IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MALICIOUS URL DETECTION IMPLEMENT MACHINE LEARNING

[1]Reshmi A.M, [2]PrameejaVimal

[1]Msc scholar, [2]Assistant Professor

[1,2]Department of computer science,

[1,2] St.joseph's college (Autonomous),Irinjalakuda,Thrissur,India

Abstract: The users accessing the internet was increased by day to day, therefore, there is also increase in various criminal activities. Everything online is your data secured? internet accessing consists of a wide range of crimes such as spam, phishing, theft the private information ,spreading of malwares and misrepresentation of data usage etc. it cause losses of billions of dollars every years. As a result malicious URL detection is of great interest now a day. The malicious uniform resource locator (URL) is the first mechanism for hosting unsolicited contents. Some of the malicious websites still escape detection of various web spam techniques. In this article, we proposed a malicious URL detection method using machine learning technique basis of our proposed URL behaviours and attributes. More over feature engineering and feature representation resources, that must be continuously reformed to handle variants of existing URL. This proposed URL increase the efficiency to detect malicious URL specifically. We includes a large set of data that containing various type of URLs and their relative labels of information.

Index Terms: malicious URL detection, feature engineering, machine learning,security of cyber concerns.

I. INTRODUCTION

The use of World Wide Web has increased day by day.Widely used applications to spread the malicious URLs. Malicious uniform resource locator (URL) is the firstly developed mechanism for hosting unsolicited contents.The growth and promotion of business spanning across many applications including online banking, ecommerce etc. Whenever unauthorized user visit the website through the URL, take them frauds that is, including the activities of malware installation and identity.From the statistics of the increasing in the number of the malicious URL distributions over the concecutive years, If there is a need to study and apply techniques. Regarding the problem of detecting malicious URLs there are two main trends based on signs or rule, malicious URL detection based on behaviour analysis techniques. The method of detecting malicious URL specifies the behaviour analysis adopted in machine learning. In this paper algorithms of machine learning used to classify the URLs based on their attributes.

In every year,malicious URLs causing billions of dollars losses of worth. In most recently, the widely used approach is applying domain knowledge to extract the lexical features of URL,it follows the machine learning models. The most commonly used engineering feature is Bag-of-words (BoW) and mostly used model of machine learning was support vector machine (SVM).although, machine learning based solution used instead of backlisting methodology,suffers from many challenges:

1) The conventional URL representation method fails to take the relation among characters and capture the sequential patterns.

2) Manual feature engineering specifes an extended domain knowledge in domain of cyber security.

To overcome the mentioned issues, this work proposes the various type of data sets used in how the models are generalizable.

People usually use machine learning method to analyse text and image information from the websites, malicious website use a variety of internet spam techniques. to solve this problem we first summarize the commonly used spam techniques for prevent the malicious website detection methods. For this instance, redirection spam and hidden Iframe spam provide false content to crawler, which provides to severe false negatives. A redirection spam gives the crawler a wrong content in webpages it automatically displays an unrelated page to the users. One of the best web spam redirection technique is JavaScript redirection. The content of hidden spam usually presents the legitimate contents. but it also contains invisible malicious information which cannot seen by browsers and users.

The rest of the sections are specified as follows. Section2 discuss the related work of malicious URL detection. Section3 discuss the deep learning techniques. Section4 provides information and overview about URL. Section5 discuss the malicious URL detection using techniques of machine learning. Section6 summarize the types of experimental results. Section7 finally, conclusions and are placed.

II. RELATED WORKS

We summarize the indicative studies in malicious URLs. We follow with a discussion of some of the patterns seen in the literature and research gaps. We accelerating the development of internet technology, malicious websites are widely spread and keep innovating as well. As machine learning techniques are advanced, more detection methods using machine learning techniques. Such as support vector machines, naive bayers,random forest etc. many approaches and classifications have been developed to detect malicious URLs and they can be categorized into many types: backlists, content based classification, URL based classification and feature engineering approach etc.

A. Backlist approach

Earliest internet filtering softwares are non machine learning based methods like backlisting. Backlist approach to detect phishing websites.Backlist are databases where the data of the URL that already confirmed as malicious are saved, and more URLs are added to the list over time. Whenever we visit the new URL,a database lookup is performed. If the URL present in the backlist, it consist of malicious and warning will be generated.Although it seems safe and effective method. But in case it is very slow, because they cannot keep up with the growing number of URL. That is database will never able to have all the malicious URLs that exit because new one created everyday and get around backlists. It identified by new phishing URL using heuristics and appropriate matching algorithms. Heuristics created new URLs by combining parts of known phished websites from available Backlist. Matching algorithms calculates the scores of the URLs.The evaluavation of the scores by matching various parts of the URL against the available URL on the backlist. These methods are not satisfying variant of existing URL. Later, machine learning algorithms are efficiently used to detect new types of malicious URL. These algorithms depends on the feature engineering approach to extracts features from URL.

B. feature engineering

Feature engineering requires extensive domain knowledge of URL in cyber security and a list of good features these are carefully chosen for feature selection. There are various types of features are widely used in the published works for malicious URL detection.These involves backlist features, lexical feature, host based features, content features etc.

Backlist features are specified through by checking its presence of URL from backlist.This serve as a strong feature in malicious URL. Lexical features are estimated through the string properties of the URL, because through the aspect of URL it should possible to identify it is malicious or not. The most commonly used features are length of the URL (components of URL are domains, sub domains), the number of special characters and each of them separated by special characters is considered as feature. Based on all the words in all the URLs a directory was built. The presence of word in URL, the value of the resource would be 1 or0.this model is also known as bag of words model. The whole bag of word resource approach seems form of backlist compatible with machine learning. Host based are estimated by the host properties of the URL. This allows us to find the location of the host. Content based features are obtained by downloading the URL page contents. Thus it being the dangerous type of feature, but it helps to prediction accuracy. We can get the HTML code of page and analyse the average number of words per line, link to remote scripts and invisible objects. Other features are related to JavaScript they used by hackers to encrypt malicious code. Moreover, malicious URL detection solution based feature engineering combines conventional machine learning can easily broken by an adversary.

II. DEEP LEARNING TECHNIQUES

Most recently more researchers start to adopt deep learning methods to detect malicious websites. Some research compares classical machine learning methods with deep learning techniques.URL lexical features and page content features are used for SVM, decision tree, naive bayes and artificial neural network (ANN) was the classification algorithm and it performs the best. Other method used for random split and time split to split the data for training, and it adopts for classical machine learning and deep learning techniques. the examples of classical machine learning involves SVM, logistic regression, naive bayes, KNN, decision tree, random forest, etc. and also deep learning includes the CNN, long short term memory models(LSTM)and CNN-LSTM) for generally deep learning methods performs better than traditional ones. In real internet world unbalanced positive and negative data cause the high level of difficulty in detection it also effects the accuracy of detection. In used convolution neural network (CNN) with character level embedding for detection of malicious URLs. This study shows unique deep learning architectures for different cyber security problems. In this work we describe the performance of various deep learning architectures for malicious URL detection.

IV. OVERVIEW OF UNIFORM RESOURCE LOCATOR(URL)

URLs ,also known as universal resource locator, as the name implies it used to find a particular resource in the internet, it also known as web address and it is the mechanism used by the browsers to retrieve any published resource on the web. A URL is nothing but, more than address of a given unique resource on the web. For example, an abstract of the location of the resources, this found in the system execute a great diversity of operations. In theory of each valid URL points to unique resource.Such resources can be HTML page, CSS document, an image etc. In reality there are some exceptions, the most common URL pointing to a resource that no longer has to be exists. As the resource represented by the URL and the URL itself are handled by the web server to carefully manage those resources and its associated URL. A URL composed of two parts, the first part defines the protocol type and the second part defines the IP address or domain name, third part defines path and its parameters to specific resource in the web.

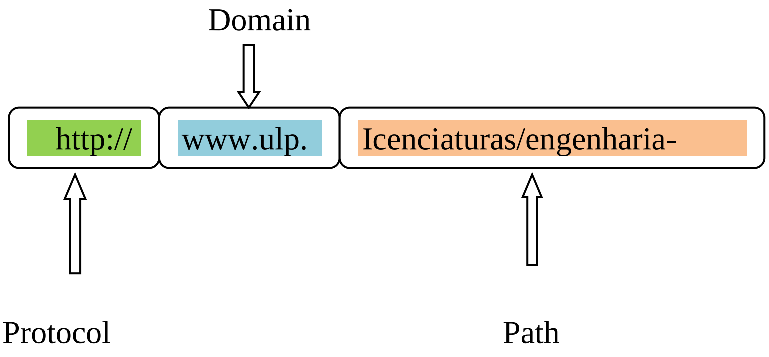

Fig4.1 representation of URL and consequently All part constitute an URL.

When a URL directs the browser to a file that can open, such as image or PDFs the browser displays the contents, we don't have to download the file, so there are many other types of files require a download. All the complexity and diversity of functions URL can made evil and attack the user.

V. MALICIOUS URL DETECTION USING MACHINE LEARNING

A. Model

This model presents the proposed malicious URL detection system using machine learning. This constrains consist of two stages: training and detection.

- Training stage: we will detect the malicious URLs it becomes necessary to collect both clean and malicious URLs. All the malicious and clean URLs are accurately labelled and proceeded attribute extraction. It is the basis for determining which URLs are malicious and which are clean. Finally, the sets of data are divided in to two subsets. Training machine learning algorithms are used training data, and testing data used for testing process. If the classification performance of machine learning model consists of high accuracy, the model will be used in phase of detection.
- Detection phase: In this phase performed on each output URL. Firstly, the URL will be going through extraction process of attributes. In next, these attributes are input to the classifier to classify whether, detect the URL is clean or malicious.

B. URL attribute extraction and selection

There are some main attribute groups for malicious URL detection as follows: lexical feature, host based features, content features etc.

Lexical feature: these features include URL length, main domain length, average path of length, average token length in domain, maximum token domain length.

Host based features: these are the features extracted from the characteristics of host of the URLs. This attributes indicate the malicious server location, identity of malicious servers and most of the host based features impacts,that contribute the URLs malicious level.

Content based features: these features are acquired when a whole web page was downloaded. But in the case of workload it consists of heavier than others. Since a lot of information needs to be extracted, and there may be security concerns about accessing the URL. The content based features of a website can be extracted primarily from its HTML content and the JavaScript will be used.

Above are the three main groups of attributes commonly used by the researchers to detect the malicious URL.Moreover, each study has its own decision on suitable characteristics and attributes for each experimental of the dataset. In this paper the use of three attribute groups is recommended. In each attribute group some new attributes and characteristics of the URL optimize to detect the malicious URL proposed.

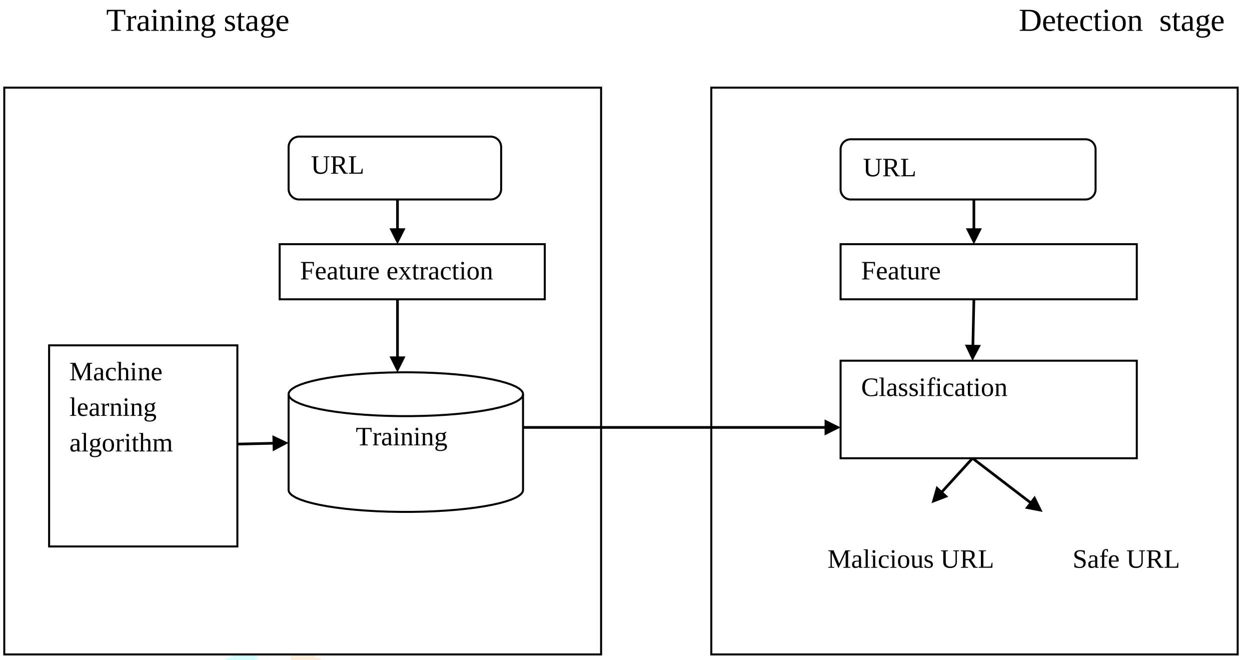

Table1.list of URL features in lexical feature group

NO	Feature group	Feature	Data type	Description
1	Lexical group	NumDots	Numeric	Number of character ‘.’ In URL
2		SubdomainLevel	Numeric	Number of sub domain levels
3		PathLevel	Numeric	The depth of URL
4		UrlLength	Numeric	The length of URL
5		NumDash	Numeric	Number of dash character ‘-‘
6		NumDashInHostname	Numeric	Number of dash character in the hostname
7		AtSymbol	Boolean	There exit a character ‘@’ in URL
8		TildeSymbol	Boolean	There exit a character ‘~’ in URL
9		NumUnderscore	Numeric	Number of the underscore character
10		NumPercent	Numeric	Number of the character %
11		NumqueryComponents	Numeric	Number of the query components
12		NumAmpersand	Numeric	Number of the character ‘&’
13		NumHash	Numeric	Number of the character’#’
14		NumNumericCharts	Numeric	Number of the numeric character
15		NoHttps	Boolean	Check if there exists HTTPS in URL of website
16		IpAddress	Boolean	Check if ip address used in the URL of website in hostname
17		DomainInSubdomains	Boolean	Check if TLD
18		DomainInPaths	Boolean	Check if TLD
19		HttpsInHostname	Boolean	Check if HTTPS is inordered in the hostname of website URL
20		HostnameLength	Numeric	Length of hostname
21		PathLength	Numeric	Length of the link path
22		QueryLength	Numeric	Length of the query
23		DoubleSlashInPath	Boolean	Here exit a slash ‘//’ in the link path
24		NumSensitiveWords	Numeric	Number of sensitive words in website
25		EmbeddedBrandName	Boolean	There exit a brand name in the

				domain
26		PctExtHyperlinks	Float	The percentage of external hyper links in the HTML source code of website

Table 2: List of URL feature in the host based feature group

NO	Feature group	Feature	datatype	Description
27	Host based feature group	PctExtResourceUrls	float	Percentage of URL external resources in HTML source code of website
28		ExtFavicon	boolean	Check if favicon is installed from a hostname that is different from the URL hostname of website
29		InsecureForms	boolean	Check if action of the form containing the content of URL without HTTPS protocol
30		RelativeFormAction	boolean	Check if the action form contains a relative URL
31		ExtFormAction	boolean	Check if action form contains external URL
32		AbnormalFormAction	boolean	Check if action form containing abnormal URL
33		PctNullSelfRedirectHyperlinks	float	Percentage of hyperlinks containing null value
34		FrequentDomainNameMismatch	boolean	Check if more frequent hostname in HTML source code
35		FakeLinkInSatusBar	boolean	Check if the html source code contains javascript command turn of right click of mouse
36		RightClickDisabled	boolean	Check if the HTML source code contains a JavaScript command start a popup window
37		PopupWindow	boolean	Check if the html source code contains “mailto”in the html
38		SubmitInfoToEmail	boolean	Check if the frame is used in the html source codes
39		IFrameorFrame	boolean	Check if the title tag is empty
40		MissingTitle	boolean	Check if the title of tag is null in html source codes
41		Src_ eval_cnt	int	Number of function eval()HTML source codes
42		Src_escape_cnt	int	Number of function escape()HTML source codes
43		Src_exec_cnt	int	Number of function exec()HTML source codes
44		Src_search_cnt	int	Number of function search()HTML source

				codes
45		ImagesOnlyInform	boolean	Check if the action in the form of HTML source code does not contain text,but only images
46		Rank_Country	boolean	Current rank of website uRL is in top 1 million of alexa
47		Rank_host	boolean	The rank of host website URL in topest 1 million of alexa
48		AgeDomain	int	The age of domain since it is registered

Table 3.List of URL features in correlated feature group

NO	Feature group	Feature	Data type	Description
49	Correlated feature group	UrlLengthRT*	-1,0,1	Correlation length of URL
50		PctExtResourceUrlsRT*	-1,0,1	Correlation percentage of external URL
51		AbnormalExtFormActionR*	-1,0,1	Correlation abnormal actions in form
52		ExtMetaScriptLinkRT*	-1,0,1	Correlation meta script link
53		SubDomainLevelRT*	-1,0,1	Correlation sub domain level
54		PctExtNullSelfRedirectHyperlinksRT*	-1,0,1	Correlation null self redirect hyperlinks

VI. EXPERIMENTAL RESULTS

A. Dataset and experiment environments

1) Experiment Dataset: experimental dataset for malicious URL detection includes; 470.000 URLs collected of which about 70.000 URLs are malicious and others are safe. All of these URLs are checked by virus Total tool to verify each URLs label.CSV format is used to store the complete dataset.

2) Experimental setup: both safe and malicious URLs of a dataset are divided into two subsets; for training and testing. The experiment is repeated with includes both SVM and RF algorithm.

B. Results and discussions

1) Evaluavation metrics: accuracy is the percentage of correct decisions among all testing samples.

Acc = (TP+TN) / (TP+TN+FP+FN)*100%

Here, TP=true positive number of malicious URLs correctly labels.

FN=false negative is the number of malicious URLs misclassified as safe.

TN=true negative number of the safe URL correctly labeled.

FP=false positive number of safe URLs misclassified as malicious.

Precision: it is the percentage of malicious URLs correctly labeled (FP) among all malicious URLs labelled by the classifier (TP+FP).

Precision =TP / (TP+FP) *100%

Recall=TP/ (TP+FN) *100%

F1 score = (2*precision*recall) / (precision + recall)

FRP is the false prediction rate it is calculated as:

FRP=FP/ (FP+TN) *100%

Training results: in this work additional small testing data set with 107 safe URLs and 118 malicious URLs are used to improve performance of best machine learning algorithm discussed above.

Table 4.confusion matrix

	Classified malicious URLs	Classified safe URLs
Real malicious URLs	TP	FN
Real safe URLs	FP	TN

Table5. Training performance of malicious URL detection system

Dataset	Algorithm and parameters	Accuracy (%)	Precision (%)	Recall (%)	Training time(s)	Testing time(s)
10.000 URLs	SVM(10Iterations)	93.39	94.67	92.51	2.32	0.01
	SVM(100Iterations)	93.35	94.84	92.71	3.11	0.01
	RF(10 trees)	99.10	98.43	97.45	2.78	0.01
	RF(100 trees)	99.77	98.75	97.85	3.34	0.01
470.000 URLs	SVM(100Iterations)	90.70	93.43	88.45	272.97	2.12
	SVM(10Iterations)	91.07	93.75	88.85	280.33	2.31
	RF(10 trees)	95.45	90.21	95.12	372.97	2.02
	RF(100 trees)	96.28	91.44	94.42	480.33	2.30

Table6. Testing results

	Predicted safe URL	Predicted malicious URL
Real safe URL(107)	96	11
Real Malicious URL(118)	9	109

II. CONCLUSIONS

In this paper, we discussed machine learning and deep learning models to detect and classifying malicious URLs. Those are wishing to incorporate machine learning technique to improve existing ones. Additionally, this study reports the specific features contained within URL may used to minimize the overhead of cost of model. The empirical results in table show the efficiency of the proposed extracted attributes. Here we use the combination between easy to calculate attributes and big data processing technologies. These factors are ensures the processing time and accuracy of the system. Finally the results of these research implemented as information security technologies in information security systems.

REFERENCES

[1]Vinayakumar R, Sriram S, Soman KP, and Mamoun Alazab, Senior Fellow,IEEE. Malicious URL detec
ction using deep learning.

[2] Dongjie liu and jong hyouk lee (senior member,IEEE).CNN based malicious website detection by invalid
ating multiple webspams.Received may3,2020,accepted may11,2020,date of publication may 18,2020,da
te of current versionjune4,2020.digital object identifier 10.1109/ACCESS.2020.2995157. IEEE access.m
ultidisciplanary. Rapid review,open access journal.

[3]shreyasa,irLabhsetwar,Tushar B.kute(student department of engineering ,S.P.P.U.university,nashik,india,
Researcher,MITU skillologies,pune,india.malicious URLs detection using machine learning.august 2019|
IJIRT|volume 6 issue 3|ISSN:2439-6002.international journal of innovative research in technology.

[4]Clayton Johnson1, Bishal Khadka1, Ram B. Basnet1*and tenzindoleck2,lcolorado mesa university, gra
nd junction,CO 81501,USA{cpjohnson,bkhadka}@mavs.coloradomesa.edu,rbasnet@coloradomesa.ed
u, 2 university of southern California.los angels,30,2020;accepted December 11,2020;published:Decem
ber 31,2020.

[5]saurabh Mittal,preeti associate professor,research scholar.department of computer science and engineerin
g, galaxy global gro Up of institutions,dinarpur,am Bala.haryana,india. Mechine learning approach for cl
assifying URLs.2016 IJEDR|Volume4|ISSN : 2321-9939.international journal of engineering Developm
ent and research.

[6] Learning Cho Do Xuanl, Hoa Dinh Nguyen1. Information Security dept, Posts and Telecommunications
Institute of Technology Hanoi,Vietnam , Information Assurance dept, FPT University, Hanoi, Vietnaml.
Tisenko victor nikolaevich3 systems of automatic Design Peter the great st. Petersburg Polytechnic u
niversity Russia,st. Petersburg Polytechnicheskaya, 29. Malicious URL Detection based on Machine lear
ning.(IJACSA)international journal of advanced computer science and applicatio ns,volume. 11 no.1,20
20. www.ijacsa.thesai.org

[7]A whoisXML API User Success Story.Malicious URL detection via machine Lesrning.

[8]Marcelo Ferreira,lusofona university of porto portugalcferreira_marcelo@outlook pt malicious URL det
ection using machine learning algorithms procedings of the digital privacy and security conference 2019.