IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Phishing Website Detection

Suraj.K.Shivan [1] , Dr. Muzameel Ahmed [2]

[1] *P.G Student, Department of Information Science Engineering, Dayananda Sagar College of Engineering, Bangalore,India* [2] *Assoc Prof, Department of Information Science Engineering, Dayananda Sagar College of Engineering, Bangalore,India*

Abstract- Phishing is a typical assault on gullible individuals by making them reveal their one of a kind data utilizing fake sites. The goal of phishing site URLs is to steal the individual data like client name, passwords and on the web banking exchanges. Phishers utilize the sites which are outwardly also, semantically like those genuine sites. As innovation keeps on developing, phishing methods began to advance quickly also, this should be protected by utilizing against phishing systems to distinguish phishing. Machine Learning is an amazing asset used to endeavour against phishing assaults. A novel idea is proposed to detect malicious and non-malicious URL links using Extreme Learning Machine Algorithm

Keywords- Phishing, Extreme Learning Machine, malicious URL, non-malicious URL

I. INTRODUCTION

Phishing is the most dangerous criminal activities in digital space. Since the majority of the clients go online to get to the administrations given by government and monetary establishments, there has been a critical expansion in phishing assaults for as long as not many a long time. Phishers began to bring in cash and they are doing this as an effective business. Different techniques are utilized by phishers to assault the weak clients, for example, informing, VOIP, mock connection and fake sites. It is easy to make fake sites, which resembles a certifiable site as far as design and substance. Indeed, the substance of these sites would be indistinguishable from their real sites. The reason for making these sites is to get private information from clients like record numbers, login id, passwords of debit and credit cards, and so forth Also, assailants ask security inquiries to answer to acting like a significant level safety effort giving to clients. At the point when clients react to those inquiries, they get without any problem caught into phishing assaults. Numerous explores have been proceeding to forestall phishing assaults by various networks around the globe. Phishing assaults can be forestalled by identifying the sites and making attention to clients to distinguish the phishing sites. Machine Learning Algorithms have been one of the amazing strategies in identifying phishing sites. In this examination, different techniques for identifying phishing sites have been examined.

II. LITERATURE REVIEW

This paper has proposed a Machine-learning technique for modelling the prediction task and supervised learning algorithms that Multi-Layer Perceptron. Decision tree and Naïve bayes classifications were used for observing [1]. It has been observed that the decision tree classifier predicts the phishing website more accurately than other learning algorithms, but it is not effective for future detection. This paper has proposed a new approach called multi-tier classification model for phishing email filtering [2]. It has a method for extracting the features of phishing email related to weighting of message content and message header and selects the features according to priority ranking. An empirical performance and analysis of the proposed algorithm have been presented. Due to rapidly evolving nature of both legitimate and phishing emails, existing corpus rapidly becomes outdated. This paper describes a Comprehensive survey on state of art of security analytics, which is its description, technology, trends and tools [3].Security analytics aims to detect previously undiscovered threats by use of analytic techniques. Common techniques of security analytics include clustering and graph-based event correlation are used, but not mentioned about the false positive rate for fraudulent transactions. Relationships between files are represented as a graph to detect malware presence. In this paper the scheme first collects from the clients the file lists which describe their mutual relationships, and determines if there are potentially malicious relationships [4]. The file associations are then used to generate an undirected weighted file relationship graph, and based on the graph a belief propagation classifier is trained, but this is unable to determine the information apart from the file contents extracted. This paper uses the Lexical Features as the classifying parameters in the Detection of Malicious URLs, by leveraging the Visible Attributes it is possible to classify the Malicious Short URLs [5].The Social Network giants such as Twitter and Facebook use mainly these kinds of primitive features to know whether to check, technically these systems are called Recommendation systems. Since this focuses only on visible features from tweets, attackers will use this kind of knowledge to spread URLs.

III. METHODOLOGY

FEATURE EXTRACTION

This idea mainly focus on how the malicious and non-malicious URL links are classified, initially feature extraction process is done from the input URL where 21 features are extracted on which it is classified as malicious and non-malicious.

URL	host	TC	RH	RC	ASN	SSWC	ATL	ND	LU	APT	IP ad	LH	Safe	ADTL	PTC	Path	LD	DTC	LP	LT
http://	www	7	-1	-1	36351	0	3.7412	3	35	5	0	21	0	3.4	1	/../	6	5	5	6

The figure above shows the 21 features extracted from the user URL input, the first feature is the URL from where the next feature (host) is taken, the next feature (TC) is the token count where the number of tokens is considered which excludes special characters such as (.,/,?,=,-,_), the next three features (RH), (RC) and (ASN) are the rank host, rank country and the asynchronous system number which is being randomly generated by the Alexa ranking for website popularity, the next feature (SSWC) is the secure sensitive word count where it checks for the secure sensitive words in the URL such as confirm, account, banking, secure, webscr, login, signin. The (ATL) feature is the average token length of the input URL, (ND) is the number of dots in the URL, (LU) is the total length of the URL which includes all the characters of the URL, (APT) is the average path token of the input URL, (IP ad) is the IP address presence in the input URL, (LH) is the length of the host of the input URL, (Safe) is the safe browsing which indicates the SSL certification, (ADTL) is the average domain token length, (PTC) is the path token count which indicates the length of the token found in the path, the next feature is the path, (LD) is the largest domain where the length of the largest token in the domain is considered, (DTC) is the domain token count which indicates the number of tokens found in the domain, (LP) is the largest path where the length of the largest path is considered, (LT) is the largest token of the input URL.

CLASSIFYING URL INTO MALICIOUS AND NON MALICIOUS

The classifier proposed is Extreme Learning Machine (ELM) classifier, where the 21 features of the input URL being extracted are classified into malicious or non-malicious by the trained classifier.

ANALYSIS-1

In the first analysis, 19 features were taken from 50 malicious and 50 non malicious URL links, the first feature is the token count (TC), the token count in malicious link is more when compared to non-malicious link, i.e. the token count of non-malicious is 3 and for malicious it is greater than 3 in most of the URLs, the next three features that is rank host (RH), rank country (RC), Asynchronous Sequence Number (ASN) and Secure Sensitive Word Count (SSWC) gives constant values for both malicious and non-malicious, in the next feature average token length (ATL), the value in non-malicious is either recurring number or a whole number but in the case of malicious the value is a rational number, the next feature is the number of dots (ND) where the number of dots is more than 1 in malicious and 1 in non-malicious for most of the URLs, the next feature that is Length of the URL (LU) is more for malicious link and less for non-malicious URLs, the next feature which is the IP address (IP ad) has no change in malicious and non-malicious, in length host (LH) majority of the value occurs in single digits in non-malicious and double digits in malicious, the next feature i.e., safe browsing(Safe) has no change in both the URLs, the average domain token length (ADTL) has a vague set of values where the ADTL comprise of more single digit values in non-malicious when compared to malicious, the features relating to path such as the Path, Path Token Count (PTC), Largest Path (LP), Average Path Token (APT) has null values in the case of non-malicious when compared to malicious because majority of the malicious URLs consists of a relative path, the largest domain (LD) has a single digit values in non-malicious URLs when compared to malicious URL, the Domain Token Count (DTC) value is 2 in non-malicious and more than 2 in malicious, the last feature is the Largest Token (LT), the value of (LT) occurs in single digits for non-malicious and double digits for most of the malicious URLs.

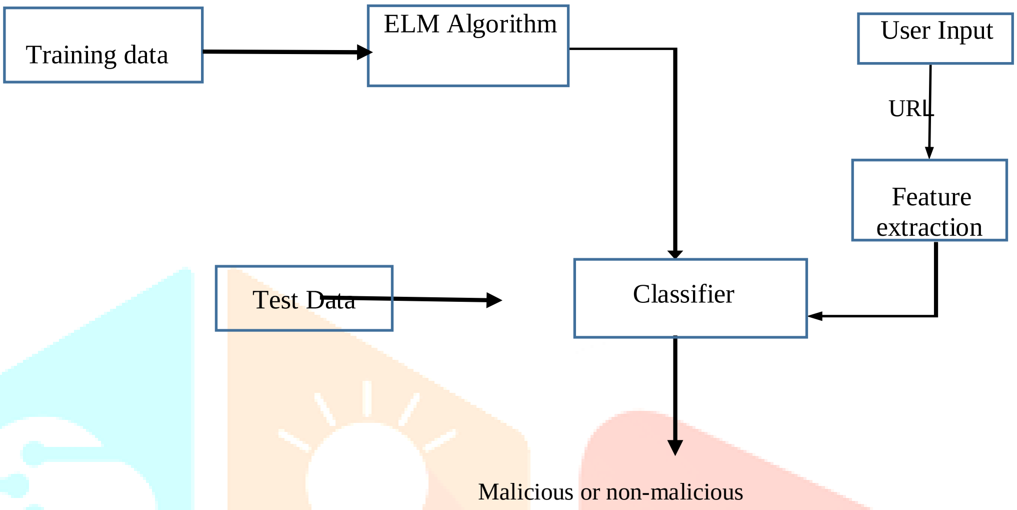
ANALYSIS-2

In the second analysis 8 features were taken from another 50 malicious and 50 non-malicious URLs, 8 features are considered because, these 8 features show less similarities when compared to the previous 13 features which was taken in the first analysis. The 8 features are Token Count (TC), Average Token Length (ATL), Number of Dots (ND), Length of URL (LU), Length of Host (LH), Largest Domain (LD), Domain Token Count (DTC), Largest Token (LT). Token Count (TC) is 3 in non-malicious and more than 3 in most of the malicious URLs, Average Token Length (ATL), the value in non-malicious is either recurring number or a whole number but in the case of malicious the value is a rational number, the number of dots (ND) is more than 1 in malicious and 1 in non-malicious for most of the URLs, Length of the URL (LU) is more for malicious link and less for non-malicious URLs, in length host (LH) majority of the value occurs in single digits in non-malicious and double digits in malicious, the largest domain (LD) has a single digit values in non-malicious URLs when compared to malicious URL, the Domain Token Count (DTC) value is 2 in non-malicious and more than 2 in malicious, the Largest Token (LT), the value of (LT) occurs in single digits for non-malicious and double digits for most of the malicious URLs.

ANALYSIS-3

This analysis further focuses on features which shows less similarities when compared to previous analysis. A new set of 50 malicious and 50 non-malicious URLSs were considered where 3 features were taken which shows negligible similarities. Those 3 features are Token Count (TC), Average Token Length (ATL), Number of Dots (ND), Token Count (TC) is 3 in non-malicious and greater than 3 in all malicious URL, Average Token Length (ATL), the value in non-malicious is either recurring number or a whole number but in the case of malicious the value is a rational number which is non-recurring. Number of Dots (ND) in non-malicious is 1 and in malicious it is greater than 1.

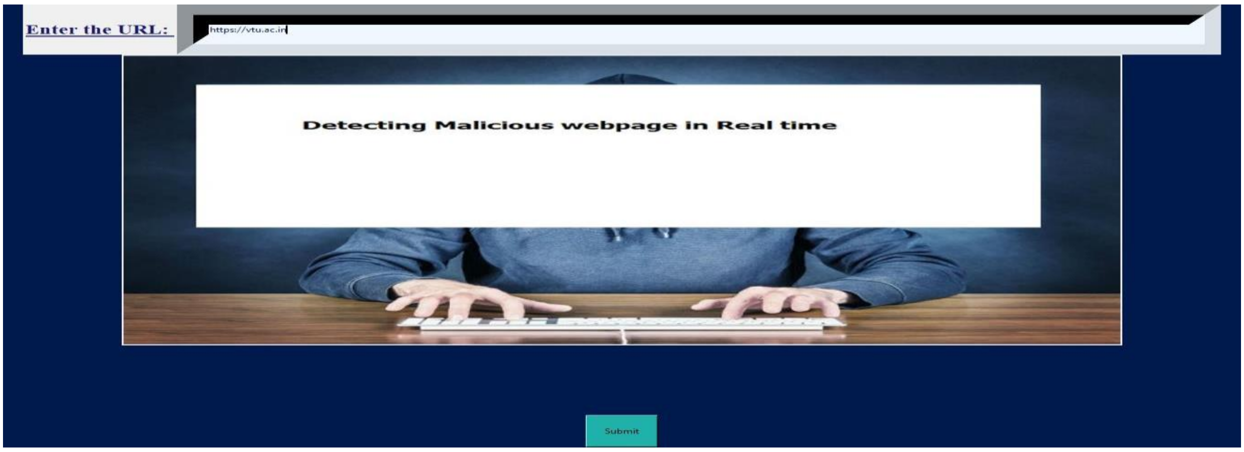
IV. SYSTEM ARCHITECTURE



The dataset used in this system is the Kaggle dataset of malicious and non-malicious URLs which consists of 1113 malicious and 1000 non-malicious URLs. The ELM algorithm is trained which invokes the classifier, the dataset is divided 80 percent into training and 20 percent into test, the user input is a URL where the 21 features are extracted from the URL and further the URL is classified as malicious or non-malicious.

V. RESULTS AND DISCUSSION

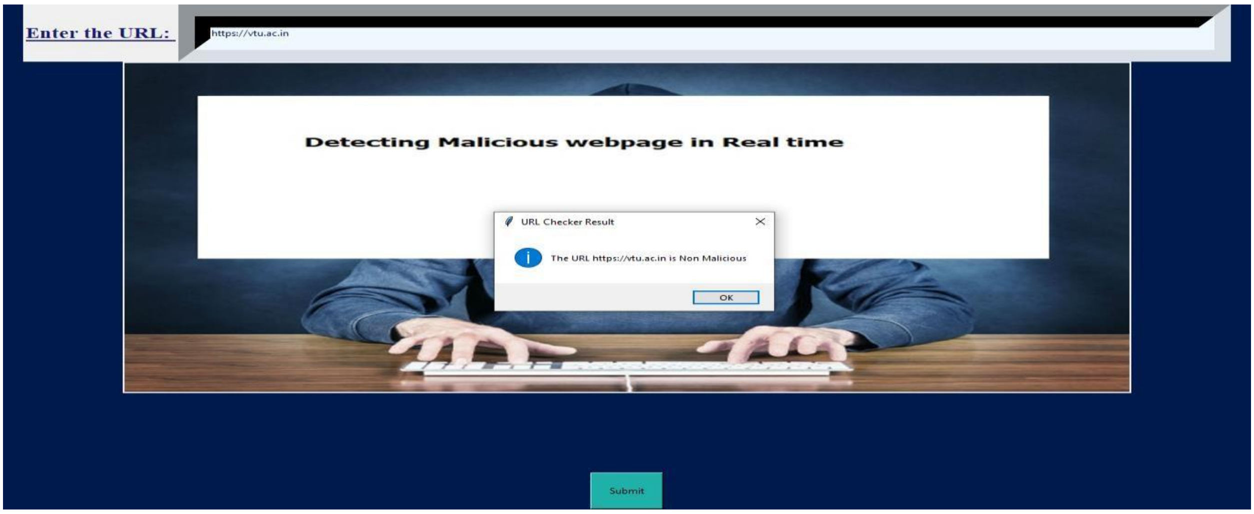
1. Taking the URL input from the user



Enter the URL: https://vtu.ac.in

Detecting Malicious webpage in Real time

2. The URL detected is non-malicious

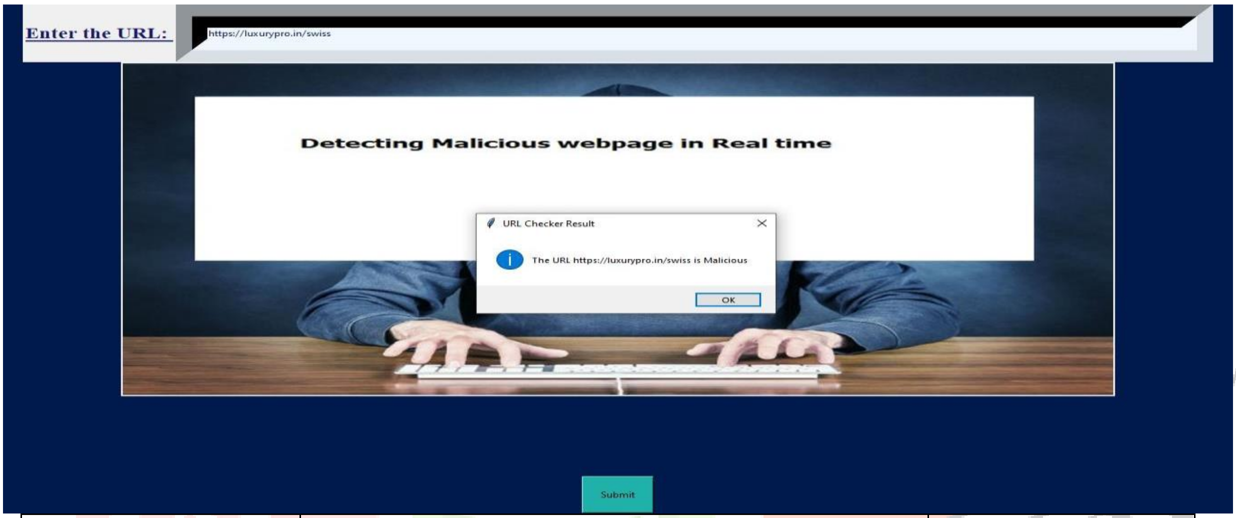


Detecting Malicious webpage in Real time

URL Checker Result

The URL https://vtu.ac.in is Non Malicious

3. The URL detected is malicious



Detecting Malicious webpage in Real time

URL Checker Result

The URL https://luxurypro.in/swiss is Malicious

Classes integrated	Tests done	Remarks
Feature extraction	Checking whether the features are extracted from URL input	Success
Classification	Class tested to check whether URL is malicious or not	Success

Here every module that includes the general framework is tried separately. Unit testing centers confirmation endeavors even in the littlest unit of programming plan in every module. This is otherwise called "Module Testing". The modules of the framework are tried independently. This testing is completed in the programming style itself. Unit testing practices particular ways in a module's control structure to guarantee finish scope and greatest blunder recognition. This test concentrates on every module exclusively, guaranteeing that it capacities legitimately as a unit. Subsequently, the naming is Unit Testing. In this progression every module is found to work acceptably as respect to the normal yield from the module. This testing is done to check for the individual piece codes for their working. It is done as such that when we do practical testing then the units which are a piece of these functionalities ought to have been tried for working.

VI. CONCLUSION

In this proposed methodology, we have implemented a malicious detection using machine learning concepts. A training data set of malicious URL and non-malicious URL is taken as training dataset. It is vectorized and then a ELM classifier is trained. Using the trained ELM classifier, the URL given by the user is classified as malicious or non-malicious. From the three analysis made we could conclude that the feature Token Count (TC) and Average token length (ATL) is not similar at each stage of analysis unlike other features, at each stage of analysis we do not consider all the features since there were similarities between them. Initially a total of 19 features were considered and in further stages of analysis it was reduced to 8 in the next analysis, and by the end of third analysis it was found only two features. A total of 300 samples were taken, 150 malicious and 150 non-malicious URLs, at each stage of analysis 50 malicious and 50 non-malicious samples were taken.

REFERENCES

[1] "Efficient prediction of phishing websites using supervised learning algorithms“, V. Santhana Lakshmi and M. Vijaya, Procedia Engineering, 30, pp.798-805, 2012.

[2] "A multi-tier phishing detection and filtering approach“, R. Islam and J. Abawajy, Journal of Network and Computer Applications, 36(1), pp.324-335, 2013.

[3] “Security analytics: big data analytics for cyber security: a review of trends, techniques and tools,” T. Mahmood and U. Afzal,” in Information assurance (ncia), 2013 2nd national conference on. Rawalpindi, Pakistan: IEEE, 2013, pp. 129–134.

[4] “Intelligent malware detection based on file relation graphs,” L. Chen, T. Li, M. Abdulhayoglu, and Y. Ye, (ICSC), 2015 IEEE International Conference on. Anaheim, California, USA: IEEE, 2015, pp. 85–92.

[5]”leveraging the visible attributes to classify the malicious short URLs“, R.k. Nepali and Y. Wang, 49th Hawaii International Conference on System Sciences (HICSS) IEEE, 2016, pp.2648-2655.

[6] “Using Supervised Machine Learning Algorithms to Detect suspicious URLs in online social networks”,Mohammed Al-Janabi, Ed de Quincey, Peter Andras, Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017,

[7] J. Shad and S. Sharma, “A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology,” pp. 425–430, 2018.

[8] Y. Sönmez, T. Tuncer, H. Gökal, and E. Avci, “Phishing web sites features classification based on extreme learning machine,” 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.

[9] T. Peng, I. Harris, and Y. Sawa, “Detecting Phishing Attacks Using Natural Language Processing and Machine Learning,” Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301, 2018.

[10] M. Karabatak and T. Mustafa, “Performance comparison of classifiers on reduced phishing website dataset,” 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.

[11] S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, “A New Method for Detection of Phishing Websites: URL Detection,” in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. 0, no. Icicct, pp. 949–952.

[12] K. Shima et al., “Classification of URL bitstreams using bag of bytes,” in 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), 2018, vol. 91, pp. 1–5.

[13] A. Vazhayil, R. Vinayakumar, and K. Soman, “Comparative Study of the Detection of Malicious URLs Using Shallow and Deep Networks,” in 2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018, 2018, pp. 1– 6.

[14] W. Fadheel, M. Abusharkh, and I. Abdel-Qader, “On Feature Selection for the Prediction of Phishing Websites,” 2017 IEEE 15th Intl Conf Dependable, Auton. Secur. Comput. 15th Intl Conf Pervasive Intell. Comput. 3rd Intl Conf Big Data Intell. Comput. Cyber Sci. Technol. Congr., pp. 871–876, 2017.

[15] X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng, “Boosting the Phishing Detection Performance by Semantic Analysis,” 2017.