International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Phishing Website Detection using Supervised Deep Learning

L Lakshmi, K Pushpa Rani, K Vijay, G.V.S.S Priyanka

***Abstract**: The website phishing is the tremendously growing problem over the internet which will lead to the loss of personal information. This process will run like, when ever user clicks a website link it will lead them to the web page that is created by the phisher to deceive the user. After this phishing has been started in order to stop it many techniques came into existence to detect the phished web site and help the user from being deceived by the attacker. Even though many techniques have adapted to stop the attackers, it is difficult because as many phished web pages are generated by the attackers within few hours. Most of the techniques to detect these phishing websites are not able to decide the fake website with legitimate one because the accuracy of getting results are much less.*

There are many supervised machine learning techniques which are supervised, where a primary set of data is given to the algorithm and depending on that set the algorithm will be trained and it will predict the results for the same. One of the most important techniques that is deep Learning classifiers is applied with significant features to detect phishing websites. By using this algorithm we can classify the phishing websites from genuine websites by using effective features. In this algorithmic approach to detect genuine websites a feature set is used so by analyzing these features using deep neural networks we can detect a website is phished or not. We can also increase the accuracy of our algorithm by adding certain more features and increasing the hidden layers in neural networks.

Keywords: Deep learning, Feature set, supervised learning, Phishing.

I. INTRODUCTION

Phishing is a fraudulent attempt done by the attackers to steal the personal information from the users by creating some fake websites. In these situations when users enter their information about anything related to transactions, user id's, passwords etc in the fake web pages that information can be miss used by the attacker which may lead to financial loss and personal information.

As the technology is being developed daily and many companies are providing huge amounts of data uploaded and used daily. Many online businesses, bank services etc, are loosing their reputation because of the generation of these fake websites. It will be very helpful for all if we can detect these web pages in the early phase only. However detecting these websites is quite difficult as many innovative techniques have been used to create them.

Even though many techniques have been suggested for the detection of websites but many are not effective enough to produce 100% accurate results and moreover many new phishing websites can be launched within seconds. Any phishing detection website can be a success when it can detect the phished website accurately and in less time. It is very important to detect the phishing websites early, so that we can warn the users before they provide their information to the attacker.

Many conventional phishing websites detection techniques have been put forth like Navie Bayes, SVM, Adagrad etc, which are based on supervised machine learning techniques, where they need to get trained with the data set before performing on the actual data set.

II. LITERATURE SURVEY

A. EXISTING SYSTEM

There are many methods and algorithms by which websites can be classified into phishing or not, but the outcome of these algorithms can provide only 67% of accuracy and above. In these cases these algorithms like SVM, Navie Bayes, Adagrad etc, can show wrong results which can put the client into danger. So, to solve these kind of problems we are proposing a supervised deep learning algorithm.

B. DISADVANTAGES OF EXISTING SYSTEM

It will be difficult if the website detection techniques based on supervised machine learning algorithms produce wrong results when it comes to that remaining 33% of accuracy.

C. PROPOSED SYSTEM

In order to overcome the above mentioned problem we are using a deep learning based anti-phishing approach that can detect phishing websites. In this algorithmic approach to detect a genuine website from phished one a feature set is used, so by analyzing these set using deep neural networks we can detect whether a website is phished or not. We can also increase the accuracy of our algorithm by adding certain more features and increasing the hidden layers in neural network.

Based on the statistics thirty features were defined and will subjected to deep neural networks which uses 2-3 hidden layers to effectively determine legitimate and phished URL.

Manuscript published on November 30, 2019.
* Correspondence Author

L Lakshmi*, Department of Computer Science & Engineering, BVRITH College of Engineering for Women, Hyderabad (Telangana), India.

K Pushpa Rani, Department of Computer Science & Engineering, BVRITH College of Engineering for Women, Hyderabad (Telangana), India.

K Vijay, Department of Computer Science & Engineering, BVRITH College of Engineering for Women, Hyderabad (Telangana), India.

G.V.S.S Priyanka, Department of Computer Science & Engineering, BVRITH College of Engineering for Women, Hyderabad (Telangana), India.

By using this method we can get 90% and above accuracy of results and can easily detect phishing websites from the genuine ones.

D. ADVANTAGES OF PROPOSED SYSTEM

The main advantage of this approach is that it doesn't need to be get trained before processing the entire data set, because neural networks will automatically process the data by extracting some features from it and then sends it to the next layers of the network and produces the results to the last layer.

III. REQUIRIMENT ANALYSIS

A. SOFTWARE REQUIREMENTS:

- Operating system - Windows 7 or 10
- Coding language - Python
- Tool - Tensor Flow
- Data base - Google Colab Notebook

B. HARDWARE REQUIREMENTS:

- Hard disk - 120 GB

IV. PROPOSED SYSTEM

There are many machine learning techniques which are supervised; one of the most important technique that is Deep Learning classifiers is applied with significant features to detect phishing websites. By using this algorithm we can classify the phishing websites from genuine websites by using effective features. In this algorithmic approach to detect genuine websites a feature set is used so by analyzing these features using deep neural networks we can detect a website is phished or not.

We can also increase the accuracy of our algorithm by adding certain more features and increasing the hidden layers in neural networks.

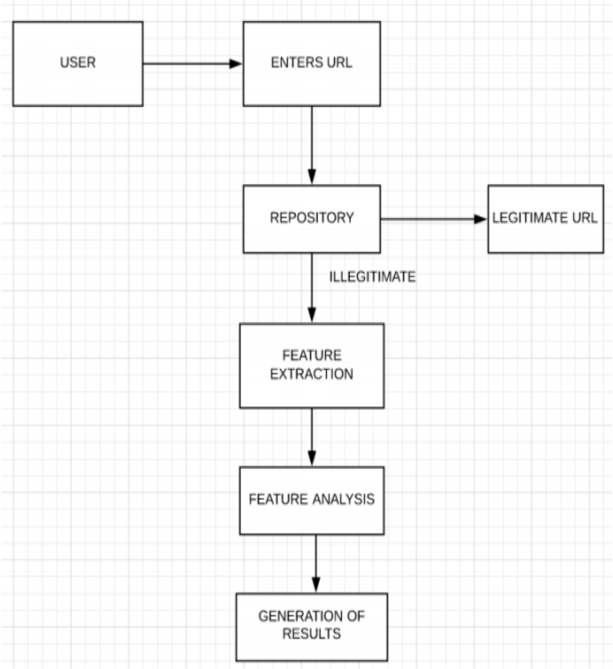

Fig 1: Architecture of Proposed System

In order to detect the websites we are using deep neural networks which can process the data on its own without any supervision where the calculated results of one hidden layer will then be sent to next layer for processing. At the beginning the input layer will be given a input from the training and the test data set and it will be process the data without any external help then the information that has been processed by it will be given to its next layers of the neural network. In this way the final results will be delivered to the output layer.

In this technique, we repeatedly iterate through the training set and update the model parameters in accordance with the gradient which is used for computing model results of error with respect to training set.

Here when implementing our model we create a forward propagation, which creates its own computational graph depending on which a model can be preceded. We don't need to implement back propagation as it implements by itself in the back end through the computational graph. These neural networks are also used for pattern recognition techniques. We can also use dropout case where there may be a chance of over fitting of data which can result in accurate and correct results.

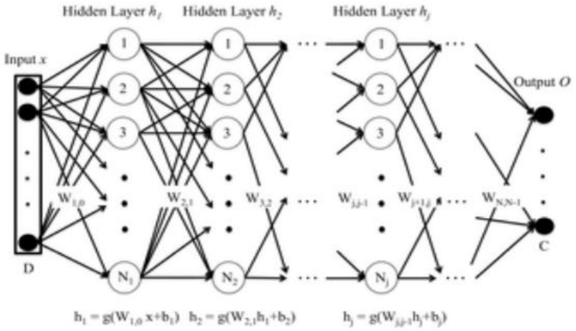

Fig 2: Architecture of deep neural network

An artificial neural network, or neural network, is a mathematical model inspired by biological neural networks. In most cases it is an adaptive system that changes its structure during learning. There are many different types of NNs. For the purpose of phishing detection, which is basically a classification problem, we choose multilayer feed forward NN. In a feed forward NN, the connections between neurons do not form a directed cycle. Contrasted with recurrent NNs, which are often used for pattern recognition, feed forward NNs are better at modeling relationships between inputs and outputs. In our experiments, we use the most common structure of multilayer feed forward NN, which consists of one input layer, and any number of hidden layers depending upon the data set and one output layer. The number of computational units in the input and output layers corresponds to the number of inputs and outputs. Different numbers of units in the hidden layer are attempted in the following experiments. For evaluation our dataset, hyperbolic tangent and sigmoid are used as activation functions. A comparison of the two is also conducted.

In deep learning, optimization is the most important part for improving and generalizing the data. Based on the situation of train and test accuracy we prefer different optimization techniques. For Example if our data is showing higher accuracy for training data but doing worse on test data then it is called Over fitting.

In order to overcome this problem we generally use dropout or L2 regularization or RMS Prop or Adam optimization. If our data is under fitting on our train data then we need to improve the architecture size and try to get more data to overcome this problem.

V. CONCLUSION

In this paper, we proposed solutions to the problem of detecting phishing websites from the genuine ones. We proposed a model of classification using supervised machine learning algorithm that is deep learning. Based on this algorithm, deep neural networks with mini batch gradient descent are applied to get accurate results by extracting the feature set. Our empirical results show that by using this neural networks we can get more accurate and fast results, if we add extra feature sets to this data set we can get perfect results.

REFERENCES

1. *Syed H (September 24, 2012).* "Editorial: Why You Should Go Nexus". *Droid Lessons.* Retrieved April 17, 2013.
2. Jayanthi A., Ramu G., Pushpa Rani K.," A review on personalized medicine technique using cognitive computing",In International Journal of Engineering and Technology(UAE) 2019.
3. Lakshmi L., Bhaskara Reddy P., Shoba Bindu "SLOLAR: Scalable listwise online learning algorithm for ranking",In Journal of Advanced Research in Dynamical and Control Systems 2018.
4. Divya Jyothi G.,Sowmya G., Navya K.," Machine learning and mining for social media analytics",In Advances in Intelligent Systems and Computing 2019.
5. "The Android Source Code". *source.android.com*. Retrieved February 2, 2017.
6. Code Samples, https://developers.google.com/maps/documentation/javascript/examples.
7. Google Maps APIs, Documentation, https://developers.google.com/maps/documentation/android-api/start.
8. Responsive Web design Templates, https://www.w3schools.com/css/css_rwd_templates.asp.
9. Android Studio, " Add App Resources" , https://developer.android.com/studio/rite/addresources.html.