DETECTION OF PHISHING WEBSITES USING ML

CHAVA NAVEEN Student (MCA), NRI INSTITUTE OF TECHNOLOGY, A.P., India.

M.V.P.UMA MAHESWARA RAO Assistant Professor, Dept. of MCA, NRI INSTITUTE OF TECHNOLOGY, A.P., India.

Abstract —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.Here proposed a multidimensional element phishing recognition approach dependent on a quick discovery method by using deep learning (MFPD). In the initial step, character succession highlights of the given URL are separated and utilized for snappy characterization by profound learning, and this progression doesn't need outsider help or any earlier information about phishing. In the subsequent advance, we consolidate URL measurable highlights, website page code highlights, site page content highlights and the brisk characterization consequence of profound learning into multidimensional highlights. The methodology can diminish the identification time for setting an edge. Testing on a dataset containing a huge number of phishing URLs and genuine URLs, the exactness arrives at 98.99%, and the bogus positive rate is just 0.59%. By sensibly changing the limit, the test results show that the discovery effectiveness can be improved

INTRODUCTION

The Internet is widely used among people and it has become an inseparable part of our life. Therefore, huge amounts of data are exchanged. Those users could be more or less experienced using the web. But, nevertheless, nobody is safe from the huge threat that is available there outside. Those threats are phishing websites that are hard to differentiate from the original ones. These websites are used to collect personal and confidential user data that usually should be protected. Later, information is misused and people are experiencing consequences. Some of the consequences could be identity loss or financial debts.Statistics for 2019 states that 15% of those who were successfully attacked will be attacked at least one more time within a year. Number of phishing attacks increased by 65% in respect to 2018 and around 1.5 million of phishing websites were created each month [1]. Almost one third of all data breaches in 2017 were due to phishing attacks. Approximately 55% of phishing websites in 2019 used SSL certificates. Research also shows that 33% of people closed their business after a phishing attack .The problem with phishing attacks is not only that they are increasing, but also, they are improving and becoming more sophisticated. Due to that, it is necessary to develop systems that will help in detection of these phishing websites to prevent negative outcomes. Therefore, in this work we want to develop an intelligent system that will be used to detect phishing websites. We are going to use machine learning algorithms for classification such as K-Nearest Neighbor (KNN), Decision Tree and Random Forest (RF).The rest of the work is organized as follows: in section two, we are giving overview of several works related to the phishing websites detection.

LITERATURE SURVEY

1.Yi et al. proposes a method which detects phishing websites. Detection model is based on a deep belief network (DBN). Two types of features are used: original and interaction features. Original features are those directly related to the websites, while interactive features include features related to the interaction between websites such as in-degree and out-degree of URL. To test DBN real IP flows data are used. Dataset includes traffic flow for 40 minutes and 24 hours. Features like IP address, access time, URL, request page source, user agent and user cookie are extracted. Information about each node is collected and connected to the graph. Contrastive Divergence algorithm is used as a training algorithm. In the final experiment there are three parameters that were changed to find the best combination. Those are the number of layers, number of iterations per layer and number of hidden layers. Approximate true positive rate is approximately 90%. The highest detection rate is achieved with two layers, as the number of layers increases, accuracy decreases. Optimal number of iterations is 250 and the number of hidden layers is between 20 and 40.

2.Mahajan and Siddavatam present a method for improvement of phishing websites detection. Dataset contains URLs of legitimate and phishing websites. Legitimate URLs are collected from www.alexa.com and phishing URLs are collected from www.phishtank.com. Python program is used to extract features from these URLs. Extracted features are presence of IP address in the URL, presence of @ in the URL, number of dots in hostname, prefix or suffix separated by - to domain, URL redirection, HTTPS token in URL, length of host name, number of slash in URL, presence of unicode in URL, age of SSL certificate, URL of anchor, iframe and website rank. Applied algorithms are decision trees, random forests and support vector machines. Dataset is divided into training and testing dataset in ratio 50:50, 70:30 and 90:10. For implementation of the experiment, the authors used the Scikit-learn tool. The highest accuracy 97.14% is achieved using Random Forest. Also, accuracy increased by increasing the number of instances in the training dataset. For the future work, authors are planning to implement hybrid solutions which will combine machine learning algorithms and blacklist methods.

3) A. Ahmad Y, M. Selvakumar, A. Mohammed, A. Mohammed and A. S. Samer, “TrustQR: A Detection of Phishing Attacks on QR Code,” Adv. Sci. Lett., vol. 22, no. 10, pp. 2905-2909, Oct. 2016

Graphic black and white squares, known as Quick Response (QR) code is a matrix barcode, which allows easy interaction between mobile and websites or printed material by getting rid of the need of physically composing a URL or contact data. From the pages of magazines to the sides of transports and announcements, QR code innovation is being utilized progressively in cell phones. Lamentably, Phishers have begun utilizing QR code for phishing assaults by utilizing a few highlights of QR code. This paper presents another methodology called "TrustQR" which identifies URL phishing on QR code. It uses QR code specific features and URL features to detect if the QR code content has a phishing URL. Some of the QR code specific features use QR code content and its characteristics like length, type, and level of error correction to generate the cryptography key. This technique uses the machine learning classification technique.

PROPOSED SYSTEM

- ❖ A multidimensional component phishing identification approach dependent on a quick recognition technique by utilizing profound learning. In the _rst step, character grouping highlights of the given URL are removed and utilized for snappy classi_cation by profound learning, and this progression doesn't require thirdparty help or any earlier information about phishing. In the subsequent advance, we consolidate URL factual highlights, page code highlights, website page content highlights, and the brisk classi_cation aftereffect of profound learning into multidimensional highlights. The methodology can decrease the location time for setting an edge. Testing on a dataset containing a large number of phishing URLs and genuine URLs, the precision arrives at 98.99%, and the bogus positive rate is just 0.59%. By sensibly altering the limit, the exploratory outcomes show that the recognition ef_ciency can be improved.

IMPLEMENTATION

- Data Acquisition: Upload the URL data from the local host

- Data Preprocessing: In this module, we will perform label encoding, convert the text data into token counts and quantify a word in documents, we generally compute a weight to each word which signifies the importance of the word in the document and corpus.

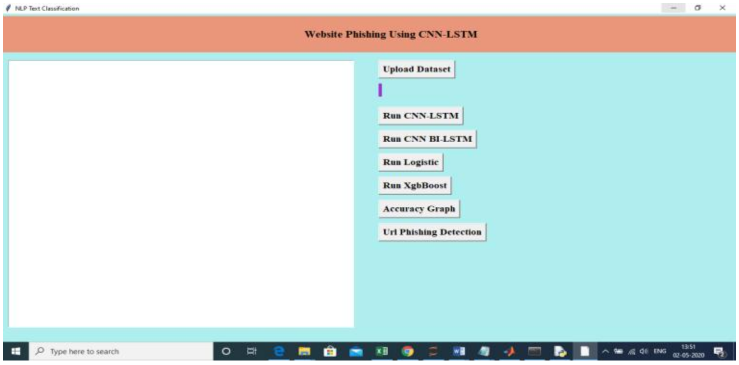
- Spliting: In this module we will split the data into train and test data. x Train and y Train become data for the machine learning, capable to create a model.Once the model is created, input x Test and the output should be equal to y Test. The more closely the model output is to y Test: the more accurate the model is.

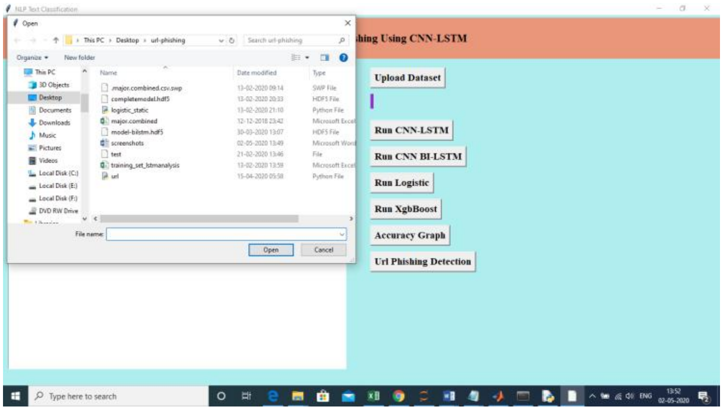
- Modelling: in this module, we will apply the CNN-LSTM and CNN-BiLSTM on URL text and we will apply the ,machine learning algorithms on the features of URL.

- Compariosn: Visualize the varies accuracy of modeling

- Prediction: Url phishing detection on the new site.

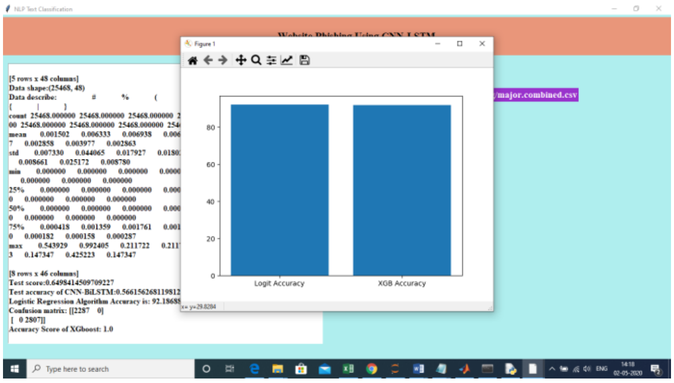
SAMPLE OUTPUT SCREENSHOTS





Name	Date modified	Type
.major.combined.csv.swp	13-02-2020 09:14	SWP File
completemodel.hdf5	13-02-2020 20:33	HDF5 File
logistic_static	13-02-2020 21:10	Python File
major.combined	12-12-2018 23:42	Microsoft Excel
model-bilstm.hdf5	30-03-2020 13:07	HDF5 File
screenshots	02-05-2020 13:49	Microsoft Word
test	21-02-2020 13:46	File
training_set_lstmanalysis	13-02-2020 13:59	Microsoft Excel
url	15-04-2020 05:58	Python File

File name:



```
[5 rows x 48 columns]
Data shape:(25468, 48)
Data describe:              #         %         (
{         |         }
count  25468.000000  25468.000000  25468.000000  2
00  25468.000000  25468.000000  25468.000000  254
mean      0.001502      0.006333      0.006938      0.006
7     0.002858      0.003977      0.002863
std       0.007330      0.044065      0.017927      0.0180
   0.008661      0.025172      0.008780
min       0.000000      0.000000      0.000000      0.0000
    0.000000      0.000000      0.000000
25%       0.000000      0.000000      0.000000      0.000
0     0.000000      0.000000      0.000000
50%       0.000000      0.000000      0.000000      0.000
0     0.000000      0.000000      0.000000
75%       0.000418      0.001359      0.001761      0.001
0     0.000182      0.000158      0.000287
max       0.543929      0.992405      0.211722      0.211
3     0.147347      0.425223      0.147347

[8 rows x 46 columns]
Test score:0.6498414509709227
Test accuracy of CNN-BiLSTM:0.566156268119812
Logistic Regression Algorithm Accuracy is: 92.18688
Confusion matrix: [[2287    0]
 [   0 2807]]
Accuracy Score of XGboost: 1.0
```

CONCLUSION

In this paper, It is well known that a good phishing website detection approach should have good real-time performance while ensuring good accuracy and a low false positive rate. Our proposed MFPD approach is consistent with this idea. Under the control of a dynamic category decision algorithm, the URL character sequence without phishing prior knowledge ensures the detection speed, and the multidimensional feature detection ensures the detection accuracy. We conduct a series of experiments on a dataset containing millions of phishing and legitimate URLs. From the results, we _nd that the MFPD approach is effective with high accuracy, low false positive rate and high detection speed. A future development of our approach will consider applying deep learning to feature extraction of webpage code and webpage text. In addition, we plan to implement our approach into a plugin for embedding in a Web browser.

REFERENCES

[1] (2018). *Phishing Attack Trends Re-Port-1Q*.Accessed: May 5, 2018.
[Online]. Available:
https://apwg.org/resources/apwg-reports/

[2] (2017). *Kaspersky Security Bulletin: Overall Statisticals For*. Accessed:
Jul. 12, 2018. [Online]. Available: https://securelist.com/ksb-overallstatistics-2017/83453/

[3] A.Y. Ahmad, M. Selvakumar, A. Mohammed, and A.-S. Samer, ``TrustQR:
A new technique for the detection of phishing attacks on QR code," *Adv.*
Sci. Lett., vol. 22, no. 10, pp. 2905_2909, Oct. 2016.

[4] C. C. Inez and F. Baruch, ``Setting priorities in behavioral interventions:
An application to reducing phishing risk," *Risk Anal.*, vol. 38, no. 4,
pp. 826_838, Apr. 2018.

[5] G. Diksha and J. A. Kumar, ``Mobile phishing attacks and defence mechanisms:
State of art and open research challenges," *Comput.Secur.*, vol. 73,
pp. 519_544, Mar. 2018.

[6] *Google Safe Browsing APIs*.Accessed: Oct. 1, 2018.[Online]. Available:
https://developers.google.com/safe-browsing/v4/

[7] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang,
``An empirical analysis of phishing blacklists," in *Proc. 6th Conf. Email*
Anti-Spam (CEAS), Jul. 2009, pp. 59_78.

[8] A. K. Jain and B. B. Gupta, ``A novel approach to protect against phishing
attacks at client side using auto-updated white-list," *EURASIP J. Inf.*
Secur., vol. 2016, no. 1, Dec. 2016, Art. no. 34.

[9] M. Zouina and B. Outtaj, ``A novel lightweight URL phishing detection
system using SVM and similarity index," *Hum.-Centric Comput. Inf. Sci.*,
vol. 7, no. 1, p. 17, Jun. 2017.

[10] E. Buber, Ö. Demir, and O. K. Sahingoz, ``Feature selections for the
machine learning based detection of phishing websites," in *Proc. IEEE*
Int. Artif. Intell.Data Process.Symp.(IDAP), Sep. 2017, pp. 1_5.

[11] J. Mao *et al.*, ``Detecting phishing websites via aggregation analysis of page layouts,'' *Procedia Comput. Sci.*, vol. 129, pp. 224_230, Jan. 2018.

[12] J. Mao,W. Tian, P. Li, T.Wei, and Z. Liang, ``Phishing-alarm: Robust and ef_cient phishing detection via page component similarity,'' *IEEE Access*, vol. 5, no. 99, pp. 17020_17030, Aug. 2017.