International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 03, 2020
ISSN: 1475-7192

Malicious URL Detector for low power consumption using Blacklist

[1]Nandhini S, [2]Nanda Kishore, [3]Josephus Andrew, [4]Shubham Godara

Abstract— *There have been many attacks on websites and information has been hacked using using viruses and Trojans through the URL or Universal Resource Locator of the website. This problem can be controlled by using a program called a URL detector that uses a blacklist which basically denies access to the website when malicious activity is detected. A convoluted neural network is used to deny access to the URL if it does not match with the existing URLs in the blacklist. This method of scanning for malicious URLs is efficient and power saving in most aspects.*

Keywords—*URL, convoluted, Trojans, blacklist.*

I. INTRODUCTION

The ceaseless improvement of web attack has caused a lot of problems for users over the years because of the hacking of valuable information. The security identification of URLs has reliably been the point of convergence of Web security. Many web application resources can be gotten to by essentially entering a URL or clicking an association in the program. An assailant can lead different web assaults by utilizing Structured Query Language (SQL) and Cross- webpage scripting(XSS) code to be inserted in the source code of the site. Accordingly, it is essential to improve the trustworthiness and security of web applications by correctly perceiving pernicious URLs.This project is designed to reduce the URL attacks on the websites gradually by denying access as it detects the malicious code in the URL. A blacklist is a list of existing malicious URLs in the form of a dataset. When a URL is to be scanned it is cross referenced with the blacklist to check if the URL exists. If it does then the URL is denied access. If it does not exist on the blacklist then it is sent through a convoluted neural network that separates the malicious code from the original URL using a Gated Recurrent Unit as a pooling layer. This system is efficient in detection, low power - consumption and accuracy.

II. RELATED WORK

The machine learning method is useful to detectmalicious URLs but manually detection of URLs is time-consuming and requires constant adjustments according to the code and it requires steady alteration of highlights to oblige changes in URLs to blend in with human information; this restricts the precision of the characterization model somewhat Of late, AI has been applied to interruption revelation and sur-passed customary acknowledgment

[1] Assistant Professor ,SRM Institute of Science and Technology, Ramapuram
[2] Assistant Professor ,SRM Institute of Science and Technology, Ramapuram
[3] Assistant Professor ,SRM Institute of Science and Technology, Ramapuram
[4] Assistant Professor ,SRM Institute of Science and Technology, Ramapuram

Received: 03 Dec 2019 | Revised: 21 Jan 2020 | Accepted: 18 Feb 2020

techniques. The different models for URL recognition are discovered that intermittent neural systems don't require truly made highlights, and its location results are superior to irregular discovery strategies. Saxe's work is firmly identified with our exploration on character-level URL implanting. This paper depicts how to distinguish vindictive URLs, record ways and utilizations character level installing and a convolutional neural system to separate quality and arrangement. It proposed the neural framework and differentiated it and various models that concentrate incorporates genuinely. There have been a ton of improvement in AI utilizing convolutional systems to recognize malevolent code in pages to forestall hacking by utilizing implanted pernicious code that is inserted in the website page. There are methods of identifying the code in the order module and calculations. It is basic to improve the steadiness and security of web applications by applying AI classifiers on static highlights.

III. MODULES

Module 1: Data Evaluation

Data Evaluation process looks for the unexpected in an information driven way, it is crucial that these tools are used by professionals can chose any tools according to the to the each stage of analysis based on the information. Barely any systems have composed all of these limits either fundamentally or at the UI level. Appearance's information driven approach grants coordination among various application UIs. It uses a plan that screens the mapping of visual articles to information in shared databases.

It will ordinarily produce handfuls, if not hundreds, of exploratory diagrams throughout investigating a dataset. Of these charts, you may wind up distributing a couple in a build up an individual comprehension of the information, so the entirety of your code and diagrams ought to be equipped towards that reason. Important details that you might add if you were to publish a graph2 are not necessary in an exploratory graph.

Visual information analysis techniques being used in a wide range can be followed back to numerous hundreds of years prior, it is on the grounds that that human eyes and brains have solid structural ability to identify such significant situation in data exploring.

Module 2: Feature Extraction

Feature extraction and selection methods are utilized secluded or in blend with the intend to improve execution, for example, evaluated exactness, representation of learned knowledge. For the most part, features can be ordered as: pertinent, insignificant, or excess. In feature selection procedure a subset from accessible feature information are chosen for the way toward learning algorithm. The best subset is the one with least number of measurements that most add to learning precision.

Information gain of a feature is estimation the distinction of entropy whether it shows up in the content. The bigger data gain, the more prominent commitment the qualities to the content. Qualities with high data increase will be chosen as feature.

Dimensionality decrease is wide spread preprocessing in high dimensional information analysis, perception and demonstrating. Perhaps the least difficult approaches to decrease dimensionality is by Feature Selection; one chooses just those information measurements that contain the pertinent data for taking care of the specific issue.

Some methods like Filter methods are used for high dimensional data sets because of their low cost and efficiency. Hybrid/embedded techniques are as of recently created which use focal points of the two filters and

wrappers approaches. A hybrid approach utilizes both a free test and execution assessment capacity of the component subset. Filters strategies can be additionally classified into two gropus, specifically feature weighting algorithm and subset search algorithm. Feature weighting algorithm helps to feature independently and rank them dependent on their importance to the objective idea.

Module 3: Prediction

Neural systems are multi-layer systems of neurons that are used to compute data. they can also have multiple neurons and multiple connects in between them if required and this complex set of neural connections is used to form complicated relationships to interpret data.

In a neural system, changing the weight of any one association (or the bias of a neuron) has a resounding impact over the various neurons and their initiations in the consequent layers.

One component that is very important is the Long Short- Term Memory Network or LSTM neural framework that is set up by dim proliferation. Through Time and overcomes the vanishing slope issue. Taking everything into account, it might be used to make immense intermittent frameworks that thus can be used to address irksome progression issues in AI and achieve top tier results. Instead of neurons, LSTM frameworks have memory blocks that are related through layers.

A square has sections that make it more splendid than an old style neuron and a memory for late progressions. A square contains doors that manage the square's state and yield. A square works upon an information succession and each entryway inside a square uses the sigmoid incitationnits to control whether they are actuated or not, revealing the improvement of state and expansion of data moving through the square restrictive.

IV. PROPOSED SYSTEM

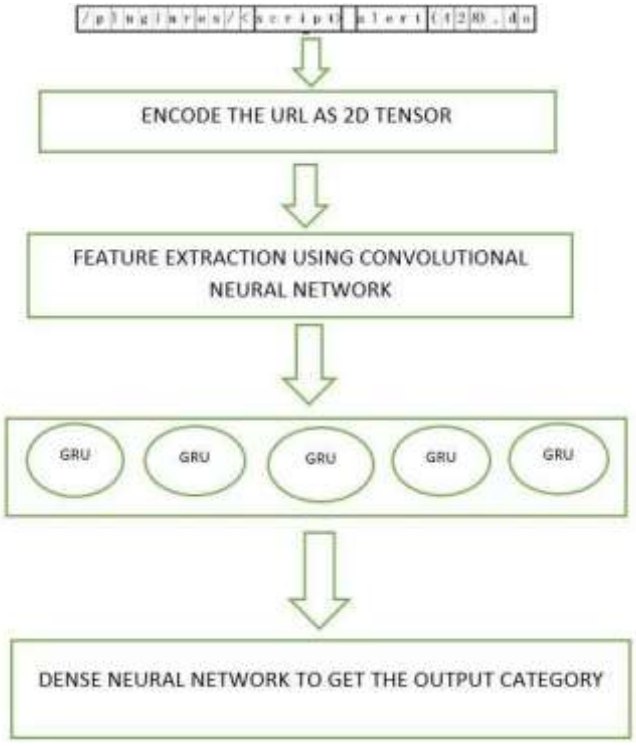

Fig 1. System architecture

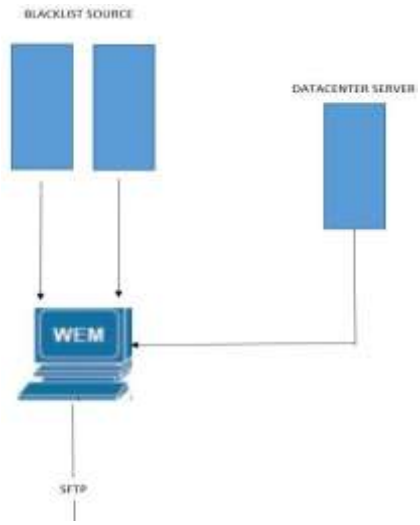

Fig 2. Blacklist architecture

A blacklist of URLs is used in the first step to check whether the URL given matches any of the malicious URLs that are present in the blacklist. A blacklist source is used to store the malicious URL. It is connected to a dataserver which is directly connected to the system. SFTP or SSH File Transfer Protocol is used for a secure connection. If it matches the URLs or keywords present then the URL is blocked and cannot be accessed.It is done by matching the IP address, host name and directory structure. Blacklisting also takes up less memory and hence saves power consumption by eliminating the malicious URL without having it pass through the convolutional neural network. If the URL is not present in the blacklist dataset then it is furthur examined as given below.

Keyword based embedding: The character module makes the original URL into a low dimensional vector. The malicious keyword is extracted from the URL. This is advantageous as it allows more feature extraction to extract the representative feature more quickly. This is the first phase.

Feature Extraction Module: This is the second phase of the system. The module uses the neural network to extract the features from the URL and uses Gated Recurrent Unit as a pooling layer. It also is used to retain some of the features for other purposes. This is an important phase in the extraction of the malicious code. It basically consists of windows which are used to fully extract the features from the URL. Then these features which are extracted are finally merged.

Classification Module: The neural network is used for the classification of the detected features. The neural network is used to arrange the detected features. In case of monitoring the accuracy rate, precision and recall a score is used to rate the progress of the CGRU or Convolutional Gated Recurrent Unit. In addition to this the malicious URL which is then separated from the original URL is then stored in the blacklist so that the program grows more efficient and becomes more power saving as the blacklist will deny access to the URL in the beginning and it will not have to go through the convolutional neural network each time a malicious URL is scanned.

Received: 03 Dec 2019 | Revised: 21 Jan 2020 | Accepted: 18 Feb 2020

V. EXPERIMENTAL RESULTS

381	http://optmd.com
382	http://quikr.com
383	http://xcar.com.cn
384	http://workercn.cn
385	http://p5w.net
386	http://jqw.com
387	http://google.no
388	http://tudou.com
389	http://cbssports.com
390	http://google.com.my
391	http://tinyurl.com
392	http://goodgamestudios.com
393	http://drudgereport.com

Fig 3 Blacklist dataset

	text	label
0	http://facebook.com	0
1	http://youtube.com	0
2	http://yahoo.com	0
3	http://baidu.com	0
4	http://wikipedia.org	0

Fig 4 Output results of original URL after extraction

Figure 3 depicts the dataset called the blacklist that contains all the URLs which have malicious code embedded in them and Figure 4 explains the output which is the URL after the feature extraction process and classification of the URL which is also the original URL which is free of any malicious code or content. A blacklist of URLs is used in the first step to check whether the URL given matches any of the malicious URLs that are present in the blacklist. A blacklist source is used to store the malicious URL. It is connected to a data server which is directly connected to the system. SFTP or SSH File Transfer Protocol is used for a secure connection. If it matches the URLs or keywords present then the URL is blocked and cannot be accessed.

VI. FUTURE SCOPE

This system is capable of performing efficiently in terms of accuracy , low power consumption and detection. The one important area that can be worked upon can be the memory consumption area. Now as the detector keeps scanning malicious code and storing the ones that don't exist on the blacklist the memory keeps increasing on the dataset which in turn increases the memory consumption of the system.

Received: 03 Dec 2019 | Revised: 21 Jan 2020 | Accepted: 18 Feb 2020

VII. CONCLUSION

This project explains an efficient system to detect malicious URLs using a neural network and security using a blacklist for low power consumption.Though there are many methods to extract malicious URLs from a particular URL, the proposed model has been efficient in most aspects and saves power making it more efficient in detection as well as optimization. The power saving aspect is obtained by the use of a blacklist or blocklist which is used to store the pre-existing malicious URLs to cross reference with the URL that is currently being scanned to immidietely deny access and protect the website from being attacked and hacked.

REFERENCES

[1] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "Phishnet: Pre- dictive blacklisting to detect phishing attacks," in 2010 Proceedings IEEE INFOCOM. San Diego, CA, USA: Citeseer, 14-19 Mar 2010, pp. 1–5.

[2] D. Sahoo, C. Liu, and S. C. H. Hoi, "Malicious URL detection using machine learning: A survey," arXiv:1701.07179 [cs.LG], 2017.

[3] J. Saxe and K. Berlin, "eXpose: a character-level convolutional neural network with embeddings for detecting malicious URLs, file paths and registry keys," arXiv:1702.08568 [cs:CR], 2017.

[4] B. Cui, S. He, X. Yao, and P. Shi, "Malicious url detection with fea- ture extraction based on machine learning," International Journal of High Performance Computing and Networking, vol. 12, no. 2, 2018, DOI:10.1504/IJHPCN.2018.10015545.

[5] B. Sun, M. Akiyama, T. Yagi, M. Hatada, and T. Mori, "Automating url blacklist generation with similarity search approach," IEICE TRANSAC- TIONS on Information and Systems, vol. 99, no. 4, pp. 873–882, 2016.

[6] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in Sixth Conference on Email and Anti-Spam(CEAS). California, USA, 2009.

[7] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond black- lists:learning to detect malicious web sites from suspicious URLs," in Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC), 2017 IEEE International Conference on, vol. 1. IEEE, 2017, pp. 143–150.

[8] A. C. Bahnsen, E. C. Bohorquez, S. Villegas, J. Vargas, and F. A. González, "Classifying phishing urls using recurrent neural networks," in Electronic Crime Research (eCrime), 2017 APWG Symposium on. IEEE, 2017, pp. 1–8.

[9] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in Advances in neural information processing systems, 2015, pp. 649–657.

Received: 03 Dec 2019 | Revised: 21 Jan 2020 | Accepted: 18 Feb 2020