iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VII **Month of publication:** July 2023

DOI: https://doi.org/10.22214/ijraset.2023.54563

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VII Jul 2023- Available at www.ijraset.com

Malicious URL Detection Using Logistic Regression

Harshitha S[1], Vidhyashree N P[2], Swati S Shetty[3]
Assistant Professor, Department Of Computer science and Engineering, Malnad College Of Engineering
UG Student, Department Of Computer science and Engineering, Malnad College Of Engineering

Abstract: ***Over the past few years, there has been a considerable expansion in both the amount and variety of web services. Online services like social networking, online gaming, and banking have quickly advanced along with people's reliance on them for routine chores. As a result, a lot of information is added to the Web every day. These web services open up new ways for people to engage, but they also give thieves new chances. URLs serve as jumping off points for all types of web attacks, making it possible for any user with bad intent to submit a malicious URL and steal the identity of a legitimate individual.***
Malicious URL detection is a crucial task in ensuring the security of internet users. This study describes a novel logistic regression technique for identifying malicious URLs.. The proposed method leverages a dataset consisting of various features extracted from URLs and their associated labels indicating whether they are malicious or not. To extract features, we consider both structural and content based characteristics of URLs. Structural features include domain length, path length, and presence of special characters, while content based features involve examining the lexical composition of the URL, such as the presence of suspicious keywords or uncommon words. Using a labelled dataset, the logistic regression model is trained using the retrieved features. The likelihood that a given URL is malicious is then predicted using the trained model. Results from experiments show how effective is the suggested strategy. When identifying fraudulent or benign URLs, the logistic regression model performs with high accuracy.
Overall, by offering a dependable and effective method for identifying fraudulent URLs, this research makes a contribution to the field of cybersecurity. To improve defence against online attacks and give internet consumers a safer surfing experience, the proposed logistic regression model can be incorporated into current security systems.
Keywords: URL, Malicious, Logistic Regression, Features, Labelled dataset

I. INTRODUCTION

Drive-by downloads and the theft of personal information, including credit card numbers and passwords, are both done through phishing websites. Since it is simpler to deceive a victim, phishing is popular among thieves. The majority of the time, such bothersome activity diverts network resources meant for other purposes into hitting a malicious link that seems authentic rather than attempting to get past a computer's security measures.
Such unpleasant behaviour typically interferes with network resources meant for other purposes and almost always jeopardises the network's and/or its data's security. Using machine learning methods like logistic regression is one way to find harmful URLs. When attempting to predict which of two classes an input belongs to, binary classification problems are frequently solved using the supervised learning process known as logistic regression. The two classes used for malicious URL detection are usually malicious and benign.

A. Uniform Resource Locator(URL)

A URL serves as a distinctive identifier for finding resources on the internet. It's also known as a web address. Instructing a web browser on how and where to access a resource, URLs include numerous components such a protocol and domain name. End users can access URLs by directly putting them into their browser's address bar or by clicking a hyperlink they discover on a website, in a favourites list, in an email, or in another application.
The type of Uniform Resource Identifier (URI) that is used the most frequently is a URL. To locate a source via a network, typescript strings called URIs are utilised. The internet cannot function without URLs. The URL contains both the name of the protocol needed to access a resource and the name of the resource itself. A URL's first part specifies the protocol to utilise as the main access method. The resource's location is identified by its IP address, domain name, and sometimes a subdomain in the second part.

Following the domain, a URL can also include the following information:

1) A route to a specific page or file inside a domain;
2) A network port to utilise when creating the link.
3) Commonly used request or search parameters seen in search URLs.

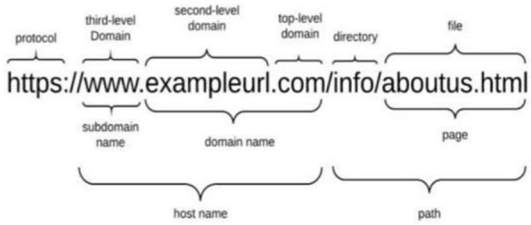


Fig. 1. Structure Of URL

A malicious URL is a link made specifically to spread fraud, attacks, and scams. Malicious URLs that are clicked on can launch phishing or spear phishing emails, download malware, or trigger other types of cybercrime. Malicious URLs pose a huge threat to the digital world since they are frequently concealed and simple to overlook. Users should avoid clicking on strange links or downloading files from dubious emails or websites to prevent being exposed to harmful URLs.

B. Logistic Regression

In classification problems, where the objective is to estimate the likelihood that a given instance belongs to a certain class, the major application of the supervised machine learning method known as logistic regression is used. Logistic regression is a technique used in classification algorithms. It employs a sigmoid function to estimate the probability for the given class and is known as regression since it inputs the output of a linear regression function.

II. RELATED WORK

Machine learning-based malicious URL detection is a crucial responsibility in the cybersecurity industry. On this subject, numerous studies and research articles have been written. I'll give an overview of some of the pertinent and current research in this topic in this literature review.

Mohammed Nazim Feroz and Susan Mengel [1] offer a method that automatically categorises URLs based on their lexical and host-based characteristics. These techniques can extract and automatically create highly analytical models. For such scalable machine learning issues, Mahout was created, and online learning is preferred to batch learning. The classifier finds a lot of phishing hosts and maintains a low false positive rate in order to reach 93–95% accuracy.

Stefan Savag, Justin Ma, Lawrence K. Saul, Geoffrey M. Voelker, and [2] outlines a solution for solving this issue that relies on automated URL classification, employing statistical techniques to identify the telltale lexical and host-based characteristics of malicious Web site URLs. By extracting and regularly analysing tens of thousands of variables that could be suggestive of questionable URLs, these technologies are able to learn highly analytical models. The resulting classifiers achieve 91–94%accuracy, [3]detection of a significant number of hazardous Web sites from their URLs, with only a small amount of false positives

III. IMPLEMENTATION

A. Problem statement

The goal is to create a logistic regression model that can precisely predict if a particular URL is dangerous or not using a dataset of URLs. Malicious URLs are ones that are intended to steal sensitive information, exploit holes in a user's computer or network, or reroute people to dangerous websites. The model should consider a number of URL characteristics, including its length, the existence of particular keywords, the domain name, and other metadata. The idea is to spot trends in the data that point to dangerous URLs and use those patterns to predict the future with accuracy.

The performance of the model will be evaluated using standard metrics such as accuracy and F1-score. The ultimate aim is to develop a robust and reliable system for detecting and preventing malicious URLs, thereby protecting users from cyber attacks and safeguarding their sensitive information.

B. Objective

The main goal of this task is to detect and prevent users from accessing malicious websites, which can be used to distribute malware, steal sensitive information, or engage in other nefarious activities. The below mentioned steps will be performed to achieve the stated objective:

1) Developing dataset consisting of correct and malicious URL along with labels.
2) Processing the model using Logistic regression.
3) Feature extraction is performed using TF-IDF Vectorizer.
4) Training and testing the developed model.

C. Methodology

1) *Data Gathering*: Compile a list of URLs that have been classified as harmful or benign. The size and variety of this dataset should permit the logistic regression model to generalise well to previously unreported data.
2) *Preprocess the data*: The preprocessing step plays a crucial role in malicious URL detection. It involves cleaning and transforming the Raw URLs into a format that can be effectively used for feature extraction and analysis. Here are some common preprocessing steps for malicious URL detection:
a) *URL Normalization:* Convert the URLs to a standard format by removing unnecessary components like protocol (e.g., ”http://”, ”https://”) and www prefix. Normalize the URLs to ensure consistent representation for similar URLs.
b) *Tokenization:* Split the URLs into individual components or tokens, such as domain, path, query parameters, and file extensions. This step helps in capturing important structural elements of the URL.
c) *Lowercasing:* Convert all characters in the URL to lowercase. This step ensures that the model does not treat the same words with different cases as different features.
d) *Removal of Special Characters:* Remove special characters, punctuation marks, and other noise from the URL. This helps in eliminating irrelevant information that may not contribute to the classification task.
e) *Handling of Numbers and Digits:* Decide how to handle numbers and digits in the URL. You may choose to replace them with a special token or remove them altogether, depending on your analysis goals.
f) *Feature Engineering:* Depending on your specific requirements, you can engineer additional features from the URL, such as the length of the URL, presence of specific keywords or patterns, or the frequency of certain tokens. These features can provide valuable information to the classification model.

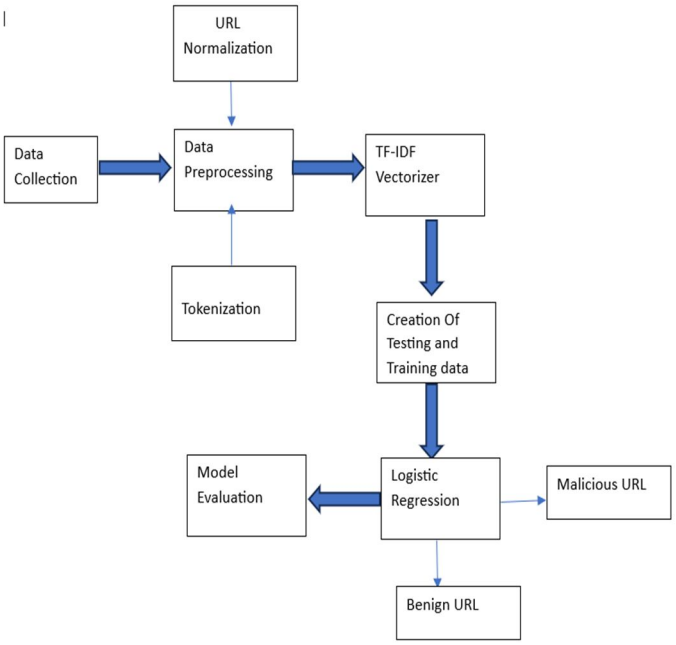


Fig. 2.Proposed model of Malicious URL Detection

g) *Create TF-IDF Vectors:* Use the TF-IDF vectorizer to convert the pre-processed URLs into numerical feature vectors. The TF-IDF vectorizer calculates the importance of each term (word or token) in a URL within the context of the entire dataset. It assigns higher weights to terms that are more informative and discriminative. By using the TF-IDF vectorizer, you can represent each URL as a high-dimensional vector where each component corresponds to a specific term. These feature vectors can then be used as input to machine learning algorithms for training classifiers or performing other analysis tasks to detect malicious URLs.

h) *Training and Testing:* Create training and testing sets for the dataset before beginning to train the logistic regression model. Use a suitable algorithm or library to then train a logistic regression model on the training data. The logistic regression model will discover the correlation between the input features and the maliciousness or benignity of the target variable.

i) *Model Evaluation:* Use the testing set to gauge how well the trained model performed. Accuracy, precision, recall, and F1 score are typical evaluation criteria. The model's ability to generalise to fresh, untested data is evaluated in this step.

IV. RESULT and ANALYSIS

The table I gives the summary of the various entities compared for the different community detection algorithms.

TABLE I
COMPARISON OF VARIOUS DETECTION ALGORITHMS

Algorithm	Accuracy
Logistic Regression	96
Decision Trees	95

As observed in the table I, Logistic Regression algorithm generates the highest accuracy score. Thus it is most suitable algorithm for malicious URL detection.

The below figure represents the amount of malicious and benign URLs present in the dataset

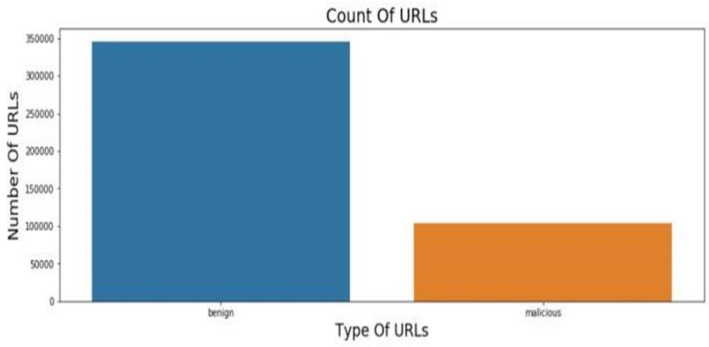
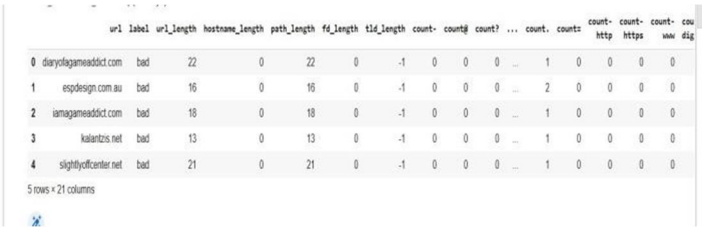


Fig. 3. Representaion of Malicious and Benign URLs



	url	label	url_length	hostname_length	path_length	fd_length	tld_length	count-	count@	count?	...	count.	count=	count-http	count-https	count-www	cou dig
0	diaryofagameaddict.com	bad	22	0	22	0	-1	0	0	0	...	1	0	0	0	0	
1	espdesign.com.au	bad	16	0	16	0	-1	0	0	0	...	2	0	0	0	0	
2	iamagameaddict.com	bad	18	0	18	0	-1	0	0	0	...	1	0	0	0	0	
3	kalantzis.net	bad	13	0	13	0	-1	0	0	0	...	1	0	0	0	0	
4	slightlyoffcenter.net	bad	21	0	21	0	-1	0	0	0	...	1	0	0	0	0	

5 rows × 21 columns

Fig. 4. Representaion of dataset along with features

```
193
194 #checking accuracy of model using test data
195 print("Accuracy of model is:",logistic.score(X_test,y_test))
196
197 #checking accuracy of model using train data
198 print("Accuracy of model is :",logistic.score(X_train,y_train))
199
200 #prediction with model created
201 X_predict = ["google.com/search=faizanahmad",
202 "pakistanifacebookforever.com/getpassword.php/",
203 "www.radsport-voggel.de/wp-admin/includes/log.exe",
204 "ahrenhei.without-transfer.ru/nethost.exe ",
205 "www.itidea.it/centroesteticosothys/img/_notes/gum.exe"]
206 X_predict = vectorizer.transform(X_predict)
207 New_predict = logistic.predict(X_predict)
208 print(New_predict)
209
210
```

Fig. 5. Input for Prediction

```
Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
Accuracy of model is: 0.9618160845730322
Accuracy of model is : 0.9723281733562049
['good' 'good' 'bad' 'bad' 'bad']
```

Fig. 6. Output of the model

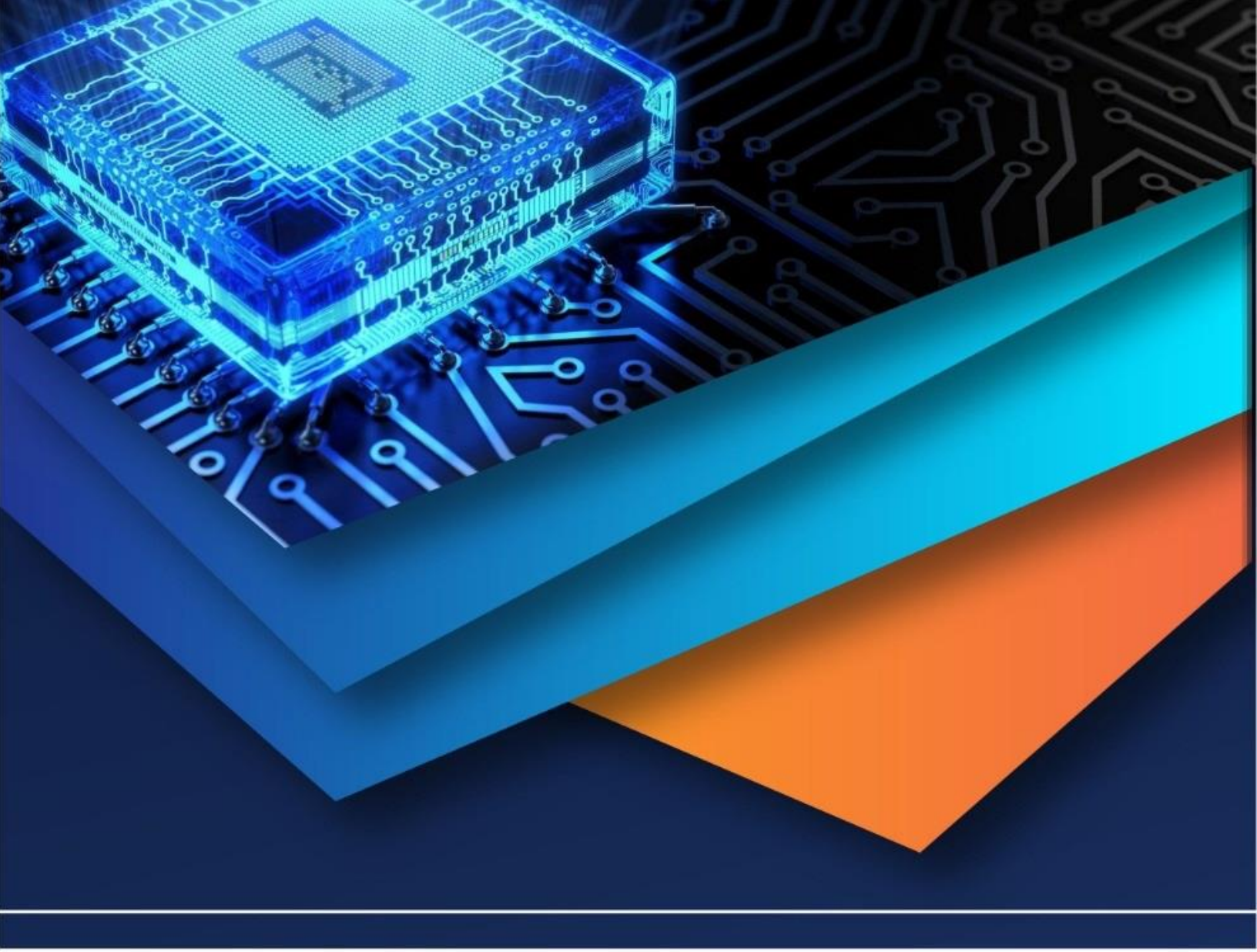
V. CONCLUSION

In this Project, we found the effectiveness of logistic regression for detecting malicious URLs. Our analysis included a comprehensive dataset of URLs labelled as either malicious or benign. We employed various features extracted from these URLs, such as domain characteristics, path information, and lexical attributes. Our findings are summarised as follows:

1) The logistic regression model demonstrated promising performance in distinguishing between malicious and benign URLs.
2) The feature analysis revealed that certain characteristics, such as the presence of suspicious keywords and the length of the URL, contributed significantly to the model's predictive power.
3) It achieved an accuracy of 96% of test set.

REFERENCES

[1] Rupa Chiramdasu, Gautam Srivastava, Thippa Reddy Gadekallu, Sweta Bhattacharya, and Praveen Kumar Reddy. Using logistic regression, find malicious urls. Pages 1-6 of the COINS 2021 IEEE International Conference on Omni-Layer Intelligent Systems. IEEE, 2021.
[2] R Vinodini, A Kavitha, and A Saleem Raja. Malicious url detection utilising lexical characteristics and machine learning methods. Proceedings: Materials Today, 47:163–166, 2021.
[3] V Vinodhini and N Vanitha. Using logistic regression to detect malicious URLs. IJEMR, 9(6):108–113, 2019. International Journal of Engineering and Management Research.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY