JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2373 - 2378
https://publishoa.com
ISSN: 1309-3452

Using Machine Learning to Find Phishing Websites

Venkatram Vennam [1], Rayan Iqbal Abdul Hafeez [2], P Samiullah Khan [3], Mohd fareeduddin Faraz [4], Syed Naveed [5]

[1] *Assistant Professor, Department of Computer Science and Engineering, Lords Institute of Engineering and Technology, Hyderabad, Telangana, India.*
[2, 3, 4, 5] *Research Scholar, Department of Computer Science and Engineering, Lords Institute of Engineering and Technology, Hyderabad, Telangana, India.*

Email : [1] *venkatram1246@gmail.com*

ABSTRACT

Phishing is one of the easiest ways to get delicate data from users who aren't doing anything wrong. The goal of phishing emails is to get important identities of users, pass codes, and bank statement data [1]. People who work in data protection nowadays are searching for reliable and consistent ways to find phishing sites. This article is about using ml algorithms to find phishing Websites. It does this by gathering and analysing different characteristics of both real and fake URLs. To find malicious urls, methodologies like Logistic Regression, Random Forest, as well as Support Vector Machine have been used. The aim of the research is to find phishing URLs and find the best machine learning technique [13] by making comparisons every method's overall accuracy, true positive rate, and negative predictive value.

Index Terms— Extraction Of features, Phishing Website, Malicious Scams, and Phishing Detection.

I. INTRODUCTION

In the twenty-first generation, technology is an essential part of being human. A few of these techniques is the online world, that also grows quickly every year has a big effect on health outcomes. It became an important as well as easy way to promote social purchases like online banking shopping. Because of this, online consumers think it's easy to give their personal data to a Website. For this, fraudsters had also begun going after such data, which is a big security issue. [2] People think that phishers WebPages were one of those troubles. Those who have used a media manipulation technique, which means that they are tricking the consumer in and out of offering, people their own private data by taking advantage of weaknesses in people instead of operating system. Results show that the majority of phishing attempts keeps rising, which puts customer data in the hands, according the [8] Anti-Phishing Workgroup (APWG) and registered phishing scams by Security Researchers, which said that it has improved by 47.48 percent from most of the phishing websites that were found in 2016. In recent decades, there have been a body of research that attempted to figure out how to stop phishers. Some authors was using the URL to make comparisons it to established watch lists of fraudulent websites, which they've been making, while others was using it to start comparing it to a white list of internet sites. [2] The second method uses a signature large database attack vectors that complement the fingerprint of something like the methodology sequence to start deciding if a webpage is phishing or not. [1] Researchers have also used Alexia to measure how much traffic a website gets as another way to find phishing websites. Also, ml algorithms have been used by other researchers. Machine learning is an application of computer science and artificial intelligence (AI) that allows machines to do tasks and learn or act in a smart way. It has two ways to learn: learning with a teacher and learning on your own [11]. In supervised methods, a classifier is constructed by having given it a collection of analysis should be undertaken of information that is linked to a target label. That once classifier is developed, it can use extracted features to come up with a new target label. From the other side, unsupervised learning that is based on producing new information without offering whatever intended labels during training.

II. RELATED WORK

A way to find phishing websites based on their content

Phishing is a big issue that involves fake emails as well as WebPages which mislead people and giving out individual data. Inside this article, designers propose the development, deployment, as well as assessment of CANTINA [10], a new way to find malicious activity content - based and the TF-IDF method for information retrieval. We also talk about how we designed and tested the methodologies we made to cut down on false - positive results. From our tests, we know that CANTINA is great at finding phishing websites because it appropriately labels about 95% of them.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2373 - 2378
https://publishoa.com
ISSN: 1309-3452

Detection methods for zero day phishing sites

A technique known as "phishing" involves tricking people into sharing personal information on phoney websites that look real. Phishing identification techniques are plentiful, although existing practises were lacking. Many phishing websites are only around for several hours, therefore web pages depend on a blacklisted of recognised bad websites. The online browser needs a speedier algorithm to recognise zero-day phishing websites [4], which are newly found malicious websites.
Use cascading style sheets to detect fraudulent web sites, which is a new strategy that is not currently employed by major anti-phishing software (CSS). Hundreds of well-known phishing sites [9] are tested against a new detection system that is compared to the most effective ones currently in use.

Anti-phishing software for the PC that uses the Heuristic Approach

Phishing is the act of creating a fictitious website in order to get access to a victim's personal data. Text messages, voicemails and emails are all used by the attacker to trick the user into believing they've received a message.
A desktop tool named Phish Shield was developed [5] in this project that focuses on the URL and the contents of phishing pages. Phish Shield uses a URL as an input and returns a result indicating if the URL points to a phishing site or an official one. To identify scamming, navigation bar connections with such a default values, blank links in the content of HTML, copyrights contents, title information, as well as the identification of the webpage were utilised as hunches [8]. Zero-hour phishing attacks, which blacklists will be unable to recognize, can be detected by Phish Shield, which is quicker then image processing and analysis techniques used to identify phishing. Because it detects so many different types of fake websites with such high precision, Phish Shield has less false negatives and false positives than other products.

A heuristic technique that is based on URLs is taken to find phishing websites

E-commerce has emerged as a significant force in today's world, thanks to the rapid development of the internet. As a direct consequence of this, scamming, which is the practise of acquiring private customer information being used in the transactions of e-commerce, has developed into an urgent issue in culture today. A great number of strategies, such as blacklists and page ranks, were suggested as ways to safeguard consumers of the internet. On the other hand, the majority of casualties have been steadily growing due to the ineffectiveness of the defensive response. Phishers frequently attempt to create websites that have URLs that are essentially equivalent to those of legitimate websites. This is why this occurs. In this research, we want to propose a new technique to analyze phishing websites by making use of the characteristics of URLs. In specifically, we extract a variety of elements from the URL and then calculate a metrics specific to each extracted component. After that, the webpage rankings would be factored in alongside the metrics [7] that have been attained in order to make a decision whether or not the websites in question are scamming sites. The recommended method for detecting hacking was validated using a dataset that included 9,661 fraudulent websites and 1,000 genuine websites. According to the findings, the method that we have proposed is capable of identifying more than 97 percent of phishing websites.

III. METHODOLOGY

The use of a phishing attack is one of the simplest ways to get sensitive information from unsuspecting customers [1]. The objective of the phishing scams is to acquire fundamental information such as the login, decryption key, other block chain and distributed nuances. Those responsible for internet security were currently looking for consistently reliable localization mechanisms for the identification of phishing websites.
In a digital communication, phishing is a dishonest attempt to get sensitive information, such as identities, passwords, and credit card numbers, by posing as a trustworthy person [1]. Multiple kinds of phishing attacks could be carried out, such as the use of phishing emails or webpages, the use of phishing scam, the use of Whale hunting or Tab Taking naps, or the use of Evil Twin phishing. Anti-phishing measures should be used in order to prevent this type of assault. Blacklisted, heuristic, visual similarity and machine learning are just a few of the anti-phishing technologies that exist.

A. Blacklist method

A database of phishing URLs is kept, and if the URL is located in the directory, it is classified as a phishing URL and a notification is issued; [6] alternatively, it is considered to be legit. As it determines whether or not the URL is stored in a database, this method is simple and quick to build. Restrictions include the fact that a little changes in URL is enough to defeat the list-based strategy, and that the list must be updated frequently in order to protect against new attacks.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2373 - 2378
https://publishoa.com
ISSN: 1309-3452

B. Heuristic based method

New attacks can be detected by extracting features from the phishing site and using them to identify phishing attacks [8]. When an attacker knows the algorithms or features of the system they are using, it is much easier for them to get around security measures. The website seems not to have similar traits, which makes detection difficult.

C. Visual similarity

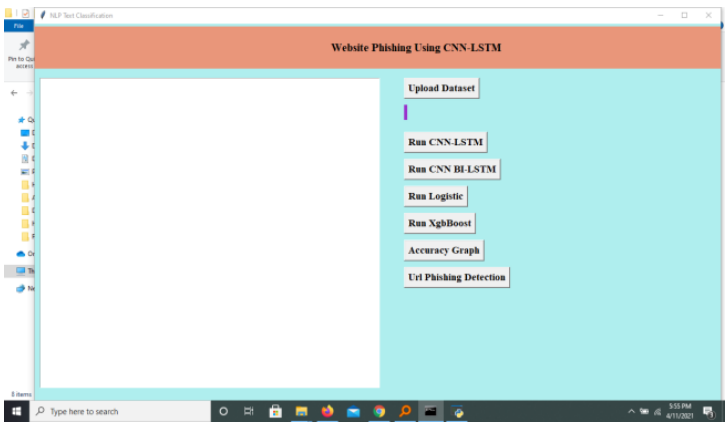
Using a picture from a respectable website to trick users is an effective way to deceive them. However, there is a drawback to this method [3], which is that it requires more time and storage space to compare images. a high rate of false negatives and a failure to identify even modest alterations in visual appearance

D. Machine learning

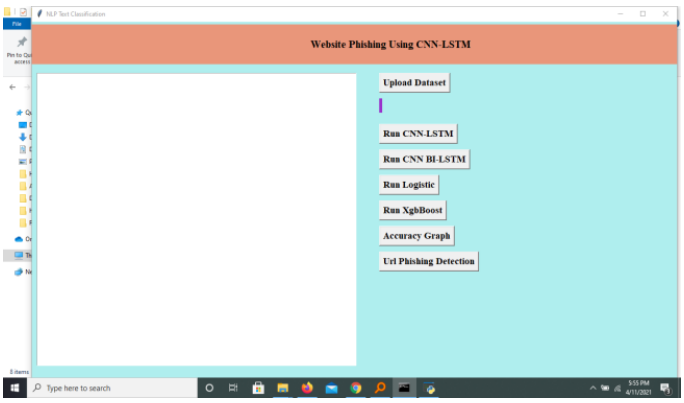
Large datasets are no problem with this method. As an added benefit, this method is capable of detecting zero-day attacks. Classifiers based on machine learning are effective and have an accuracy rate of over 99 percent. Training data amount, feature set, and classifier type all affect performance. However, this has a flaw in that it is unable to detect when an attacker is hosting their site on a hacked domain. There has been a lot of study done on phishing detection. [14] The majority of research has focused on enhancing the detection accuracy of phishing websites by utilising different classifiers. KNN, SVM, Decision Tree, ANN, Nave Bayes, PART, ELM, and Random Forest are a few of the Classifiers in use. DT and RF are the best of the tree-based classifiers, according to my literature review. Tree-based classifiers will be used to detect phishing websites in the proposed approach. F-measure, precision, recall, accuracy, AUC, ROC curve, etc. Are some of the performance measures used to evaluate the best algorithm?
Maintains artificial intelligence technology for the purpose of discovering phishing URLs by extracting and analysing various features of both authentic and phishing URLs [12]. Calculations based on the Choice Tree, the Irregular Woodland, and the Support Vector Machine [14] are applied to differentiate phishing websites. The purpose of this work is to detect phishing URLs using the best AI calculation possible by comparing the precision rate, false positive rate, and false negative rate of each calculation.
The following modules are being utilised in the execution of this project:
Dataset Upload: With the help of this module, we are able to upload datasets to the application.
Run CNN-LSTM: With the help of this module, we would feed all of the images obtained into the CNN [8], and with the help of this component, we would trained the LSTM algorithms using the processing database, and maybe we'll prediction using the testing data.
Run CNN BI-LSTM: The character-level features are brought about by the employment of the CNN component. In order to derive an unique feature representation using the per-character feature vectors, including such characters extracted features and (optionally) characteristics of a particular, the method makes use of a convolution operation and a max-pooling for every text that is being analysed.
Run Logistic: The continuous and categorical parameter can be predicted with the help of this module by utilising a predetermined group of private elements.
Run XgbBoost: By employing this module, we are able to raise the accuracy of the gradient decision trees, which will result in more accurate predictions.
Accuracy Graph: With the help of this module, we would plot the correctness graphs of all three techniques for the characteristics that have been chosen.
Detection of phishing through URL: When an arriving URL is processed by this module, it is checked against a predefined list of phishing characteristics to determine whether or not it should be categorised as malicious.

IV. RESULT AND DISCUSSION

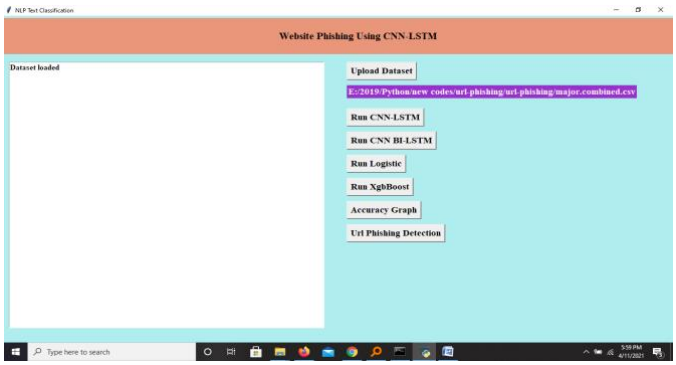
Run the project to get the below result



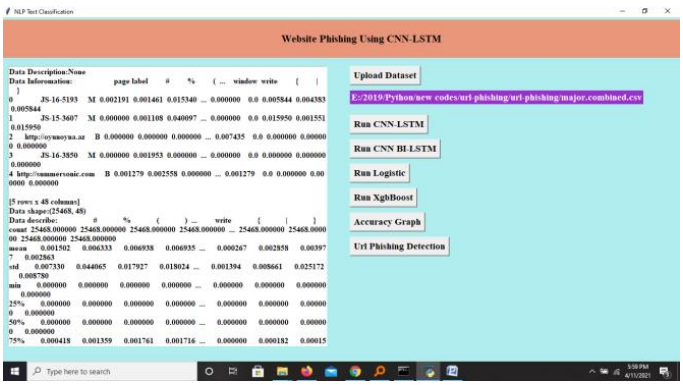
JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2373 - 2378
https://publishoa.com
ISSN: 1309-3452



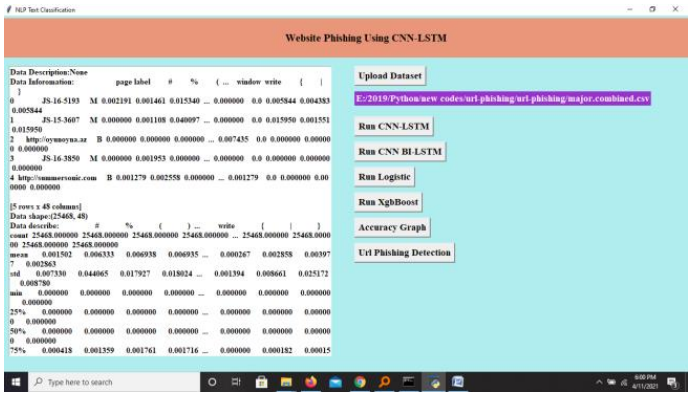
After that, we will upload the dataset in order to obtain the results below.



Next To see the CNN-LSTM results, click the run CNN-LSTM tab.

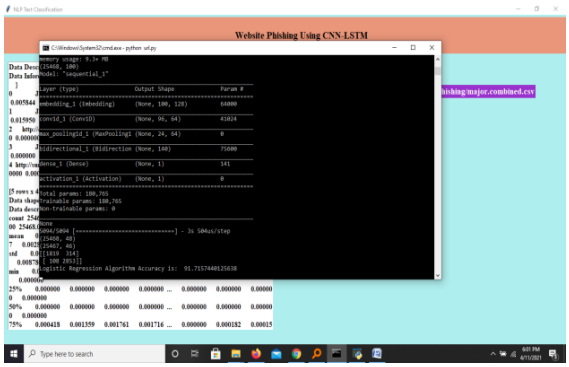


The accuracy of the CNN-LSTM model is shown in the above output. Use the RUN CNN BI-LSTM tab to obtain the following outcome.

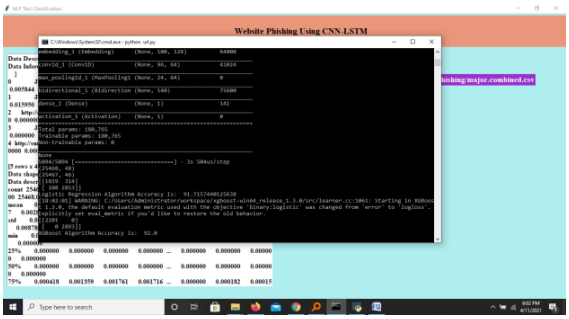


JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2373 - 2378
https://publishoa.com
ISSN: 1309-3452

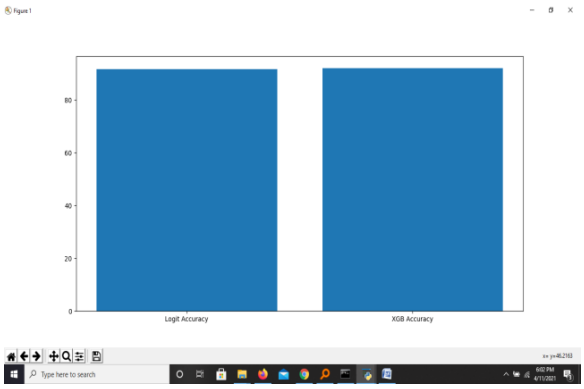
The accuracy values for CNN BI-LSTM are shown in the results above. To obtain the information shown below, select the Run Logistic tab.



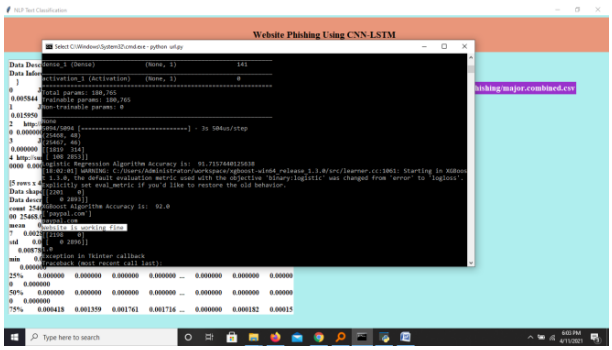
In this case, the correctness of the logistics is clearly shown. To see the results listed below, select the Run XGBOOST tab.



XGBOOST Accuracy can be shown in the above results. To see the graph below, go to the Accuracy Graph tab.



In the graph above, you can see the accuracy of all algorithms. Selecting URL Phishing Detection gives you the information shown below.



The above result indicates whether or not a website is functioning properly.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2373 - 2378
https://publishoa.com
ISSN: 1309-3452

V. CONCLUSION

Through the application of machine learning techniques, the purpose of this paper is to improve detection methods for phishing sites. To use the random forest technique, we were capable of achieving a detection performance of 97.14 percent with the least potential false positive rate [15]. Also, the results reveal that the performance of the classifiers improved when we employed a greater number of examples as training data. The blacklist method and the random forest technique of machine learning algorithms [10] would be utilised in the future when hybrids technologies are deployed to identify phishing websites more effectively. In the future, hybrid technique would be used.

REFERENCES

1. AO Kaspersky lab. (2017). The Dangers of Phishing: Help employees avoid the lure of cybercrime. [Online] Available: https://go.kaspersky.com/Dangers-Phishing-Landing-Page- Soc.html [Oct 30, 2017].
2. ”Financial threats in 2016: Every Second Phishing Attack Aims to Steal Your Money” 2017 financial-threats-in-2016. Feb 22, 2017 [Oct 30, 2017].
3. Y. Zhang, J. I. Hong, and L. F. Cranor, ”Cantina: A Content-based Approach to Detecting Phishing Web Sites,” New York, NY, USA, 2007, pp. 639-648.
4. M. Blasi, ”Techniques for detecting zero day phishing websites.” M.A. thesis, Iowa State University, USA, 2009.
5. R. S. Rao and S. T. Ali, ”PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach,” Procedia Computer Science, vol. 54, no. Supplement C, pp. 147-156, 2015.
6. E. Jakobsson, and E. Myers, Phishing and Counter-Measures: Understanding the Increasing Problem of Electronic Identity Theft. Wiley, 2006, pp.2–3.
7. L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, ”Detecting phishing web sites: A heuristic URL-based approach,” in 2013 International Conference on Advanced Technologies for Communications (ATC 2013), 2013, pp. 597-602.
8. Z. Zhang, Q. He, and B. Wang, ”A Novel Multi-Layer Heuristic Model for Anti-Phishing,” New York, NY, USA, 2017, p. 21:1-21:6.
9. N. Sanglerdsinlapachai and A. Rungsawang, ”Web Phishing Detection Using Classifier Ensemble,” New York, NY, USA, 2010, pp. 210-215.
10. G. Xiang, J. Hong, C. P. Rose, and L. Cranor, ”CANTINA+: A Feature- Rich Machine Learning Framework for Detecting Phishing Web Sites,” ACM Trans. Inf. Syst. Secur., vol. 14, no. 2, pp. 21:1-21:28, Sep. 2011.
11. R. M. Mohammad, F. Thabtah, and L. McCluskey, ”Predicting phishing websites based on self-structuring neural network,” Neural Comput & Applic, vol. 25, no. 2, pp. 443-458, Aug. 2014.
12. Pradeepthi K V and Kannan A, ”Performance study of classification techniques for phishing URL detection,” in 2014 Sixth International Conference on Advanced Computing (ICoAC), 2014, pp. 135-139.
13. S. Marchal, J. Franois, R. State, and T. Engel, ”PhishStorm: Detecting Phishing With Streaming Analytics,” IEEE Transactions on Network and Service Management, vol. 11, no. 4, pp. 458-471, Dec. 2014.
14. A. Sirageldin, B. B. Baharudin, and L. T. Jung, ”Malicious Web Page Detection: A Machine Learning Approach,” in Advances in Computer Science and its Applications, Springer, Berlin, Heidelberg, 2014, pp. 217-224.
15. R. Verma and K. Dyer,”On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers,” New York, NY, USA, 2015, pp. 111-122.