International Journal of Computer Science & Communication (ISSN: 0973-7391)
Special Issue, Page30-34, December 2020 ------ available at www.csjournals.com

Detection of Phishing Websites Using deep Learning Context

Poornashree [1] Sharmila S P [2] Chaithanya Prakash K P [3] Prajna N M [4]

[1,3,4]Department of Information Science and Engineering, Siddaganga Institute of Technology, Tumakuru
[2]Assistant Professor, Department of Information Science and Engineering, Siddaganga Institute of Technology, Tumakuru

Abstract: Though, Web service being one of the greatest inventions of mankind so far, and there is also thoughtful manifestation of computer impact on human beings. The swift development of the Internet and the growing popularity of electronic payment in web service, main concern of the public arises in the field of internet fraud and web security. Web phishing is a fraudulent attempt to gain personal information from victims such as username, password, bank information, credit card information and so on. This paper emphases on applying a deep learning framework to detect the websites that are carrying out phishing attacks. Deep Belief Network(DBN) is used for feature extraction and Support Vector Machine is used for classification of URL. The group of features we consider are address bar related, JavaScript based, abnormal based and domain-based features. For each URL we identify phishing rules mining approach to categorize whether the URL is suspicious or phishy prone or legitimate one.

I. INTRODUCTION

Computers and internet play a very important role in our daily life. Most of the people communicate with each other either through a computer or a digital handheld or wearable device connected over the internet [3]. Web service is one of the crucial communication software facilities for the Internet. Web service encompasses World Wide Web infrastructure to provide the methods for a connectivity between one electronic device to other electronic devices [8]. The number of people using e-banking, online shopping and other online services has been increasing due to the availability of convenience and comfort. An attacker takes this as an opportunity to gain money and steals sensitive information needed to access the online service websites. There are numerous security threats to web services on the internet, Web phishing is one of them. Web phishing is a technique to accomplish internet fraud, which uses social engineering techniques through emails, links and short messages to induce users to visit fake websites and to extract sensitive information like their credit card information, private account details, token for payment, personal contacts and so on [5]. A fake web page is created similar to the legitimate web page.

Phishing attack can happen through a virus, website or email. The first phishing attack on America Online (AOL) can be traced back to early 1995. The phisher successfully obtained AOL user's personal information. The phishing activity in early 2016 was the highest ever chronicled since it began monitoring in the year 2004. The over-all number of phishing attacks in 2016 was 12,20,523. According to the Third Microsoft Computing Safer Index Report which was released in February 2014, mentioned that the annual worldwide effect of phishing could be as high as $5 billion [8]. With the pervasiveness of internetwork, phishing has become one of the most serious security coercions in the modern society, thus specifying mechanisms for detecting and providing protection against web phishing is an urgent concern and an essential research task.

II. PHISHING ATTACK AND ITS TYPES:

An adversary creates a fake website and sends enormous amount of mails to users to link to the fake website. When a user clicks on the link or email then user is asked to enter some sensitive information, thus the adversary captures the sensitive data from the user. There are various types of phishing attacks. Phishing attacks are basically observed in four categories [10][11][12].

Table 1: Types of Phishing attacks

Attack	Description
Malware-Based Phishing	execution of terrific software on the user's PC, through which Malwares are encroached along with an attachment in the email, the downloadable files can trace and track the inputs specified through keyboard.
Deceptive Phishing	text sent to a user showing concerns about the need of verification of account details, system failure.
Data Shoplifting	Insecured PCs may consist of susceptible information being stored or warehoused on protected servers.
Content-Injection Phishing	Manipulation of the contents of a genuine sites with untruthful content in order to misdirect the user into forsaking their confidential information to the hacker
Phishing through Search Engines	Many unwanted additions of products and services are introduced into the search

	engines offering products or services at a discounted or economical rate.
Spear phishing	that influences friendship information from social networks, generated over 70 percent achievement rate in experiments.
Whaling	directed specifically at elderly or senior executives and other high-profile targets
Cat Phishing	online trickery that involves getting to acquainted with someone closely in order to gain access to their information
Clone Phishing	The attachment or link within the email is substituted with a malevolent version and then directed from an email address hoaxed to pretends to come from the original sender
Voice phishing	calls that claimed to be from a bank telling that users need to dial a phone number regarding complications or any issues with their bank accounts.
SMS phishing or Smishing	uses cell phone text messages to deliver the attraction or temptation to induce people to reveal their personal information. These attacks typically offer the user to call a phone number, or click a link, or contact an email address specified by the attacker via text message.

Some possible solutions to prevent phishing were introduced which includes, detection of phishing websites that generally includes categories such as detection based on the inclination as white list and black list, detection based on the features and topographies of Uniform Resource Locator (URL), detection based on web content and detection based on machine learning [3].

III. DETECTION OF PHISHING ATTACKS

Web phishing makes an attempt to acquire subtle and sensitive information from novices, often for malicious purposes and reasons, by pretending to be a trustworthy website on the Internet. Investigators present some explanations and solutions to detect this kind of web phishing. To decide whether a specific website is a phishing website or not, the direct way is to use a black list or white list technique proposed in [5]. Search the URL in some database may also be done and decide accordingly. There are two traditional methods to detect phishing websites by the blacklist technique. The first possible way is including five empirical approaches to enumerate simple combinations of already recognized phishing sites to discover new phishing URLs [2]. The second technique consists of an approximation matching algorithm that dissects a given URL into several components that are matched individually against the entries available in the blacklist. But these methods are not for the real time applications and may charge you a large slice of our time and effort.

An alternative way for detection of phishing websites is to investigate on the features of URL. Sometimes phishing URL also looks alike to the famous site URL or it may contain few distinct and special characters in the URL. The features of the URL are extracted and applied in the machine-learning-based classification algorithm to detect phishing URLs from the tangible data set.[1] This method is efficient and economical because it utilizes the previous or historical knowledge of the URL, which has a recklessly fast detection speed and a cheaper cost.

A feasible method for phishing detection is content based detection which refers to the detection of phishing sites through the pages of elements, such as field names, form information, and resource reference. The accuracy rate of this method is also found to be high [8]. But at the same time the downside of this method is there is a need to collect bulk amounts of data as a erstwhile knowledge. The problem of phishing detection can also take influences of the popularity of machine learning. machine learning algorithms can be applied to provide solution to phishing detection. This method integrates domain name features, URL text features and web content features into an amalgamated detection basis. In general, the essence of this method is to record all the features of the phishing website into the identical space and then to apply the machine learning and data mining algorithms to detect or notice the phishing activity.

IV. LITERATURE SURVEY

In [1] S. Marchal et al. used the concept of intra-URL relatedness. They evaluated the model using features extracted from the words that constitute a URL. These words are generated based on interrogation or query data drawn from Yahoo and Google search engines. These features are then casted in machine-learning based classification in order to detect phishing URLs from a real and tangible data set.

In [2] Routhu Srinivasa Rao et al. proposed a novel heuristic-based solution to detect phishing attacks and technologically advanced it as a practical application and called it as PhishShield.

In [3] Mahmood Moghimi et al. proposed an ap- proach to detect phishing URLs using the feature sets of the URL. They proposed feature sets which determines the relationship between the the URL of the page and the content of the page.

In [4] W.Niu et al. proposed a model that focuses on phishing email detection. Machine learning models, principally Support Vector Machine (SVM), have been verified and evidenced to be effective for detecting phishing websites. In order to improvise the classification accuracy, this research paper states a model, called Cuckoo Search SVM(CS-SVM). According to W. Niu CS-SVM extracts 23 features, which are used to paradigm the hybrid classifier. In

this hybrid classifier, Cuckoo Search (CS) is combined with SVM to optimize and improvise parameter selection of Radial Basis Function (RBF). Experiments are executed on a dataset consisting of 20,071 non-phishing emails and 1,384 phishing emails.

In [5] Ajay Pandey et al. proposed a model that worked on diagnosing and classifying the phishing attack based on scrutinizing phishing Blacklist, analyzing websites features, and querying WHOIS database. The domain names and features of URLs and are checked using several criteria such as IP Address, lengthy URL address, URLs having distinct symbol "@", redirection made using the symbol "//".

In [6] Baykara et al. proposed a model that focused on phishing email detection. Machine learning based detection methods, mainly Support Vector Machine (SVM), have been demonstrated to be operationally effective. To improve the classification accuracy, this paper also make use of the model Cuckoo Search SVM(CS-SVM). This model extracts 23 features to construct the hybrid classifier. In the hybrid classifier, Cuckoo Search (CS) is integrated with SVM to optimize the parameter selection process of Radial Basis Function(RBF).

In [7] HemaPriya Natarajan et al. proposed a model for detection and categorization of malicious URLs using DBN for feature selection with deep neural network for binary classification.

In [8] Ping Yi et al. proposed a model that focuses on detecting the phishing website using a deep learning framework. Deep Belief Network is used for feature extraction. These extracted features are used for training the Support Vector Machine, to perform classification of URLs.

In [9] P. Neelima et al. proposed a model that uses SVM classifier to find the phishing website. The extracted features are used to train the SVM model.

To quickly summarize the survey there is a lot of requirement to protect user against phishing using Anti-phishing techniques, by learning mechanisms to distinguish Phishing Emails, especially phishing detection system has become essential for e-banking in these days.

V. PROPOSED SYSTEM

We propose Feature Extraction process for phishing detection using Deep belief network (DBN). DBN is a generative graphical model, composed of visible layer and multiple layers of hidden units. There are connections between the layers, not between units within each layer. It is a stack of Restricted Boltzmann Machine (RBM). Input data is represented by Visible units and features that captures the correlation between the data is represented by the hidden units.

The architecture and flow of the proposed system is as depicted in the Figure 2 and Figure 3 respectively. The three major modules of the proposed model are DBN Model, SVM Model and the Classifier. The data set consisting of Phishing and Legitimate URLs is collected from the *kaggle* website. DBN is used to extract features from the URL. It is trained by using layer by layer approach that is one layer is trained at a time. We outline two kinds of features to detect web phishing [6].

- Original Feature: The features such as special characters, number of dots, age of the domain is considered as original features.
- Interactive Feature: The features such as in- degree of URL, out-degree of URL, frequency of access, cookie, browser type is considered as Interactive features.

Table 2: Phishing Features

Sl. No.	Phishing Features
1	presence of special characters
2	count of dots>4
3	age of the domain<3 months
4	low indegree of URL
5	low outdegree of URL
6	frequency of access of URL<=1
7	doesn't leave cookies to the user
8	unknown browser type

We intent to choose Contrastive Divergence (CD) as training algorithm. It calculates the gradient over k times of Gibbs Sampling. Input is given to the visible layer. CD algorithm is applied to the first RBM. Training data is used to train the first layer. After deriving the individual activation probabilities for the first hidden layer, the hidden units of the first hidden layer will be updated in parallel. This is called as Positive Phase. Visible units are reconstructed from the hidden units using Negative Phase that uses the same technique like positive phase. Then associated weights are updated. Weights are updated by multiplying the difference of positive and negative phase values with the learning rate and adding the initial value of weight.

$$UpdW_{11} = W_{11} + L * (P(H_{11} = 1|V) - P(V_1 = 1|H_1))$$ where,

L is Learning Rate

$P(H_{11} = 1|)$ is Positive Phase value

$P(V_1 = 1|H_1)$ is Negative Phase Value

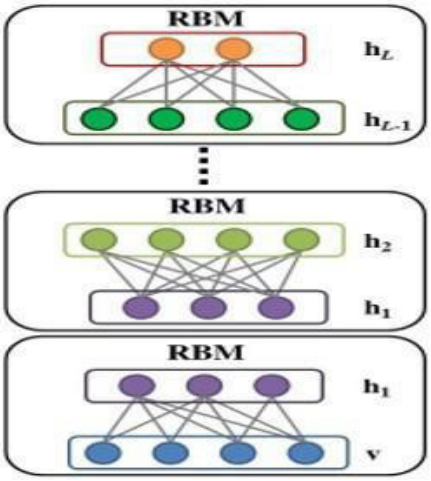

Figure 1: DBN Model

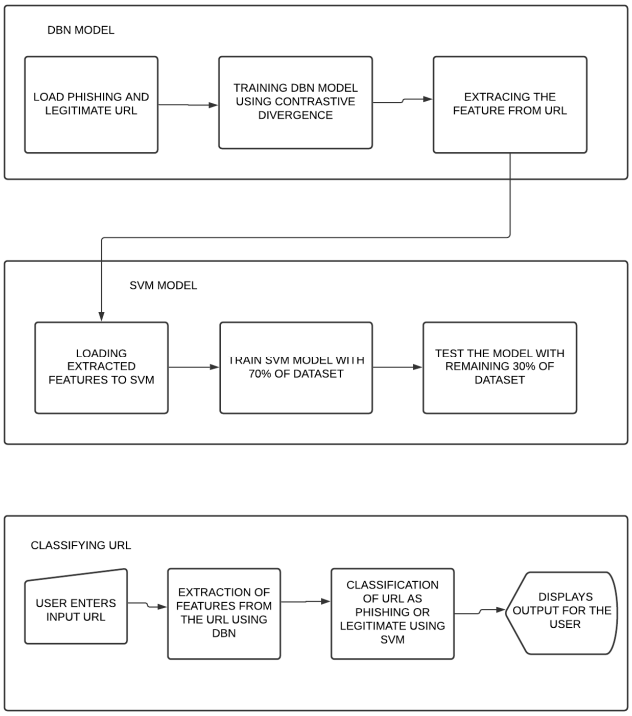

Figure 2: Architecture of the proposed system

The output of this RBM is used as the input for the following RBM. Weight of the second RBM is the transpose of the weight of the first RBM. Again, the Contrastive Divergence algorithm is applied to the next RBM just like it is applied to the first RBM. Positive and Negative phase values are calculated and

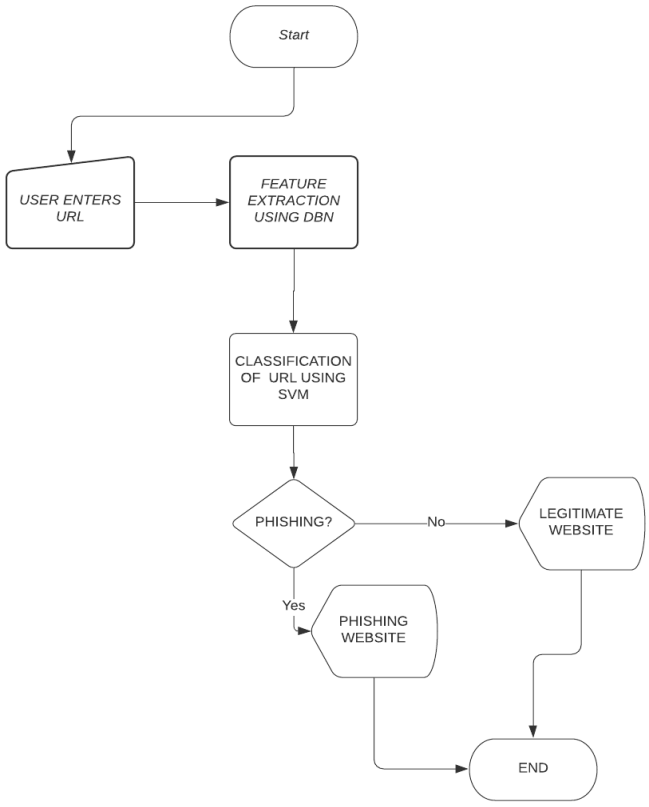

Figure 3: Flow Diagram of the proposed system

associated weights are updated. One more RBM layer is added and the process repeats. This process is repeated till we get a threshold values or an output layer.

A major concern when installing and deploying RBM is satisfactorily tuning their parameters since their performance is very sensitive and limited to the parameter selection process.

After training a model using CD we have to fine tune the parameters to get an optimal value. Activation function, number of hidden layers, learning rate, number of iterations can be considered for fine tuning.

We propose application of SVM for Classification. Support Vector Machine (SVM) is a one of the supervised machine learning algorithms that can categorize cases by finding or filtering a separator. The features extracted from DBN are given as input to the SVM classifier. The data set is split into two subsets a testing set (30%) and a training set (70%). a SVM is trained using a training set and it is tested using a testing set. SVM classifies the URL with high accuracy. Accuracy of the model is calculated by using true positive (TP), true negative (TN),false negative (FN) and false positive (FP) [8] with the following relation.

$$Accuracy = (TP+TN)/(TP+TN+FN+FP)$$

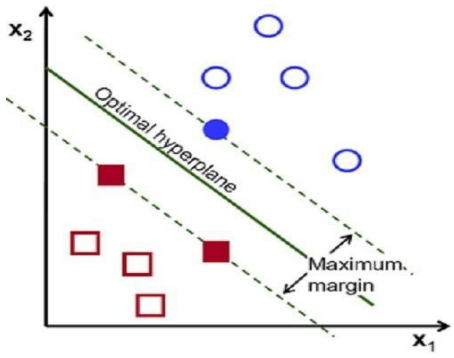

Figure 4: Classification of data using SVM

TP: a phishing website and classified as phishing website.

TN: a non-phishing website and classified as non-phishing website

FP: a non-phishing website and classified as phishing website

FN: a phishing website and classified as non-phishing website

VI. CONCLUSIONS

In this paper we aim to propose a method for detecting phishing URLs and to provide security for the public. We have presented a survey on phishing detection approaches based on URL features. This paper presents two types of features for web phishing detection: Original feature and Interactive feature. In detecting phishing URLs, there are two steps. The first step is to extract features from the URL. We propose the application of DBN as a feature extractor and to classify URLs using the model that has been developed with the help of training data sets. We also prefer the custom usage of SVM as a classifier. Detection of phishing websites with still better accuracy is still a tranquiling open challenge for further research and development. In future we can use different classifiers for classifying the URL. We can compare the accuracy of different classifiers and choose the best one among them.

REFERENCES

[1] S. Marchal, J. Francois, R. State, T. Engel, “Phish storm: Detecting phishing with streaming analytics,” IEEE Transactions on Network and Service Management, vol. 11, no. 4, pp. 458–471, 2014.

[2] Routhu Srinivasa Rao, Syed Taqi Ali,"PhishShield: A Desktop Application to Detect Phishing Web pages through Heuristic Approach",Eleventh International Multi-Conference on Information Processing- 2015 (IMCIP-2015)

[3] Mahmood Moghimi, Ali Yazdian Varjani," New rule-based phishing detection method",Expert Sys- tems With Applications 53 (2016) 231–242.

[4] W.Niu, X.Zhang, G.Yang, Z.Ma, Z. Zhuo, "Phish- ing Emails Detection Using CS-SVM," 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUAA), Guangzhou, 2017, pp.1054-1059.

[5] Ajay Pandey, Hemali Sampat, Hezal Lopes, Manisha Saharkar, “Detection of Phishing Website Using Machine Learning”, International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 03, Mar-2018.

[6] Baykara, Z. Z. Gurel, "Detection of phishing attacks," 2018 6th International Symposium on Dig- ital Forensic and Security (ISDFS), Antalya, 2018, pp. 1-5.

[7] ShymalaGowri Selvaganapathy, Mathappan Nivaashini , Hema Priya Natarajan, "Deep belief network based detection and categorization of ma- licious URLs", Information Security Journal A Global Perspective, ISSN: 1939-3555, April 2018.

[8] Ping Yi , Yuxiang Guan, Futai Zou, Yao Yao, Wei Wang, Ting Zhu, "Research Article Web Phish- ing Detection Using a Deep Learning Framework", Hindawi Wireless Communications and Mobile Computing Volume 2018, Article ID 4678746, Pub- lished 26 September 2018.

[9] P.Neelima, P.Srilatha, V.Priya Darshini, "De- tecting Phishing Website with Machine Learning", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8 Issue-3, September 2019.

[10] Gupta, Surbhi, Abhishek Singhal, and Akanksha Kapoor. "A literature survey on social engineering attacks: Phishing attack." 2016 international conference on computing, communication and automation (ICCCA). IEEE, 2016.

[11] Pratik Patil1, Prof. P.R. Devale, “A Literature Survey of Phishing Attack Technique”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 4, April 2016 DOI 10.17148/IJARCCE.2016.5450 198

[12] G. Jaspher Willsie Kathrine, Paradise Mercy Praise, A. Amrutha Rose, Eligious C Kalaivani."Variants of phishing attacks and their detection techniques", 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019