International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 11 | Nov 2022 www.irjet.net p-ISSN: 2395-0072

Malicious Link Detection System

Amar Palwankar[1], Rifah Solkar[2], Afiya Borkar[3], Shreya Khedaskar[4], Pranali Shingare[5]

[1]*Prof. Amar Palwankar, IT Engineering Dept & Finolex Academy of Management and Technology, Ratnagiri, India.*
[2] *Rifah Solkar,, IT Engineering Dept & Finolex Academy of Management and Technology, Ratnagiri, India.*
[3]*Afiya Borkar, IT Engineering Dept & Finolex Academy of Managemnt and Technology, Ratnagiri, India.*
[4]*Shreya Khedaskar, IT Engineering Dept & Finolex Academy of Managemnt and Technology, Ratnagiri, India.*
[5]*Pranali Shingare, IT Engineering Dept & Finolex Academy of Managemnt and Technology, Ratnagiri, India.*

Abstract - *Cybersecurity has recently become a serious concern for computer systems due to the rise in Internet usage. Malicious refers to a desire to cause damage. Different harmful URLs release various forms of malware and attempt to collect user data. The use of internet services to conduct business while staying at home increased and changed as a result of the global lockdown in the year 2020. As a result, there were a rising number of cybercrimes committed by cybercriminals and significant data losses for businesses. Malicious URLs must be found and threat types must be recognised in order to halt these attacks. Such websites are frequently found using signature-based approaches, and attempts have been made to impose access restrictions on detected malicious URLs using a variety of security tools. In order to increase the effectiveness of classifiers for identifying dangerous websites using the Logistic Regression Technique of Supervised Machine Learning algorithm, this chapter suggests leveraging linguistic aspects of the related URLs. The findings demonstrate that the ability to recognise harmful websites based solely on URLs and categorise them as spam URLs without depending on page content would lead to significant resource savings as well as a user-safe surfing experience.*

Key Words: Suspicious URL Detection, Machine Learning, Supervised Learning, Logistic Regression, Cybersecurity.

1. INTRODUCTION

Due to the expansion and promotion of social networking, online banking, and e-commerce, the relevance of the World Wide Web (WWW) has drawn more and more attention. New advancements in communication technology not only open up new possibilities for e-commerce, but they also give attackers new opportunities. These days, millions of such websites can be found online and are sometimes referred to as harmful websites. It was stated that the development of technology led to some strategies to target and con consumers, including spam SMS in social networks, online gambling, phishing, financial fraud, false prize claims, and fake TV shopping (Jeong, Lee, Park, & Kim, 2017).

Resources on the Internet are referred to by their Uniform Resource Locator (URL). The features and two fundamental parts of a URL are described by Sahoo et al. These are the protocol identifier, which determines the protocol to use, and the resource name, which identifies the IP address or domain name where the resource is located. Each URL has a distinct structure and format, as can be seen. Attackers frequently attempt to alter one or more URL structural elements in an effort to trick users into sharing their malicious URL. Links that hurt people are referred to as malicious URLs. These URLs will reroute users to websites or resources where hackers can run malicious software on users' computers, send users to undesirable websites, harmful websites.

The assaults using the distributing malicious URL strategy are ranked first among the 10 most popular attack strategies in 2019. According to this figure, the threat level and frequency of assaults using the three primary URL spreading techniques—malicious URLs, botnet URLs, and phishing URLs—increase.

Based on the statistics showing a rise in the distribution of malicious URLs over the course of several years, it is obvious that approaches or methods must be studied and practised to identify and stop these bad URLs. The research also presents a novel technique for extracting URL attributes.

There are now two primary tendencies when it comes to the challenge of identifying malicious URLs: malicious URL identification based on indicators or sets of rules, and malicious URL detection based on behaviour analysis approaches. Malicious URLs can be rapidly and precisely detected using an approach based on a collection of markers or criteria. This strategy, however, is unable to identify new dangerous URLs that do not match the specified indications or guidelines. Based on behaviour analysis approaches, the method for identifying malicious URLs uses machine learning or deep learning algorithms to categorise URLs according to their actions. In this study, URLs are categorised according to their properties using machine learning techniques. The publication also has a brand-new URL attribute extraction method.

In our study, URLs are categorised using machine learning algorithms based on their characteristics and behaviours. The properties are unique to the literature and are taken from the static and dynamic behaviours of URLs. The research's key contribution is those newly suggested features. The entire malicious URL detection system uses

machine learning algorithms. Logistic Regression is the only supervised machine learning algorithm employed.

2. LITERATURE SURVEY

In order to have more information about Malicious Link Detection Systems which are already used to detect suspicious domains, IP's and URL, we referred Research papers based on Malicious Link Detection System using Machine Learning. It gave us information about different techniques used to detect malwares and other breaches with their advantages and disadvantages.

2.1 Overview of Earlier Research Work Done

[1] Mr. Mohammed Alsaedi student of CSE Engineering dept, Mr. Fuad A. Ghaleb a faculty of University Technology Malaysia , Mr. Faisal Saeed student from Birmingham City University and Mr. Jawad Ahmad from Edinburgh Napier University, named "Cyber Threat Intelligence-Based Malicious URL Detection Model Using Ensemble Learning" had put forth their focus and published their International article in (Sensors 2022, 22, 3373. https://doi.org/10.3390/s22093373). This paper describes that a malicious website detection model was designed and developed with a hypothesis stating that cyber threat intelligence is an effective and safer alternative to improve the detection accuracy of malicious websites.

[2] Mr. Shantanu & Mr. Janet B. from Department of Computer Application National Institute of Technology Tiruchirappalli, India. and Mr. Joshua Arul Kumar, Department of ECE MAM College of Engineering Tiruchirappalli, India had studied on the concept of detection of malicious URLs as a binary classification problem and evaluated the performance of several well-known machine learning classifiers entitled "Malicious URL Detection" which was published in (International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/ ©2021 IEEE).

[3] Mr. Zhiqiang Wang, Mr. Xiaorui Ren, Shuhao Li, Bingyan Wang, Jianyi Zhang and Tao Yang from Beijing Electronic Science and Technology Institute researched on malicious URL detection model based on deep learning such that the system model uses word embedding method based on character embedding by combining it, titled "A Malicious URL Detection Model Based on Convolutional Neural Network" published in research paper (Hindawi Security and Communication Networks Volume 2021, Article ID 5518528, https://doi.org/10.1155/2021/5518528).

[4] Mr. Jino S Ganesh, Mr. Niranjan Swarup.V, Mr. Madhan Kumar.R, Mr. Harinisree.A students of P.G under the guidance of Prof. Dr. Giri Raj.M of Mechanical Engineering, Vellore Institute of Technology, Tamil Nadu, India, had worked and made a system by using four different machine learning algorithms, namely logistic regression, decision tree, random forest, multilayer perceptron neural networks to detect malwares and phishing sites entitled "Machine Learning based Malicious Website Detection" published in (International Journal of Scientific & Engineering Research Volume 11, Issue 7, July-2020).

[**5] Mr. Doyen Sahoo, Mr. Chenghao Liu, Mr and Mr. Steven C.H. Hoi** from School of Information Systems, Singapore Management University described that Malicious URL detection plays a critical role for many cybersecurity applications by categorizing them into Blacklist or Heuristic Approach and also used ML approach to classify different spams and malwares named "Malicious URL Detection using Machine Learning: A Survey", published in International article (Vol. 1 August 2019, https://doi.org/10.1145/nnnnnnn.nnnnnnn).

[6] Mr. Ayon Gupta and Mr. Sanghamitra Giri under the guidance of Prof. R. Naresh from Dept. of CSE, SRMIST, Chennai, India researched on the concept that malicious URLs can be detected in real time by using ML algorithms like Support Vector Machine and Logistic Regression to train datasets and detect malicious link entitled "Malicious URL Detection System using combined SVM and Logistic Regression Model" published in (International Journal of Advanced Research in Engineering and Technology, IJARET Volume 11, Issue 4, April 2020).

[7] Mr. Cho Do Xuan & Mr. Hoa Dinh Nguyen of Posts and Telecommunications Institute of Technology, Hanoi, Vietnam and Mr. Tisenko Victor Nikolaevich from Peter the Great St. Petersburg Polytechnic University Russia described that malicious URLs can be detected using two machine learning algorithms RF and SVM by analysing and extracting static behaviour of URLs titled "Malicious URL Detection based on Machine Learning" published in (IJACSA International Journal of Advanced Computer Science and Applications, Vol. 11, No. 1, 2020).

[8] Last but not the least we preferred a book by **Mr. Ferhat Ozgur Catak** professor of University of Stavenger, Ms. Kevser Sahinbas from Istanbul Medipol University and Mr. Volkan Dortkardes from Turkey titled "Malicious URL Detection using Machine Learning" by using Random forest and Gradient boosting ML algorithms to detect malicious URL which was published in book (USA by IGI Global Engineering Science Reference).

2.2 Summary

So basically, after studying the research and review papers of various authors we found that various authors have created a URL/website which is system eco-friendly to Analyse suspicious domains, IPs and URLs to detect malware and other breaches. There are many online websites available for detection of spam and phishing URLs which can be done by entering the link in their system. Therefore, our system will scan and analyse the URL based on the ML

approach and update the result, whether the URL is malicious or not on a single click either it may be from social site or email.

3. METHODOLOGY

A suspicious link is a malicious URL that is designed to promote virus attacks, fraudulent activities, scams and phishing attacks. The method is used to analyze suspicious URLs and prevent the users from being attacked by them.

By clicking on an infected URL, the malware such as virus, trojan, ransomware gets downloaded and can take control of your devices by compromising your machine. Whenever the user clicks on any link provided in the email or any social networking site, the trained model will identify whether the link is suspicious or not. The goal is to classify URLs given as inputs to predict if they are dangerous or inoffensive.

To build this model, we will use a dataset with URLs labelled both bad and good. We selected BAD as a label for the malicious users and GOOD for the legitimate ones. We will train the model using a dataset with many URLs as text already labelled as good and bad. To provide a quick and better view of the data, it is handled and explored to the users in graphical form. For this, data exploration is performed to identify the good and bad URLs using data visualization techniques like bar graph, pie chart.

The learning algorithms would provide the feature extraction of the URLs present in the dataset. The provided URL will read one by one for extracting the features such as suspicious characters, no. of dots and slashes, etc. The technique used in this model is "Bag of words" for extracting features. The URLs are composed of words such as domain name, path, file, extension. This technique works with numerical features and helps to convert words into numerical vectors. It is done by applying Natural Language processing.

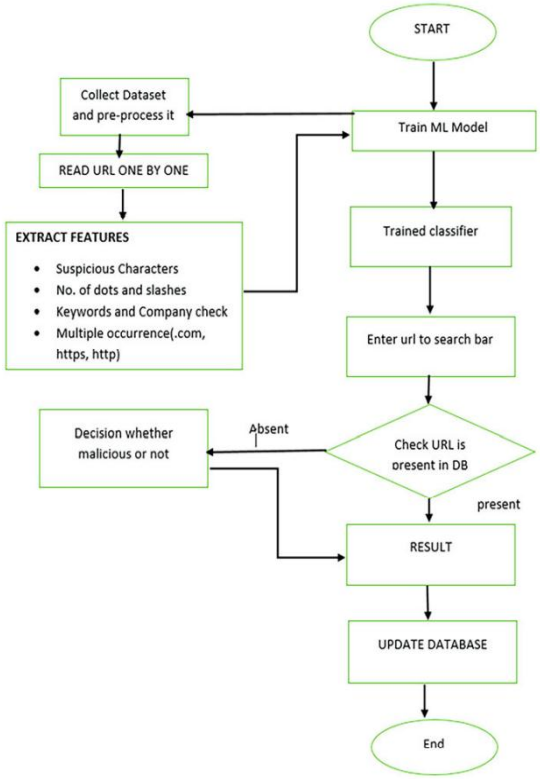
The model used is Logistic regression to find the probability of a certain class. After initializing the algorithm, we will fit the algorithm into our training dataset for learning purposes. We divided the dataset in a training test used to fit the features and feed the model. The URL is checked in the database and if it is present in the database, it has been already checked as malicious or not. But if the URL is not present in the database, then it will go through all the operations and provide the result to the user as malicious or not.

Next step is the identification of the URLs. The system will check and shortlist the link based on directories. The standard URLs will apply to the whitelist and the blacklist directory will include malicious URLs. Another way of checking the URL as malicious is through the filtering process. The model will check for keywords like "com," "www," etc. inside the invalid domain name and if it contains more than four numbers in the domain name it is likely to be a malicious URL. Also, the presence of special characters and any of some famous domains in the URL would also lead to malicious content.

Then with the test set we validate our model with an unbiased evaluation. The model learns during the training phase from the dataset and is used to make predictions. Good URLs have been correctly predicted as authentic URLs and bad URLs as malicious URLs and are labelled as GOOD and BAD respectively. The results obtained classify the URLs as good or bad to predict if they are suspicious or legitimate to use.

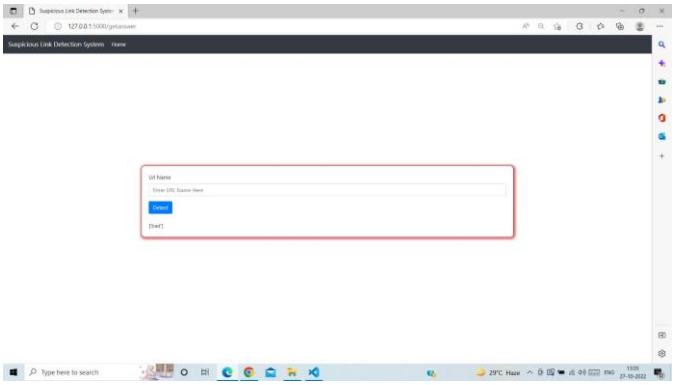
4. SYSTEM ARCHITECTURE

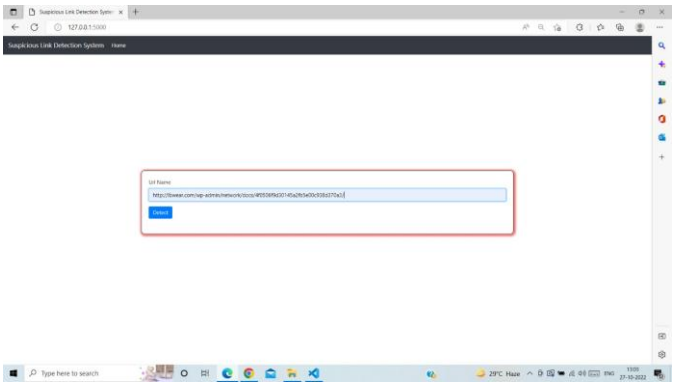
This is the flowchart that how the system will proceed and simulate its process for detection of malicious link.

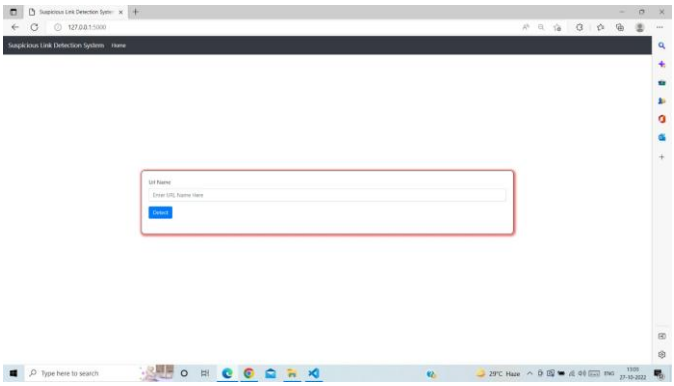


5. RESULTS AND OUTPUT

Following are the screenshots of our results obtained from our system.







The resulting outcome of the training algorithm about how efficient it is to detect the malicious URL website and algorithm with good accuracy gives us better classification results of any website either it is Good or Bad. This is easy to understand by any user and it makes them alert that they are using a malicious website and which can save them from an online scam, online attack, hacking, phishing and credentials detail steal.

6. CONCLUSION

With the evolution in system technology and emerging rise on the internet, millions of people exchange their information over the social sites and also do many activities related to their daily life. During these processes, users have intelligence and critical information such as descriptive username & passwords and mostly networks detect their users with them. Most of the users are unaware about their saved information which can be exploited by attackers and thus they can become victim of such malicious and phishing websites. Therefore, the detection of harmful web pages has become very important to protect the users of the web environment from these threats. So, to overcome such with such situations Malicious URL detection plays a critical role for many cybersecurity applications. The techniques used in machine learning are promising method.

We provided an organised system for malicious URL detection with the help of machine learning. Also, we provided detailed explanation of current research on the detection of malicious link, by creating representation of feature sets and new learning algorithms plotting for dealing with the detection of malignant URL. We had used Logistic Regression machine learning algorithm which is more convenient than Random Forest or Naive Bayes for suspicious link detection. The experimental results of the proposed method indicate that the performance of ML model in processing large dataset and predicting the website as benign or malicious is significantly good. This indicates we can quickly build deployable and reliable machine learning models for malicious link detection.

7. FUTURE SCOPE

The research team successfully proposed a method where URLs can be used directly to extract features and classify them as good or bad. The study is inspired by this and focuses on this methodology only hence in order to grasp the global features of malicious URLs and extract high-dimensional features based on pre-processed data, deep learning algorithms are potentially worthwhile topics for future research studies that we will be taking under consideration. Deep learning has become the mainstream malicious URLs detection system these days. Deep learning can automatically extract features which frees up the time and feature engineering.

Implementing Deep Learning will not only allow us to process the data faster but also will improve the performance of malicious detection. In order to keep behind the drawback of time-consuming and labour-intensive machine learning implementation which extracts shallow features, deep learning implementation will be preferred.

The future scope solely is focused on study findings and implementation of deep learning into our major project in the given time being.

REFERENCES

[1] Mohammed Alsaedi, Fuad A. Ghaleb, Faisal Saeed, Jawad Ahmad_(2022) Cyber Threat Intelligence-Based Malicious URL Detection Model Using Ensemble Learning. International article in (Sensors 2022, 22, 3373. https://doi.org/10.3390/s22093373).

[2] Shantanu, Janet B, Joshua Arul Kumar R_(2021)Malicious URL Detection.(International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281- 9537-7/20/ ©2021 IEEE).

[3] Zhiqiang Wang, Xiaorui Ren, Shuhao Li, Bingyan Wang, Jianyi Zhang, Tao Yang_(2021) A Malicious URL Detection Model Based on Convolutional Neural Network. (Hindawi Security and Communication Networks Volume 2021, Article ID 5518528, https://doi.org/10.1155/2021/5518528).

[4] Jino S Ganesh, Niranjan Swarup.V, Madhan Kumar.R, Harinisree.A and Dr. Giri Raj.M_(2020) Machine Learning based Malicious Website Detection. (International Journal of Scientific & Engineering Research Volume 11, Issue 7, July-2020).

[5]Doyen Sahoo, Chenghao Liu, Steven C.H. Hoi_(2019)Malicious URL Detection using Machine Learning: A Survey International article (Vol. 1 August 2019, https://doi.org/10.1145/nnnnnnn.nnnnnnn).

[6] Ayon Gupta, Sanghamitra Giri, R. Naresh_(2020)Malicious URL Detection System using combined SVM and Logistic Regression Model.(International Journal of Advanced Research in Engineering and Technology, JARET Volume 11, Issue 4, April 2020).

[7] Cho Do Xuan, Hoa Dinh Nguyen_(2020) Malicious URL Detection based on Machine Learning. (IJACSA International Journal of Advanced Computer Science and Applications, Vol. 11, No. 1, 2020).

[8] Mr. Ferhat Ozgur Catak professor of University of Stavenger, Ms. Kevser Sahinbas from Istanbul Medipol University and Mr. Volkan Dortkardes from Turkey_(2020) (USA by IGI Global Engineering Science Reference).