ISSN: 2278-4632
Vol-11 Issue-01 2021

A SURVEY ON PHISHING WEBSITES DETECTION USING MACHINE LEARNING

MANDA JAYASRI Student, M.Tech (CSE), VIKAS COLLEGE OF ENGINEERING & TECHNOLOGY, A.P., India.

Mr. G. YAMINI SATISH Assistant Professor , Dept. of Computer Science & Engineering, VIKAS COLLEGE OF ENGINEERING & TECHNOLOGY, A.P., India.

Abstract — Phishing is a common attack on credulous people by making them to disclose their unique information using counterfeit websites. 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

Phishing is popular among attackers, since it is easier to trick someone into clicking a malicious link which seems legitimate than trying to break through a computers defense systems. The malicious links within the body of the message are designed to make it appear that they go to the spoofed organization using that organizations logos and other legitimate contents. 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

ISSN: 2278-4632
Vol-11 Issue-01 2021

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

S. Marchal et al., (2017) proposed this technique to differentiate Phishing website depends on the examination of authentic site server log knowledge. An application Off-the-Hook application or identification of phishing website. Free, displays a couple of outstanding properties together with high preciseness, whole autonomy, and nice language-freedom, speed of selection, flexibility to dynamic phish and flexibility to advancement in phishing ways.

Mustafa Aydin et al. proposed a classification algorithm for phishing website detection by extracting websites' URL features and analyzing subset based feature selection methods. It implements feature extraction and selection methods for the detection of phishing websites. The extracted features about the URL of the pages and composed feature matrix are categorized into five different analyses as Alpha-numeric Character Analysis, Keyword Analysis, Security Analysis, Domain Identity Analysis and Rank Based Analysis. Most of these features are the textual properties of the URL itself and others based on third parties services.

Samuel Marchal et al. presents PhishStorm, an automated phishing detection system that can analyze in real time any URL in order to identify potential phishing sites. Phish storm is proposed as an automated real-time URL phishingness rating system to protect users against phishing content. PhishStorm provides phishingness score for URL and can act as a Website reputation rating system.

ISSN: 2278-4632
Vol-11 Issue-01 2021

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

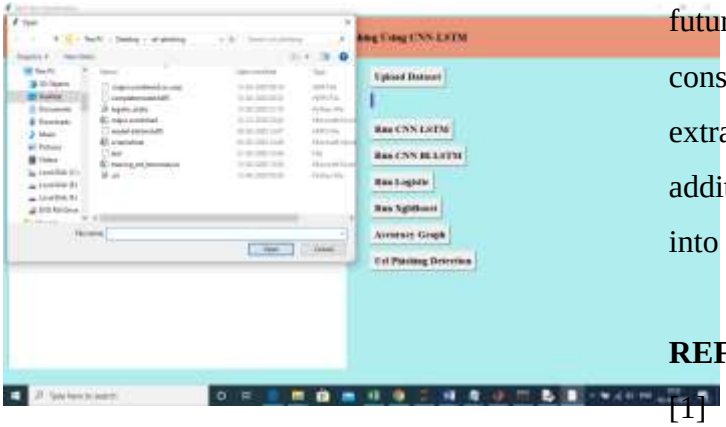
- Data Acquisition: Upload the URL data from the local host
- Data Preprocessing: In this module, we will perform label encoding, convert the text data into token counts and quantify a word in documents, we generally compute a weight to each word which signifies the importance of the word in the document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

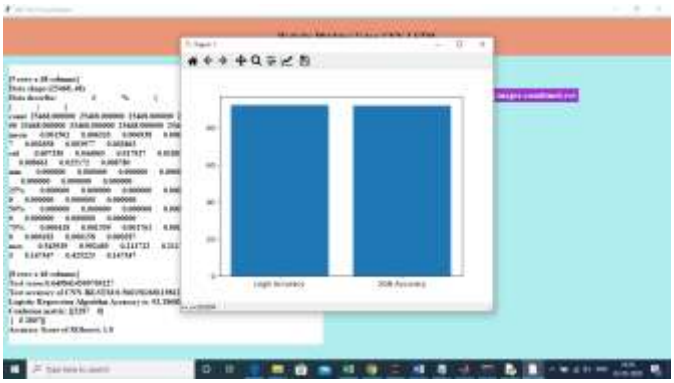
ISSN: 2278-4632
Vol-11 Issue-01 2021

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.

SCREENSHOTS





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.