DEEP LEARNING APPROACH FOR PHISHING ATTACKS

Aliya CH[1], Dr.D.Loganathan [2]

[1]PG Scholar,Dept of Computer Science & Engineering,SVS College of Engineering Coimbatore,Tamil nadu, India
[2]Professor,Department of Computer Science & Engineering,SVS College of Engineering Comibatore, Tamil nadu, India

---***---

Abstract - Phishing websites are one of the biggest threats to the entire internet users of the world. Those kind of websites tries to steal sensitive data such as personal identity information, banking details credit card details and login credentials etc. So, we can implement an intelligent system to detect phishing websites. Basic system is based on machine learning method, particularly supervised learning, in which the Random forest technique is used to classify which can shows good performance and accuracy rate of 98.8%. For improving performance and other metrics we are going to implement a new method to the current system. That is deep learning method, particularly on supervised learning. In this system instead of Random forest method we are using CNN CSTM method for the detection technique. The benefit of LSTM is the data timing and recording long-term dependencies. LSTM has a strong capacity for learning, powerful potential face of large, high-dimensional complex data. Experimental results suggest that the 99.1 percent precision of this pattern is that of other algorithms in the neural network.

Key Words: CNN LSTM, Deep Learning, Phishing URL Detection, Machine Learning, Malicious URL

1.INTRODUCTION

This Physical attacks have been prevalent in the Internet world over the past several decades, but we knew how to solve the issues caused by it. Syntactic attacks came into play later, taking advantage of the software vulnerabilities to compromise the systems. Today, semantic URL attacks are commonly used by attackers to obtain a user's sensitive information. .It is still the most frequent attack in cyberspace. This system is used by attackers mainly because people always fail to corroborate the veracity of the information they get through the internet and make it a simple prey to the hands of online violations. Usually, such attacks are carried out through spam emails and phishing websites and trick the user into providing the user's credentials and other sensitive data. The most unsafe criminal activity in cyberspace is phishing. As most users go online to access the services provided by government and financial institutions, phishing attacks have increased significantly in recent years. Phishers have begun to earn cash, and they do this as a successful business. It can occur in two ways, either by receiving suspicious emails that cause the fraudulent site or by users accessing links that go on to a phishing website.

Various methods are phishers to attack the vulnerable users messaging, VOIP, spoofed link and counterfeit websites. The reason for creating phishing websites is for users to obtain private information such as account numbers, login ID, debit and credit card passwords, etc. In addition, attackers ask security questions to respond to posing as a user-providing high-level safety measure. They are easily trapped in phishing attacks when users respond to those questions. Many researches phishing attacks by different communities world. Email is considered to be the number one vehicle to deliver all kinds of malicious attacks.

To align it with the signature of a heuristic pattern, the heuristic solution uses the signature databases of any known attacks. Novel attacks are not identified by the trade-off in using heuristics, so it is possible to circumvent the signatures by obfuscation. Updating the signature database is also slow given the emergence of new attacks, especially zero-day attacks. Content analysis content-based approach, using well known algorithms term frequency/inverse document frequency, to detect phishing websites (TF-IDF). It analyses a page's text-based content itself to determine whether or not the website is phishing. Additionally, measuring website traffic using Alexa is another method that has been implemented by researchers to detect phishing websites. Machine learning takes advantage of its predictive power. It learns the characteristics of the phishing website then predicts new phishing characteristics. There are several techniques, such as nave Bayes (NB), decision tree (DT), support vector machines (SVM), RF, artificial neural network (ANN), and Bayesian net (BN). The accuracy of phishing detection varies from one algorithm.

We offer anti-phishing industry a solution that can detect more sophisticated phishing attacks as well as detecting simple phishing attacks and also to study and analyse various security issues in Phishing Attacks. On Internet with an attempt to suggest a model of security implementation, which will cover the advantages of the available process used to handle the security issues in Internet Computing and will be able to provide more reliable aspect for implementation of Security in Internet. The insights gained from the research would form a set of guidelines for designing less vulnerable Internet Computing Security, in the form of a structured framework for risk evaluation. Phishing detection techniques do suffer low detection accuracy and high warning especially when novel phishing approaches are introduced. Phishing attacks are often prevented by detecting the websites and creating awareness to users to spot the phishing websites. One of the powerful techniques for detecting phishing websites is machine learning algorithms. Phishing detection tools play a vital role in ensuring a safe online experience for users because of the importance of protecting online users from becoming victims of online fraud, disclosing tips to an attacker, among other effective uses of phishing as an attacker's tool. Deep learning is a branch of supervised machine learning which learn from the data by itself and design a model for future use. It has greater possibility to detect newly generated phishing URLs and also require manual feature engineering. This is done by combination of CNN with long short term memory (LSTM) to obtain the accuracy in classifying the phishing URLs.

2. PHISHING DETECTION USING DEEP LEARNING

An Phishing detection method uses deep learning approach which focus on convolutional neural network (CNN) and combination of CNN with long short term memory(LSTM) to obtain the accuracy in classifying the phishing URLs.

2.1 PHISHING DETECTION

The deep learning model is the method of detecting phishing websites, building the input required by the model and extracting the features through the deep learning model to complete the detection of the phishing website URL Phishing websites using LSTM Recurring neural networks[1][2][10] benefit from data timing capture and long-term dependencies. LSTM has a strong, strong potential face of complex high-dimensional massive data. The URL dataset is collected and used for data training and testing in this strategy. As a result of the accuracy obtained for each and every one, the phishing and non-phishing sites will result from training and testing.

2.2 SYSTEM ARCHITECTURE

The URL dataset samples are obtained from sources like PhishTank and OpenPhish .Nearly 4000 URL samples are taken. Out of which the data set is divided into it is divided into 2 sections 70% for training and 30% for testing the data. After that it is moved to CNN network for training the data.. Analysis phase is taken in the next step through LSTM.

Classifier is used for classification of the system and certain models are obtained as per the accuracy of the each data and in the prediction phase we can predict the phishing and on phishing site as per the high and low accuracy respectively. LSTMs have over conventional feed-forward neural networks and RNN.

This is because of their property of selectively remembering patterns for long duration of time. The graph presentation will show a clear idea of accuracy and loss of CNNLSTM.

2.3 PROCESS FLOW

The detection process in the system will take as per the following steps:-

Step 1: Data set is split into 2 sections

Step 2: 70% train data set and 0% test data set

Step 3: Training the data set and models are created

Step 4: Graph presentation is obtained

Step 5: Test data set is done manually

Step 6: Accuracy of tests is obtained

Step 7: Prediction of phishing website or not

A user interface is also shown for clear clarification of the process. In UI when we type the URL to check and if the URL is authorized one it will directly move to the website and if it is a Phishing site an dialog box with phishing will appear on the screen.

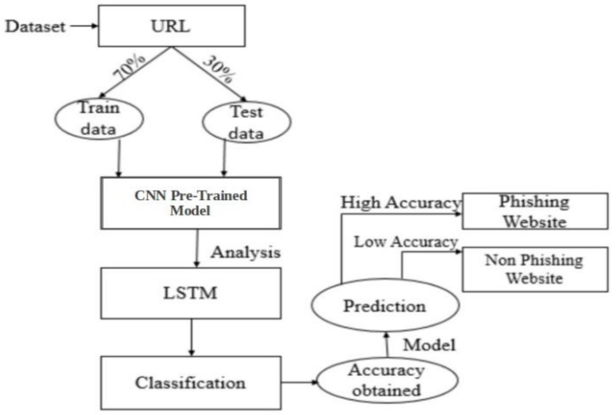


Fig-1: System architecture CNN-LSTM

3. EXTRACTION AND ANALYSIS OF URL FEATURES

Confusing the Uniform Resource Locator is very common for phishing; an important part of phishing is to attract users to click on the URL to visit their phishing website. The standard resource address on the Uniform Resource Locator, entrance to increase the likelihood of users visiting phishing sites, phishing attackers often use deceptive URLs that are visually similar to fake ones. It is a specific type of Uniform Resource Identifier (URI) that is used to locate a web resource n a computer network and to retrieve it.

Two-part URL: protocol. The protocol will be specified via the domain name or Internet Protocol (IP) address as http, https and resource location. A colon will separate the protocol and location and two forward slashes will follow. Web search results by entering the URL obtained through emails, Users visit a website and other modes of web page connections. If the URL used is compromised and an attack is imposed on the user, the user will be redirected to web pages containing malicious scripts. These compromised URLs are referred to as malicious URLs. Thus, finding a URL through some mechanism resolves the attacks mentioned as below:

Protocol://hostname[:port]/path/[;parameters][query]#.

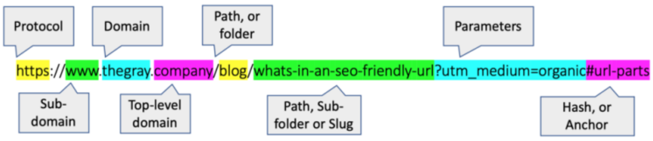
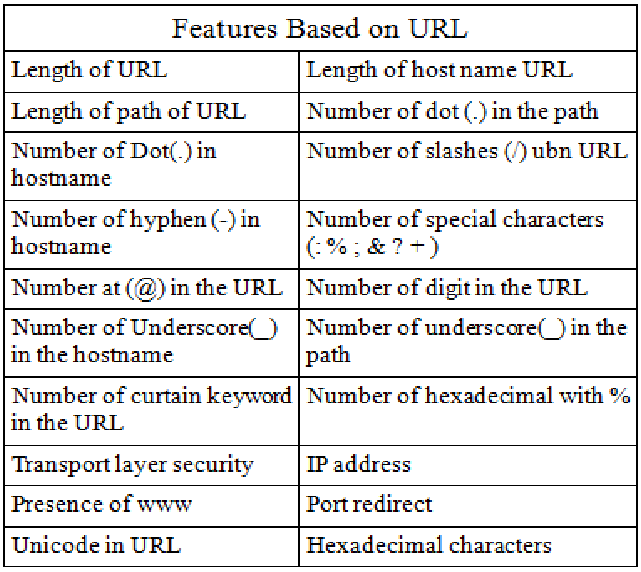


Fig-2: Uniform Resource Locator structure

The popular way to confuse URLs is to create a phishing URL to confuse the user by partially modifying and removing the host name portion and the route portion based on the destination URL. For example, the attackers using "www.flip0kart.com" as a fake Amazon website (the real URL is "www.flipkart.com"),or using the "www.interface-transport.com/ www.paytm.com/" as a fake Paytm website (the real URL is "www.patym.com") and so on.

4. CHARACTERISTICS OF UNIFORM RESOURCE LOCATOR

The attacker's phishing URL is designed to persuade the user that this is a legitimate website. The user's personal and leaked financial details can therefore be accessed by cyber criminals. In order to achieve this goal, attackers use some common methods to disguise phishing ties. Some of the attributes are:



Features Based on URL	
Length of URL	Length of host name URL
Length of path of URL	Number of dot (.) in the path
Number of Dot(.) in hostname	Number of slashes (/) ubn URL
Number of hyphen (-) in hostname	Number of special characters (: % ; & ? +)
Number at (@) in the URL	Number of digit in the URL
Number of Underscore(_) in the hostname	Number of underscore(_) in the path
Number of curtain keyword in the URL	Number of hexadecimal with %
Transport layer security	IP address
Presence of www	Port redirect
Unicode in URL	Hexadecimal characters

Fig-3: Feature extracted from URL

URLs Length: Phishing URLs appear to be longer than URLs that are real. By covering the questionable portion of the URL, which can redirect user-submitted information or redirect uploaded web pages to suspicious domain names, long URLs increase the risk of misleading users.

Prefixes and suffixes in URLs: Phishers fool users by re-modeling URLs that appear like real URLs.

Length ratio: Calculate the ratio between the URL's length and the path's length. A higher proportion of legitimate URLs are often available on phishing sites.

The “@” and “-” counts: The '@' and '-' numbers in the URL. The '@' sign in the URL allows the browser to ignore previous entries and then redirects users to the typed links.

Punctuation counts: The number of “! # $% &” in the URL. Phishing URLs normally have more punctuation.

Other TLDs: The number of TLDs that the URL route shows. By using domain names and TLDs on the path, phishing web links imitate valid URLs.

IP address: Host name: The URL component uses an IP address rather than a domain name. .

Port Number: Verify that the port is included in a list of established HTTP ports, such as 21, 70, 80, 443, 1080 and 8080, if a port number appears in the URL. If there is no port number in the list, mark it as a potential URL for phishing.

URL Entropy: The larger the URL entropy is, the more dynamic it is. Since phishing URLs appear to have random text, their entropy helps us to try to find them.

5. CONCLUSIONS

To detect phishing URLs, the phishing detection analysis was performed using CNN-LSTM models. Deep learning techniques such as CNN and CNN LSTM are preferred to machine learning techniques, since they have the potential to achieve optimal feature representation themselves by taking the raw URLs as their input. Using the LSTM method, high precision is obtained, solving the problem that it is difficult for other machine learning techniques to extract valid features from the data.

The prediction approach has been shown to be successful in practice and can overcome the issues that conventional approaches are difficult to solve. At the same time, the methodology of LSTM deep learning, in conjunction with the characteristics of the RNN optimize the model's method of training. The deep learning model training time is generally possible from hours to days, and the optimization convergence time is strict on the timeliness of power dispatching and other problems. This is of great importance.

REFERENCES

[1] Alejandro Correa Bahnseny, Eduardo Contreras Bohorquez, Sergio Villegasy, JavierVargasy and Fabio A. Gonz´alez, “Classifying Phishing URLs Using Recurrent Neural Networks” 978-1-5386-2701-3/2017 IEEE.

[2] Amani Alswailem, Norah Alrumayh, Bashayr Alabdullah, Dr.Aram Alsedrani , “Detecting Phishing Websites Using Machine Learning”, 978-1-7281-0108-8/19/$31.00 2019 IEEE.

[3] Bo Wei , Rebeen Ali Hamad , Longzhi Yang , Xuan He , Hao Wang , Bin Gao and WaiLok Woo , “A Deep-Learning-Driven Light-Weight Phishing Detection Sensor” Sensors 2019, 19, 4258; doi:10.3390/s19194258 www.mdpi.com/journal/sensors

[4] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735– 1780(1997)

[5] Huaping Yuan, Zhenguo Yang, Xu Chen, Yukun Li, Wenyin Liu ,“URL Modeling with Character Embeddings for Fast and Accurate Phishing Website Detection” 2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications.

[6] Phishing Activity Trends Report: Phishing Activity Trends Report 1st Half. Methodology (2017)

[7] PHISHTANK: Free http://www.phishtank.com/ community site for anti-phishing service.

[8] Ping Yi ,Yuxiang Guan, Futai Zou, Yao Yao, WeiWang, and Ting Zhu, “Web Phishing Detection Using a Deep Learning Framework” Hindawi Wireless Communications and Mobile Computing Volume 2018, Article ID 4678746, 9 pages

[9] Sinha, S., Bailey, M., Jahanian, F.: Shades of grey: on the effectiveness of reputation based “blacklists”. In: International Conference on Malicious and Unwanted Software, pp. 57–64. IEEE (2008)

[10] Wenwu Chen, Wei Zhang, and Yang Su, “Phishing Detection Research Based on LSTM Recurrent Neural Network”, Springer Nature Singapore Pte Ltd. 2018