MALICIOUSURLDETECTIONUSINGMACHINELEARNINGALGORITHMS

Dr.G.Sai Chaitanya Kumar, Associate Professor, , Department of Computer Science and Engineering, DVR & Dr HS MIC College of Technology,Kanchikacherla, Andhra Pradesh, India
Department of Computer Science and Engineering.
G. Pratyusha, Assistant Professor, , Department of Computer Science and Engineering, DVR & Dr HS MIC College of Technology,Kanchikacherla, Andhra Pradesh, India
Department of Computer Science and Engineering.
A .Navya, SK.Faiz, B.VinayKumar, P.Ranjitha,UG Student, Department of Computer Science and Engineering, DVR & Dr HS MIC College of Technology,Kanchikacherla, Andhra Pradesh, India
Department of Computer Science and Engineering.

Abstract—Currently,theriskofnetworkinformationinsecurityisincreasingrapidlyinnumber and level of danger. The methods mostly used by hackers to day is to attack end to end technology and exploit human vulnerabilities. These techniques include social engineering, phishing, pharming, etc. One ofthestepsinconducting these attacks is to deceive users with malicious Uniform Resource Locators (URLs). As a results, malicious URL detection is of great interest now a days. There have been several scientific studies showing a number of methods to detect malicious URLs based on machine learning and deep learning techniques. In this paper, we propose a malicious URL detection method usingmachinelearningtechniquesbasedonourproposedURLbehaviorsandattributes.Moreover,bigdatatechnologyisalsoexploitedtoimprovethecapability of detection malicious URLs based on abnormal behaviors. In short, the proposed detection system consists of a new set of URLs features and behaviors, a machine learning algorithm, and a big data technology. The experimental results show that the proposed URL attributes and behavior can help improve the ability to detect malicious URL significantly. This is suggested tha the proposed system may be considered as an optimized and friendly used solution formalicious URL detection.

Keywords—URL;malicious URL detection; feature extraction; feature selection; Machine learning

1. INTRODUCTION

Uniform Resource Locator (URL) is used torefer to resources on the Internet. In [1], Sahoo etal.presentedaboutthecharacteristicsandtwobasiccomponentsoftheURLas:protocolidentifier, which indicates what protocol to use, and resource name, which specifies the IP addressor the domain name where the resource is located. It can be seen that each URL has a specific structure and format. Attackers often try to change one or more components of the URL's structure to deceive users for spreading their malicious URL. Malicious URLs are known as links that adversely affect users. These URLs will redirect users to resources or pages on which attackers can execute codes on users' computers, redirect users to unwanted sites, malicious website, or other phishing site, or malware download. Malicious URLs can also be hidden in download links that are deemed safe and can spread quickly through file and message sharing in shared networks. Some attack techniques that use malicious URLs include [2, 3, 4]: Drive-by Download, Phishing and Social Engineering, and Spam.

According to statistics presented in [5], in 2019, the attacks using spreading malicious URL technique are ranked first among the 10 most common attack techniques. Especially, according to this statistic, the three main URL spreading techniques, which are malicious URLs, URLs, there are two main trends at present as malicious URL detection based on signs or sets of rules, and malicious URL detection based on behavior analysis techniques[1,2].The method of detecting
malicious URLs based on a set of markers or rules can quickly and accurately detect malicious URLs.

 DOI: 10.36893.DRSR.2023.V13I04.019-0024

However, this method is not capable of detecting new malicious URLs that are not in the set of predefined signs or rules. In our research, machine learning algorithms are used to classify URLMachine learning algorithms are a part of the whole malicious URL detection system. Two supervised machine learning algorithms are used, Support vector machine (SVM)and Random forest(RF).

The paper is organized as follows. Section II reviews some recent works in the literature on malicious URLdetection.TheproposedmaliciousURLsdetectionsystemusingmachinelearningispresentedinSectio n

III. In this section, the new features for URLs detection process are also described in details. Experimental results and discussions are provided in Section IV. The paper is concluded by Section V.

2. RELATEDWORK

2.1 Signature based Malicious URL Detection

Studies on malicious URL detection using the signature sets had been investigated and applied long time ago[6,7,8].Most of these studies often use lists of known malicious URLs.; otherwise URLs will be considered as safe. The main disadvantage of this approach is that it will be very difficult to detect new malicious URLs that are notinthe given list.

2.2 Machine Learning based Malicious URL Detection

There are three types of machine learning algorithms that can be applied on malicious URL detection methods, including supervised learning, unsupervised learning, and semi supervised learning. And the detection method sare based on URL behaviors.

The behaviors and characteristics of URLs can be divided into two main groups, static and dynamic. In their studies [9, 10, 11] authors presented methods of analyzing and extracting static behavior of URLs, including Lexical, Content, Host, and Popularity-based. The machine learning algorithms used in these studies are Online Learning algorithms and SVM. Malicious URL detection using dynamic actions of URLs is presented in [12, 13]. In this paper, URL attributes are extracted based on both static and dynamic behaviors. Some attribute group sare investigated, including Character and semantic groups; Abnormal group in websites and Host-based group; Correlated group.

2.3 MaliciousURLDetectionTools

- URLVoid:URLVoidisaURLcheckingprogram using multiple engines and blacklists ofdomains.SomeexamplesofURLVoidareGoogle Safe Browsing, Norton Safe Web andMy WOT. The advantage of the Void URL toolis its compatibility with many different browsersaswellasitcansupportmanyothertestingservices.ThemaindisadvantageoftheVoidURL tool is that the malicious URL detection process relies heavily on a given set of signatures.
- Dr.WebAnti-VirusLinkChecker:Dr.WebAntiVirusLinkCheckerisanadd-onforChrome, Firefox, Opera, and IE to automatically find and scan malicious content on a downloadlinkonallsocialnetworkinglinkssuchasFacebook,Vk.com, Google+.

ComodoSiteInspector: This is amalwareand security hole detection tool. This helps users checkURLsorenableswebmasterstosetupdaily checksby

- downloadingallthespecifiedsites.andrunthemina sandbox browserenvironment.
- Someothertools:Amongaforementionedtypicaltools,therearesomeotherURLcheckingtools, such as UnShorten.it, Virus Total, NortonSafeWeb,

Site Advisor (by McAfee), SucuriBrowserDefender, Online Link Scan, and Google SafeBrowsingDiagnostic.

From the analysis and evaluation of malicious

 DOI: 10.36893.DRSR.2023.V13I04.019-0024

URLdetectiontoolspresentedabove,itisfoundthatthemajority of current malicious URL detection tools aresignature-based URL detection systems. Therefore, theeffectivenessofthesetoolsis limited.

3. ProposedMethod

3.1 TheModel

Fig.1presentstheproposedmaliciousURLdetectionsystemusingmachinelearning.ThemaliciousURLdetection model using machine learning contains twostages:trainingand detection.

- Training stage: To detect malicious URLs, it isnecessarytocollectboth maliciousURLsandclean URLs. Then, all the malicious and cleanURLs are
- correctlylabeledandproceededtoattributeextraction. These attributes will be the best basisfor determining which URLs are clean and whichare malicious. Details of these attributes will bepresented in details in thispaper. Finally, thisdatasetisdividedinto2subsets:trainingdatausedfortrainingmachinelearningalgorithms,andtestingdatausedfortestingprocess.Iftheclassificationperformanceofthemachinelearningmodelisgood(highclassificationaccuracy), the model will be used in the detectionphase.
- Detectionphase:ThedetectionphaseisperformedoneachinputURL.First,theURLwillgothroughattributeextractionprocess.Next,theseattributesareinputtotheclassifiertoclassifywhethertheURL iscleanormalicious.

3.2. URLAttributeExtractionandSelection

In [1], the authors listed some main attribute groupsformalicious URL detectionas follows.

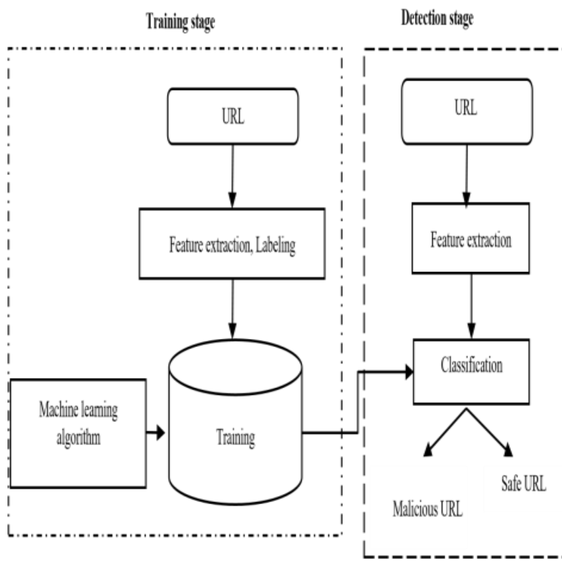

Fig. 1. Malicious URL Detection Model using Machine Learning.

Lexical features: these features include URL length,main domain length, maximum token domain length,path average length, average token length in domain.Host-based Features: these features are extracted fromthe host characteristics of the URLs. These attributesindicatethelocationofmaliciousservers,theidentityofmalicious servers, the degree of impact of several host-basedfeaturesthatcontributetheURL'smaliciouslevel.

Above are the three main attribute groups commonly used by researchers todetectmaliciousURLs.However,eachstudyhasitsowndecisiononsuitableattributesandcharacteristics for each particular experimental dataset..However, in each attribute group some new attributesand characteristics of the URL to optimize the ability todetect malicious URLs are proposed. The new attributesformaliciousURLdetectioninthisresearcharelistedinTablesI, II, and III

 DOI: 10.36893.DRSR.2023.V13I04.019-0024

no	Featuregroup	Feature	Datatype
1	**Lexicalgroup**	Num Dots	numeric
2		Subdomainlevel	numeric
3		Pathlevel	numeric
4		Urllength	numeric
5	**Host-basedfeaturegroup**	PctExtResourceUrls*	float
6		ExtFavicon*	boolean
7		InsecureForms*	boolean

Table1:ListofUrl

Allattributesmarked“*”inTablesI,II,IIIarenewlyextracted and selected in this research. Besides, inpreviousresearches,authorstendtousefeatureextraction and selection method based on a group ofpredefined features. However, those recommendedfeatures are specialized and not popular. As a results,it is usually difficult to implement those features inotherworks,andtore-evaluatethedetectionperformance of those features. In this work, we try tocombinebasic featurestoformulatenew ones.

3.3. MachineLearningAlgorithmSelection

The application of machine learning algorithms indetectingmaliciousURLshasbeenstudiedandappliedwidely[1].Inthispaper,twocommonlyusedsupervisedmachinelearningalgorithms,RFandSVM[15, 21], areused.

In this research, machine learning algorithms arethe last puzzle to complete our proposed maliciousURL detection system. Those algorithms are suitabletoutilizedtheusefulnessofournewfeaturesselectedfor malicious URL detection. The machine learningalgorithmsarealreadywellinvestigatedintheliterature. In this work, SVM and RF are selected asan example to illustrate the good performance of thewhole detection system, and are not our main focus.Readersareencouragedtoimplementsomeotheralgorithms such as Naïve Bayes, Decision trees, k-nearestneighbors, neural networks,etc.

Inordertoexploretheeffectivenessofusingthesetwo algorithms, different adjustments of parametersareimplemented.

3.4Random ForestAlgorithm

Step-1: SelectrandomKdatapointsfromthetrainingset.
Step-2: Build the decision trees associated with theselecteddatapoints (Subsets).
Step-3: Choose the number N for decision trees thatyouwant to build.
Step-4:Repeatsteps1&2.
Step-5: For new data points find the predictions ofeach decision tree and assign the new data points tothecategorythat wins themajorityvotes.

4. ResultsandDiscussion

4.4DatasetandExperimentEnvironments

1) *Experiment dataset:* The experimental datasetformaliciousURLdetectionmodelincludes:470.000URLscollectedfrom[16,17,22,23],ofwhichabout 70.000 URLs are malicious and 400.000 URLs aresafe. All these URLs are checked by Virus Total tooltoverifythelabelsofeachURL.Thecompletedatasetisstored using CSVformat. Each URL

2) *Experimentalsetup:*Thedatasetofbothsafeandmalicious URLs mentioned above is divided into

 DOI: 10.36893.DRSR.2023.V13I04.019-0024

2subsets. About 80% of the dataset, 470.000 URLs(400.000 safe URLs, 70.000 malicious URL), is usedfortraining,andabout20%ofthedataset,about
10.00 URLs(5.000maliciousURLs,5.000safeURLs),isusedfortesting.Theexperimentisrepeatedmanytimeswithboth SVMandRFalgorithm.Differentparametersettingsareusedindifferentruns.

3) Experimentdataset

- Setup environment: Python version 3.6; Sparkversion2.3.0; Hadoop version 2.7; Java (JDK)8; Ubuntu 18.04.
- Hardware:RAM16GB;Intel(R)Xeon(R)CPUE52640v3 @ 2.60GHz.

4.2. Evaluation

1) *Evaluation metrics:* Accuracy: the percentageofcorrectdecisions amongall testingsamples*acc**_______________*TPT***N**%

*TP***TN***FP***FN* (1) where:*TP*- Truepositive is the number of malicious URLs correctlylabeled;*FN*-FalsenegativeisthenumberofmaliciousURLsmisclassifiedassafe;*TN*-TruenegativeisthenumberofsafeURLcorrectlylabeled;*FP*-FalsepositiveisthenumberofsafeURLsmisclassifiedas malicious.

Confusionmatrix:isatwo-wayTableIVrepresentinghowmany samplesareclassified into which labelaccordingly.

Precision: is the percentage of malicious URLscorrectlylabeled(TP)amongallmaliciousURLslabeledby theclassifier (TP+FP).

TP

*Precision**Re*call*

FPR(Falsepredictionrate)iscalculatedas:

FP

*FRP** *100%

*FP***TN*

2) Results

- Trainingperformance

Toevaluatethetrainingperformanceofthemachinelearningalgorithm,bothtwodatasubsetsareusedindividually.Eachofthesedatasubsetshasdifferent data size as well as different distribution ofdata labels, which may result in different trainingperformances.Theresultsarepresentedin TableV. Experimental results show that the RF with 100trees gives the best predictive result. In return, thetrainingtimeoftheRFisslightlylongerthanSVM,butthetesting timeisnot muchdifferent. The accuracyoftheseconddatasetisreducedduetotheunbalance between safe and malicious URLs of thedata. As expected, RF algorithm, with its fast speedand high accuracy, is very suitable for classificationproblem.Besides,inourresearch,whenmachinelearningalgorithmsarecombinedwithsparklibraries,thetrainingandtestingtimecanbereducedsignificantly.SparkMLMachineLearningisandsupportsmany machine learning.

5.Conclusion :

Thispaperpresentsamachinelearning-basedsolution for malicious URL detection. The empiricalfindings in Tables V and VI have demonstrated theefficacy of the extracted characteristics. Unlike manyothertraditionalarticles,wedon'tusespecialqualitiesin this study or try to build enormous datasets toincrease the accuracy of the system. The processingspeed and accuracy of the system are determined bythecombinationofsimple-to-calculatequalitiesand largedataprocessingtechnologiestoensurethebalance of the two elements. The findings of this study can be used and put intopractiseininformationsecuritytechnologiesandsystems.Afreeprogrammeto identify fraudulent URLs

 DOI: 10.36893.DRSR.2023.V13I04.019-0024

on websites has beencreated[20] on the findingsof this paper.

6.References:

[1] D. Sahoo, C. Liu, S.C.H. Hoi, "Malicious URLDetectionusing Machine Learning: ASurve [2]y". CoRR,abs/1701.07179, 2017.

M.Khonji,Y.Iraqi,andA.Jones,"Phishingdetection:aliteraturesurvey,"IEEECommunicationsSurveys & Tutorials, vol. 15, no. 4, pp. 2091–2121,2013.M. Cova, C. Kruegel, and G. Vigna, "Detectionandanalysisofdriveby-downloadattacksandmaliciousjavascriptcode,"inProceedingsofthe19thinternational conference on World wide web. ACM,2010, pp. 281– 290.R. Heartfield and G. Loukas, "A taxonomy ofattacksandasurveyofdefencemechanismsforsemanticsocialengineeringattacks,"ACMComputing Surveys (CSUR), vol. 48, no. 3, p. 37,2015.InternetSecurityThreatReport(ISTR) 2019–Symantec.https://www.symantec.com/content/dam/symantec/docs/re ports/istr-242019-en.pdf[Last accessed10/2019].S.Sheng,B.Wardman,G.Warner,L.F.Cranor, J.Hong, and C. Zhang, "An empirical analysis ofphishingblacklists,"inProceedingsofSixthConferenceonEmailandAnti-Spam(CEAS),2009.

C.Seifert,I.Welch,andP.Komisarczuk,"Identification of malicious web pages with staticheuristics,"inTelecommunicationNetworksandApplicationsConference,2008.ATNAC2008.Australasian.IEEE,2008, pp. 91–96.

S. Sinha, M. Bailey, and F.Jahanian, "Shades ofgrey:Ontheeffectivenessofreputation-based"blacklists"," in Malicious and Unwanted Software,2008.MALWARE2008.3rdInternationalConferenceon.IEEE,2008, pp. 57–64.

J.Ma,L.K.Saul,S.Savage,andG.M.Voelker,"Identifying suspicious urls: an application of large-scale online learning," in Proceedings of the 26thAnnualInternationalConferenceonMachineLearning.ACM, 2009, pp. 681–688.

B. Eshete, A. Villafiorita, and K. Weldemariam,"Binspect:Holisticanalysisanddetectionofmalicious web pages,"in Security and PrivacyinCommunicationNetworks.Springer,2013,pp.149–166.S.Purkait,"Phishingcountermeasuresandtheireffectiveness–literaturereview,"InformationManagement&ComputerSecurity,vol.20,no.5,pp.382–420, 2012.

Y. Tao, "Suspicious url and device detection bylog mining," Ph.D. dissertation, Applied Sciences:Schoolof Computing Science, 2014.

G. Canfora, E. Medvet, F. Mercaldo, and C. A.Visaggio, "Detection of malicious web pages usingsystem calls sequences," in Availability, Reliability,andSecurityinInformationSystems.Springer,2014,pp. 226–238.

LeoBreiman.:RandomForests.MachineLearning45 (1), pp.5-32, (2001).

ThomasG.Dietterich.EnsembleMethodsinMachineLearning.InternationalWorkshoponMultiple Classifier Systems, pp 1-15, Cagliari, Italy,2000.

Developer Information.https://www.phishtank.com/developer_info.php.[Lastaccessed 11/2019].

URLhaus Database Dump.https://urlhaus.abuse.ch/downloads/csv/.[Ngàytruynhập11/2019].

DatasetURL.http://downloads.majestic.com/majestic_million.csv.[Lastaccessed 10/2019].

Malicious_n_Non-MaliciousURL.https://www.kaggle.com/antonyj453/urldataset#data.csv.[Lastaccessed11/2019].

chrome.zip.https://drive.google.com/file/d/13G_Ndr4hMFx_qWyTEjHuOyJmHFWD0Gud/view?fbclid=IwAR0SLVCrvjHHGmoHZH97nXN3BmDMY7jG4SOsKZYLAZjTFgeoJADfli64-g. [Last accessed12/2019].

[21] G.Sai Chaitanya Kumar, Dr.Reddi Kiran Kumar, Dr.G.Apparao Naidu, "Noise Removal in Microarray Images using Variational Mode Decomposition Technique " Telecommunication computing Electronics and Control ISSN 1693-6930 Volume 15, Number 4 (2017), pp. 1750-1756

[22] G. S. C. Kumar, D. Prasad, V. S. Rao and N. R. Sai, "Utilization of Nominal Group Technique for Cloud Computing Risk Assessment and Evaluation in Healthcare," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 927-934, doi: 10.1109/ICIRCA51532.2021.9544895

[23] V. S. Rao, V. Mounika, N. R. Sai and G. S. C. Kumar, "Usage of Saliency Prior Maps for Detection of Salient Object Features," *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2021, pp. 819-825, doi: 10.1109/I-SMAC52330.2021.9640684