4th quarter of 2018 and 4th quarter of 2018 is slightly decreased when compared to 3rd quarter of 2018 was 151,014.

There were increases in SAAS/webmail targeted sector with 33% of overall phishing attack and other sectors like payments with 21%, Financial Institution with 19%, cloud storage/File Hosting with 4%, eCommerce/retail with 4%.

As per the security threat reports we can understand that there is a continuous increase in web attacks in the forms

spams, phishing and malware distribution, which shows more impact on business, educational institutes, banking sector etc.,

One response to this solution is to employ a blacklist of malicious URLs produced by anti-virus organizations.

The difficulty with this strategy is that the blacklist can't be comprehensive, as new malicious URLs tend to pop up all the time. Therefore, approaches that can automatically identify a new, previously unknown URL as either a phishing site or a legitimate one are required. These solutions are usually machine-learning approaches in which a program can categorize new phishing sites through a model built using established attack training sets and other approaches for detecting web based attacks are:

- Static analysis
- Dynamic analysis
- Heuristic-based

Types of Phishing Attacks:

1. Traditional Phishing:

It also called as deceptive phishing or cloned phishing. It is one of the utmost common type of phishing.

The cyberpunk profess to be other person to obtain either personal data or login credentials. The two most commonly used modalities are:

(A) Victim receives a email from fraudster. The fraudster impersonate to be a representative of a reputable organization in this email, and makes an attempt to steal the personal details of victim.

(B) The victim receives an email in which the hacker connects to a harmful website. Either the URL is nearly similar to the real one but has a loophole that the hacker takes advantage by adding an iframe and waits for the user to access that information and enter it.

2. Malware-Based Phishing:

The crook launch some virus into the email, or a link indicating to harmful site, When the victim accesses it automatically installs a piece of virus to his computer. This form of attack is mainly common for small and medium-sized businesses because the software they use is not always updated to the latest version.

3. Spear phishing:

This form of phishing attack is generally much more customized than in previous cases. Within such emails, hackers normally provide some personal details like
Person name, his/her position in the business or his/her mobile number.
The reason for this is to obtain his/her trust and therefore gain the information
they need to penetrate the corporate network and access the sensitive data they are after.

4. Smishing (SMS):

In this type, the attacker concentrates on both the ways i.e email and through mobile phone. The hacker pretends to be working on behalf of a well known organization and sends an message to the victim, either announcing he has won a reward or offering him to take part in a lottery or contest. To redeem the award, the victim must:

- Victim has to click on a link
- Call a number
- Victim should send back a text message with some data

5. Vishing:

The word "vishing" is a combination of voice and phishing. In this attack it involves the usage of telephone. Phisher tries to do fake calls to the victim , seems to be a bank employee, manufacturer, operator with the purpose of collecting personal details like bank accounts etc.,

6. Pharming:

Cybercriminals exploit a company's hosts files or its domain name system (DNS) during a pharming attack. Consequently, a false address is returned when a URL is requested, and the user is escorts to a fake website. Hackers therefore get victims to enter information on a fake website which they monitor.

II. Related Work:

Various approaches are developed for identifying a harmful URL. Here are few approaches which have been used by researches for identifying a harmful or phishing URL. They are:

- Based on Machine learning
- Based on Neural Networking

(A)Based on Machine Learning:

This approach works more effectively in larger datasets. In addition to that this removes the drawbacks of present approach and proficient in identifying zero day attack. It has proved that nearly 99% of data is accurate using machine classifiers. Output relays on the size of training data, feature set, and classifier type. Limitation of this approach is it fail to identify when attackers use compromised domain to host their website.

In this area a lot of research work has been carried out. Some work has used for various classifiers to increase the accuracy of phishing website detection. KNN, SVM,

Fig.1: processing cycle of phishing attack.

Decision tree, ANN, Naïve Bayes, PART, ELM and Random Forest are different classifiers are used. Among all these tree based DT and RF classifiers , it is possible to increased the data set as per my survey.

By utilizing ML techniques to identify phishing URLs some features or set of features extract from it. There are two general types of features that can be extracted from URLs, namely:

- Host based features
- Lexical features

Host based features describe the characteristics of a website such as location of the website, by whom it operates and when the site was designed. In addition to that textual properties of URL are described by lexical features because URL consists of text strings which can be split into subparts like protocol, hostname and path and system may describe the persuasiveness of a site based on any combination of those components.

In inclusion to URL- based applications, various types of applications are used in the detecting process of academic studies in machine learning algorithms. Features gathered

from the studies for identifying phishing domains with machine learning techniques are listed. They are:

1. Based on URL features
2. Based on Domain features
3. Based on Page features
4. Based on Content features

1. Based on URL features:

URL is a primary thing a website analyzes to verify whether it is a phishing or not. There are certain points to distinguish the phishing domain URL. Features related to certain points are acquire while processing the URL functions which are mentioned below are based on URL.

- Digit count within the address
- Total length of address
- Checking whether the URL is cybersquatted or not(www.adidas.com → www.Addidas.com)
- Evaluating whether it includes a well-known brand name or not (apple-icloud-login.com)

2. Based on Domain features:

It has the function detecting the names of phishing domains. Unassertive questions relating to the domain name, which we want to categorize as phishing or not, furnish useful information to us. Listed below are some useful domain based features.

- Number of days completed after getting registered.
- Its name or its information processing address in blacklists of well-known name services?
- Is the registrant name hidden?

3. Based on page features:

Page based features utilizes the details about pages that are rating services measured for trustworthiness. Below are some features which provides information about how secure a website is. They are :

- Global pagerank
- Country pagerank
- Position at the Alexa prime one million website.

Numerous Page-based features furnish us with details on target site behavior. Some of the features are listed below .It is not easy to acquire certain types of features. There are Some paid services to get certain kinds of features.

- Evaluates Range of Visits for the website on a daily, weekly or monthly basis.
- It calculates Number of Page views per Visit
- It calculates average Visit duration of website.
- Domain Category.
- Web traffic share per country.
- Similar Websites etc.

4.Based on Content Features:

By acquiring these type of features needs active scan to target domain. Page Contents are examined so that we can identify whether or not target domain is used for phishing.

Some data regarding about pages being processed is listed below.

- Headings of the page
- Meta Tags
- Unseen Text
- pictures etc.

By analyzing these information we can gather information such as:

- Is it necessary to login to website

- Category of website
- Data regarding audience profile etc.,

All features which are described above are used to detect a phishing site. In all cases , using these features may not be useful, and there are some disadvantages to using such functions. For example, it may not reasonable to use any of the features such as based on content features to create a mechanism of rapid detection that can evaluate the number of domains between 100,000 and 200,000. If we want to inspect new registered domains, another example would be that Page-based functionality is not very useful. Thus the feature detection mechanism can use depend on the purpose of the detection mechanism. What features Should be carefully selected for use in the detection mechanism.

Detection Process:

Detecting Phishing Domains is a classification problem, so in the training process we need labeled data which has samples as phish domains and legitimate domains. A very important point to construct effective detection mechanism is the dataset which will be used in the training process. We need to use samples whose particular groups are known. So it means that the samples classified as phishing must be identified completely as phishing. Similarly the samples identified as valid must be detected as completely legitimate. Otherwise the system does not work properly if we use samples of which we are not sure about.

In machine learning there are variety of algorithms for identifying whether it is phishing or legitimate one. So, Among all as mentioned above Decision tree is a powerful algorithm for classifying the training dataset.

Decision tree:

Decision tree includes arrows and nodes, and is initialized from the root node. If-then there are rules for any network node. Arrow is the representation of which node was the next. Also, Tree will contain various stage classifiers and internal nodes.

Different terminologies used in the decision tree are:

- **Base hub:** base hub is called the base hub from which tree is built.
- **If-then principle:** each hub inside the structure includes If-then instructions, classes and components.
- **Arrow:** by using a bolt, the next point is eluded.
- **Leaf hub:** tree closes with a leaf hub or eliminator.
- In order to generate a tree specific calculations, ID3 model is implemented to decide entropy data in order to evaluate the objective value. In C4.5 calculation tree, sub trees will be generated in which each tree hub has a parent hub and also a kid hub. In addition, the tree ends with the ending center, which contributes to the problem's objective yield.

Naïve Bayes:

Naïve Bayes (NB) is a simple but effective classifier which is used in many applications. For an NB classifier, x is the vectors of features, y Î{0,1} is a symbol representing either a phishing or a legitimate website (y = 1 for phishing and y = 0 for legitimate), and P (x) is the conditional probability of the vector given its symbol. Assuming phishing and legitimate websites are equally possible, the later likelihood of x belonging to y=1 will be as follows:

$$P(y = 1|x) = \frac{P(x|y=1)}{P(x|y = 1)+P(x|y=0)} \qquad (1)$$

Support Vector Machine:

Support Vector machines are categorized into two categories i.e., linear and non linear classifiers. It works by finding a hyperplane which seperates the training data into two classes. In other words SVM discovers the ideal hyperplane separation between two labels. This can be

expressed via the kernel function K(x, xb), which calculates the similarity of two feature vectors, and the non-negative coefficients μ3. SVM shows which examples of training lie close to the boundary of the decision. It classifies data into decision boundaries by measuring distance.

$$h(x) = \sum_1^n \propto_i (2y_i - 1)\, K(x_i, x) \qquad (2)$$

Random Forest:

A Random Forests is built with random attribute selection using bagging. In order to enhance efficiency, Random Forests employs a divide and conquer method (ensemble mechanism) The system blends multiple random subsets of trees in a random forest. The cumulative result is determined based on the sum of the individual outcomes, or weighted average. The accuracy depends on a measure of dependency between the classifier and the strength of the individual classifiers, and they enhance the problem of decision trees overfitting.

(B) Neural Network Based Approaches:

The technique of identifying fake domains based on the deep learning model is to develop a realistic deep learning model, Design the input needed by the model, and pill out the features to complete the detection of the phishing website URL via the deep learning model. For this form of approach, the selection and design of the input model would have a direct influence on whether the model is successful. Currently CNN and RNN are the widely used models to detect phishing websites.

1.Convolutional Neural Network:

Convolutional Neural Network (CNN) is a category of deep neural networks utilized for image processing research. Compared to other algorithms for the image classification, CNN needs relatively little preprocessing. This learns features itself – a significant benefit in software development, which differ from other classifications pre-specified by conventional phishing detection researchers. Since it is mainly intended for image classification, it is performed for phishing detection on embedded character-levels. CNN networks contain a layer of convolution, pooling layer and a fully connected network with non-linear activation feature.

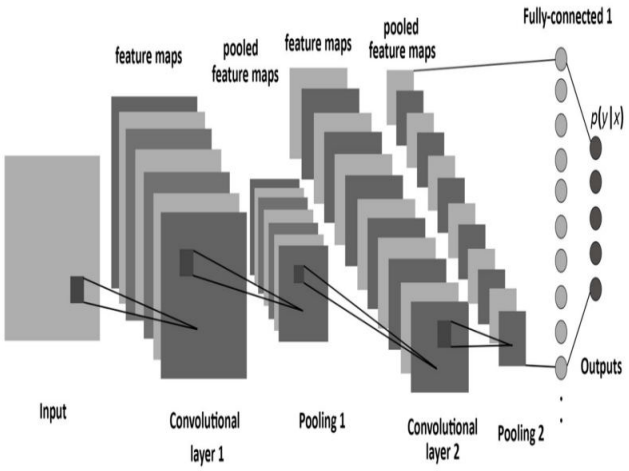


Fig 2: Structure of CNN

2. **Recurrent Neural Network (RNN):**

Recurrent Neural network(RNN) is also one of the category of Neural network. In RNN **the input of the current step depends on the output of previous step**. But in neural networks, all the inputs and outputs are not dependent to each other, some cases like when there is a need to estimate the next coming word of a sentence, the previous words are required and hence there is a need to recollect the previous words. Thus RNN came into survival, which solved this issue with the help of a Hidden Layer. The main and most significant characteristic of RNN is **Hidden state**, which remembers some information about a sequence.

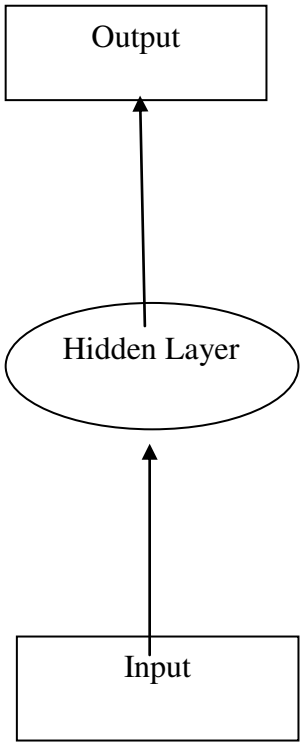


Fig 3: Structure of RNN

RNN have a **"memory"** which recollects all information about what has been calculated. It utilizes the same values for each input as it carry out the same task on all the inputs or hidden layers to generate the output. This reduces the complexity of parameters, unlike other neural networks.

3. **Long Short Term Memory(LSTM):**

LSTM was designed by Hochreiter & Schmidhuber. They are special kind of RNN. It is mainly introduced to handle the situation RNN fails. RNN depends on the output of previous step for input of current step. So, it should remember all the inputs of the previous step. By introducing LSTM we can overcome this long-term dependencies.

LSTM consists of three gates. They are:

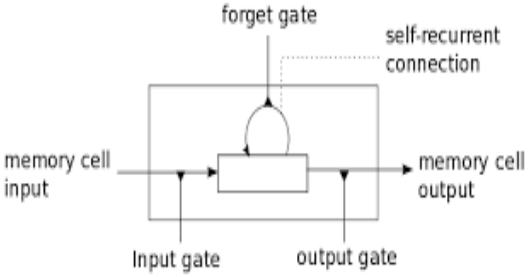
(i) Forget gate

(ii) Input gate

(iii) Output gate

- Forget gate: Forget gate decides what needs to be remember and it throws away unnecessary data.
- Input gate: Cell state is updated.
- Cell state gets multiplied by the forget vector which results in dropping a value in cell state and a pointwise addition is done which results in new cell state.
- Output gate: It decides what the next hidden state should be.

LSTM has capability of memorize the data for long period of time. It is used for predicting the data.



Search Strategy:

It includes an overview of strategy that involved in searching for the details of the topic Phishing Website detection with different techniques.

Initially we prefer searching in IEEE Xplore to get papers published on phishing website detection in conferences and journals. Further, we proceed to read paper's title, abstract and keywords to find its relevance with our proposed work. Moreover, we read the content to find out related information about any phishing website detection system and its respective research work.

III. Results and Discussions:

The different algorthims which are to detect a phishing website are reviewed individually. From the review of different survey papers.

References	Algorithm	Dataset	Accuracy
[8]	Convolution Neural Network(CNN) LongShortTerm Memory (LSTM) Recurrent Neural Nework(RNN)	Alexa and Phishtank website	97%
[10]	Neural network Deep belief network Backpropagation Technique	Not mentioned	Neural network-89% Deepbelief network-94% Random forest-91%
[11]	Neural network	UCI dataset Phishtank Website	96%
[12]	Decision Tree(DT),Random Forest(RF), Gradient Boosting(GBM), Generalized Linear model(GLM), Generalized Additive Model(GAM)		Random Forest Highest Accuracy – 98.4%
[13]	Natural language Processing	Nazario Phishing Email set	Proposed SEA Hound provides 95% Accuracy

Commonly Used Algorithms:

No	Algorithms	References
1	Random Forest	[10] [14]
2	J48	[10]
3	CNN	[8] [15] [17]
4	LSTM	[8] [14][15][17]
5	Naïve Bayes	[16] [17]
6	SVM	[16] [17]

IV.Conclusion:

Phishing attack is the one of the most problematic cyber crime facing by a internet users.

The usage of internet has been increased in the same way on the other hand threats from attackers also increased.

In this paper, a survey of different type of phishing attacks which are faced by users and this survey brings a knowledge on different machine learning and deep learning techniques, there are many techniques which are used to reduce phishing attacks from attackers.

V.References:

[1] Zinal Shukla, Kirtirajsinh Zala , Riddhi Kotak , "A Survey of Website Phishing Detection Techniques" International Journal on Future Revolution in Computer Science &CommunicationEngineering Volume: 4 Issue: 1, IJFRCSCE | January 2018, Available @ http://www.ijfrcsce.org

[2] Purvi Pujara, M. B.Chaudhari, "Phishing Website Detection using Machine Learning : A Review" International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2018 IJSRCSEIT | Volume 3 | Issue 7 | ISSN : 2456-3307

[3]https://docs.apwg.org/reports/apwg_trends_report_q3_2019.pdf

[4] Dharmaraj R. Patil1, Jayantrao B. Patil

"Feature-based Malicious URL and Attack Type Detection Using Multi-class Classification" ISeCure The ISC Int'l Journal of Information Security July 2018, Volume 10, Number 2 (pp. 141–162) http://www.isecure-journal.org

[5]https://randed.com/types ofphishing/?lang=en

[6] https://towardsdatascience.com/phishing-domain-detection-with-ml-5be9c99293e5

[7] Eint Sandi Aung ,Chaw Thet Zan, Hayato YAMANA "A Survey of URL-based Phishing Detection" DEIM Forum 2019 G2-3

[8] Weiping Wang,1 Feng Zhang, Xi Luo, Shigeng Zhang

"PDRCNN: Precise Phishing Detection with Recurrent Convolutional Neural Networks" Volume 2019 |Article ID 2595794 |https://doi.org/10.1155/2019/2595794

[9]https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/

[10] Maneesh Kumar Verma, Shankar Yadav, Bhoopesh Kumar Goyal, Bakshi Rohit Prasad and Sonali Agarawal "Phishing Website Detection Using Neural Network and

Deep Belief Network" https://doi.org/10.1007/978-981-10-8639-7_30

[11] Erzhou Zhu ; Yuyang Chen ,Chengcheng Ye ,Xuejun Li ,Feng Liu "OFS-NN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network" IEEE Access (Volume: 7) 04 June 2019

[12] J.Shad and S.Sharma, "A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology," pp. 425–430, 2018.

[13] T.Peng, I.Harris, and Y.Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301, 2018.

[14] Alejandro Correa Bahnsen, Eduardo Contreras Bohorquez, Sergio Villegas, Javier Vargas and Fabio A. Gonz´alez "Classifying Phishing URLs Using Recurrent Neural Networks" Easy Solutions Research ,MindLab Research Group, Universidad Nacional de Colombia, Bogot´a

[15] Thuy Thi Thanh Pham, Van Nam Hoang, Thanh Ngoc Ha "Exploring Efficiency of Character-level Convolution Neuron Network and Long Short Term Memory on Malicious URL Detection" Request permissions from Permissions@acm.org. ICNCC 2018, DOI: http://dx.doi.org/ 10.1145/3301326.3301336

[16] Joby James, Sandhya L., Ciza Thomas "Detection of Phishing URLs Using Machine Learning Techniques" 2013 International Conference on Control Communication and Computing (ICCC)

[17] Arun Kulkarni, Leonard L. Brown, III, "Phishing Websites Detection using Machine Learning" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 7, 2019