DETECTION OF MALICIOUS URLs USING MACHINE LEARNING

Purav Patel
Information Technology department
Shah And Anchor Kutchhi Engineering College,Chembur.
Mumbai,India.

Ansh Vaghela
Information Technology department
Shah And Anchor Kutchhi Engineering College,Chembur.
Mumbai,India.

Mr. Panjab Mane
Assistant Professor, Information Technology department
Shah And Anchor Kutchhi Engineering College,Chembur.
Mumbai,India.

Abstract— **There has been a massive growth in the sheer number of people using the world wide web in the past 15 years. Services ranging from banking to education, from social media to gaming have attracted many people to perform their daily tasks. Due to this, large chunks of information are transferred through the web on a daily basis. Along with all the advancements made there has been a massive rise in cybercriminals who prey on the wealth of information which is available online. They target unsuspecting users for different types of information ranging from private data to bank details or to even steal identity of a legal person to engage in illegitimate activities. It has been a great challenge for the service providers to keep cybercriminals off their users so that their data is secured at any given time. We know that URLs are a gateway to any service a user wants to access and is also the most vulnerable piece of the link which connects the user to the service. Attackers often spoof or provide fake URL's which leads users to unknowingly provide them with sensitive data. In this paper we have proposed a machine learning system which uses Logistic Regression to identify malicious URLs. This can be useful for users to check if a URL is safe to visit or not.**

I. Introduction

Malicious websites have become very common and they depend on sheer number of users that visit that site. For example, Facebook is widely used all over the world and attackers would try to create a phishing website which looks similar to the actual website, which resembles the original site and the user is fooled into believing that they are visiting the legitimate Facebook while their data is being compromised and stolen by the attackers in real time. Data theft, Identity theft, Monetary losses, Sensitive information loss are among multiple other losses bared by unsuspecting users every day. Damages ranging in billions of dollars are caused every year due to these internet frauds, scams and identity theft instances. The URL plays a very important role in identifying if a website is safe to visit or not. By carefully examining the URL one can make sure they are not getting tricked into visiting some other website pretending to be the original one. Regrettably, technical advancements have resulted in highly developed tactics for defrauding and attacking web users. Con sites that sell phony products, financial scams that deceive users into giving sensitive data that leads to identity or money theft, or the installation of malware on the user's device are examples of these attacks. In general, phishing sites employ social engineering to lure people in by sending them a fake link that takes them to a fake website. The bogus link is posted on well-known websites or emailed to the victim. The bogus site looks exactly like the real one. Instead of redirecting the user's request to the original web page, it will be directed to the web server of the attacking host.

To address the aforementioned flaws, we developed a system that uses machine learning techniques to detect fraudulent URLs. Feature extraction is a key part of detecting harmful uniform resource locators. We've gathered a training dataset, which is subsequently used to train for the extracted feature. We retrieved three primary attributes: host-based, lexical-based, and popularity-based. We employed one machine learning techniques in this paper: logistic regression. This strategy is then used to the dataset in order to train it and discover any harmful URLs.

To solve the drawbacks mentioned above we have developed a machine learning system which can detect malicious URLs. We tested multiple algorithms in order to achieve maximum accuracy. Support vector machine (SVM), Naïve-Baiyes, Logistic Regression are among the algorithms that we actively tested. Going by the accuracy, Logistic Regression was the most accurate and we decided to go with that. A training dataset was used to train the machine learning algorithm. The resulting accuracy obtained was around 95-97% in detecting malicious URLs with only modest false positives

II. Proposed work

Blacklist method is one of the most common and used method to detect malicious URL. Blacklist is nothing but a database of URLs that are confirmed to be malicious in the past. The blacklist database is compiled over time by crowd sourcing way when a malicious URL is encountered by any user who is accessing the web. Blacklist method is simple fast and easy to implement. At times, this method can be ambiguous and give a high false positive rate and maintaining an exhaustive list of malicious URLs in a database can be difficult because new malicious URLs are generated almost every day.

For evading blacklisting of malicious URL attackers make use of obfuscation which makes the URLs “appear” legitimate/safe. Attackers also make use of URL shortening services and generally hide the malicious URL behind a shortened URL. Once, any users visit such a URL the malicious code which is embedded in the JavaScript is executed and the attack is launched. In order to avoid the detection of malicious code by signature-based authentication tools and code is also obfuscated along with the URL. To overcome this, machine learning approach is utilized which uses a dataset of malicious and safe URL as training data and based on the corresponding statistical output a prediction model is developed which classifies the URL as safe or malicious. The proposed work in this paper represents experimental evaluation for classification of URLs using Logistic Regression algorithm.

III. Related work

In the paper Malicious URL detection using Logistic regression technique written by Vanitha N and Vinodhini V, they have proposed a system where the URLs are classified and fed into the system which is the machine learning algorithm called logistic regression which is used for binary classification. This method helped them achieve over 97% accuracy by learning phishing URLs. In order to train the model, they have used a pre-defined data set which identifies websites based on various attributes such as distance of the URL, Number of dots present, token-based diagrams to achieve the desired result.

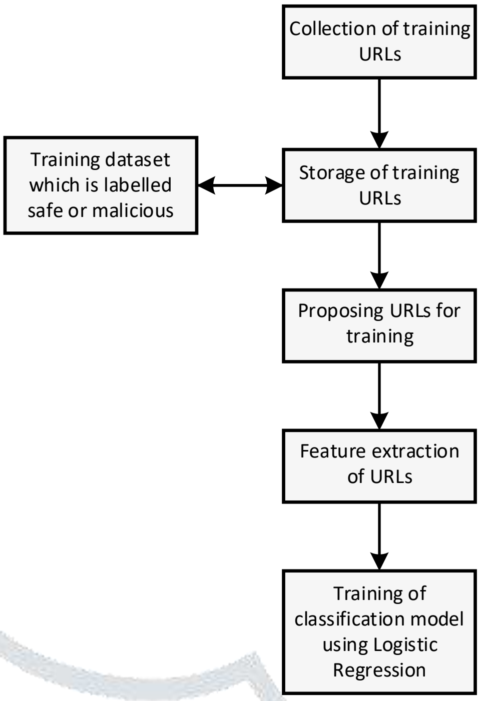
R. Naresh, Ayon Gupta, Sanghamitra Giri have described a method in their study about deploying URL lexical options, payload size and python supply options and with the help of support vector machine and logistic regression they were able to achieve over 98% accuracy.

A study conducted by Chunlin Liu, Lidong Wang, Bo Lang, Yuan hao stated that various methods used for identification of malicious URLs were employed and accuracy ranging from 95-99% was achieved when multiple testing features were used at once instead on single point testing. The problem for false negatives and false positives was considered and fixed with multiple approaches.

IV. Dataset

The malicious and safe URLs present in the dataset have been collected and compiled from various sources such as Kaggle, DMOZ Open directory and ISCX-URL2016 to name a few. It contains a total of 450,177 URLs out of which 104,438 are malicious and 345,739 are safe. This dataset is then split into testing and training keeping the “test_size” parameter as 0.2 which gives us the size of training data as 360,140 and size of testing data as 90,036.

The figure given below describes the steps involved in collection, storage and training of URLs.


Collection of training URLs
Training dataset which is labelled safe or malicious
Storage of training URLs
Proposing URLs for training
Feature extraction of URLs
Training of classification model using Logistic Regression

V. Methodology and implementation

5.1 Machine learning approach

Machine learning is a subset of Artificial Intelligence (AI) which offers skills to the model so that it can learn by itself and improve by itself based on the previous experiences without being explicitly programmed. Machine learning provides the access of the data to the model so that it can analyze and learn by themselves.
For a model to be trained by machine learning data plays a very important role, with the data that is provided to the model it can make a better conclusion without any additional human interference or support.

Supervised machine learning

Supervised learning is used in the vast majority of machine learning applications.
When you have input variables (x) and an output variable (Y), supervised learning is when you use an algorithm to learn the mapping function from the input to the output Y=f (X). The aim is to get close enough to the mapping function that you can predict the output variables (Y) for new input data (x).

In supervised learning, data for both input and preferred output is given to provide a learning basis for future data analysis, input and output data are labelled for classification. It can be thought of as an instructor supervising the learning process. We know the correct answers, so the algorithm iterates through the training data, making predictions that are then corrected by the instructor. When the algorithm reaches a satisfactory level of success, learning comes to an end.

Regression

Regression analysis is a type of supervised learning in which a set of machine learning methods for predicting a continuous outcome variable (y) based on the values of one or more predictor variables (x).
Regression is used for predicting the value of an answer (dependent) variable using one or more predictor variables, with the variable being numeric. Linear, multiple, logistic,

polynomial, non-parametric, and other types of regression exist.

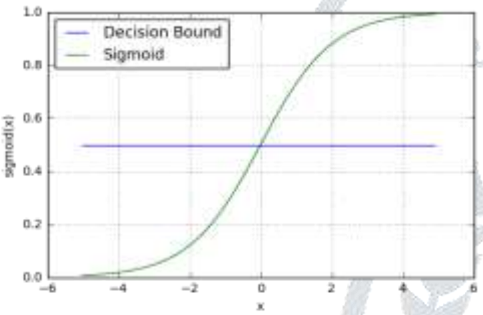
Logistic regression

Logistic regression is a statistical model that employs a logistic function to model a binary dependent variable. In logistic regression (or logit regression) is a method of estimating the parameters of a logistic model (a form of binary regression). A binary logistic model has two possible dependent variables.

A binary logistic model, mathematically, has a dependent variable with two possible values, such as pass/fail, which is expressed by an indicator variable, with the two values labelled "0" and "1."

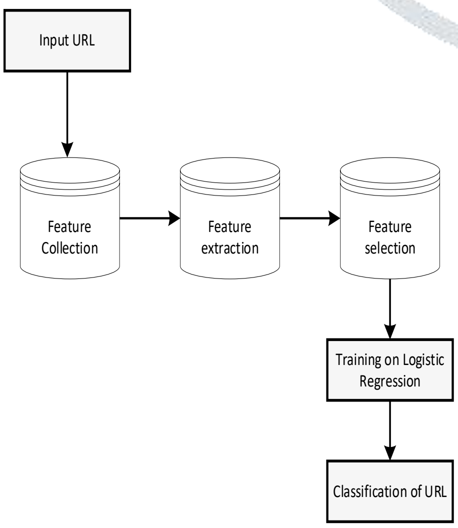
The log-odds (logarithm of the odds) for the value labelled "1" in the logistic model is a linear combination of one or more independent variables ("predictors"); the independent variables can either be a binary variable (two classes, each coded by an indicator variable) or a continuous variable (any real value).

1/(1+e^-x)



5.2 Working Methodology and Implementation

5.2.1 Architecture of the model



The proposed work in this paper represents experimental evaluation for classification of URLs using Logistic Regression algorithm. Our main aim is to make develop a model which can be trained and fitted to detect safe or malicious URL. We focus on teaching our model how to optimize our output by having a higher rate of detection and a low outcome of false positive.

After collection and compilation of dataset, the most critical step is to extract informative features from a URL which is helpful for mathematical interpretation by the model. URL is a string of characters which cannot be fed directly to a machine learning model hence we have to extract suitable features which is called as feature extraction. This encompasses lexical features, statistical properties of the URL string, bag of words, n-gram, etc., as well as host-based features such as WHOIS information and the host's geo-location properties.

The features that we have obtained by feature extraction are further processed into a numerical format using TF-IDF Vectorizer and Count Vectorizer which are fed to our machine learning model. Since the underlying assumption of machine learning (classification) models is that feature representations of malicious and benign URLs have different distributions, the ability of these features to provide relevant information is crucial to subsequent machine learning. As a result, the quality of the URL feature representation is vital to the quality of the machine-learned malicious URL predictive model.

The actual training of the model is the next step in constructing the prediction model using the training data and the required feature representation. Many classification algorithms (Naive Bayes, Support Vector Machine, Logistic Regression, and so on) can be used directly on the training data.

5.2.2 Feature representation

As previously mentioned, the quality of the training data, which is dependent on feature representation quality, is critical to the success of a machine learning model.
Given a URL a∈ U, where U denotes the domain of any valid URL strings, the goal of feature representation is to find a mapping g: U → Rd , such that g(a) → x where x ∈ R^d is a d-dimensional feature vector, which can be fed into the machine learning model.
The feature representation step is broken down into two parts:

1. Feature Collection: This is an engineering-oriented phase that collects specific and relevant URL information. This includes data such as the existence of URLs in a blacklist, features extracted from the URL String, information about the host, website material such as HTML and JavaScript, popularity data, and so on.

2. Feature Pre-processing: In this step, the unstructured knowledge about the URL (e.g., textual description) is properly formatted and transformed to a numerical vector such that it can be fed through machine learning algorithms. For e.g., numerical details may be used as is, and Bag-of-words is often used to represent textual or lexical material.

5.2.3 Feature extraction

Researchers have proposed various kinds of features that can be used to provide valuable information for malicious URL identification. These features are classified as follows: Blacklist Features, URL-based Lexical Features, Host-based Features, Content-based Features, among Others (Context and Popularity).

A. Lexical features

Lexical features are characteristics derived from the URL name's properties (or the URL string). The motivation is that the deceptive nature of a URL should be detectable depending on how it "looks." Many obfuscation techniques, for example, attempt to "sound" like innocuous URLs by mimicking their names and incorporating a minor variation to it. Malicious URLs, including those used in phishing attacks, typically have distinguishable patterns in their URL. The standard domain/path token length (delimited by ‘.', ‘/', ‘?', ‘=', ‘-',) and phishing URLs reveal entirely different lexical patterns.

B. Link popularity features

One of the most important options used in this strategy is "link popularity," which can be calculated by examining the number of incoming links from other websites. Malicious sites have a lower level of link popularity, while many benign sites have the highest level of link quality.

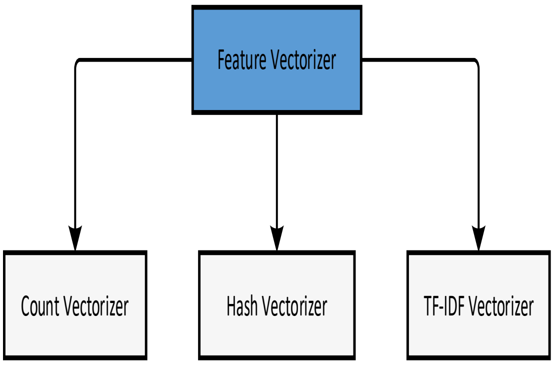
C. Host based features

The URL's host-name properties are used to obtain host-based functionality. They allow us to determine the location, identity, management style, and properties of malicious hosts. They investigated the effect of a few host-based features on the maliciousness of URLs. Some of the major findings were that phishers used Short URL services; the time-to-live from domain registration was almost instant for the malicious URLs; and many used botnets to host themselves on numerous computers in multiple countries. As a result, host-based capabilities have been a critical component in identifying malicious URLs.

D. By content of webpage

Content-based features are those that can be accessed by downloading the entire website. As opposed to URL-based features, these are "heavy-weight," since a large amount of information must be collected, and safety issues can occur. With the recent development and advancement in the field of dynamic webpage technology attackers have been able to inject malicious code in to the website which can lead to cross site scripting attack, buffer overflow and SQL injection to name a few.

5.2.4 Feature Vectorizer



The URLs present in the dataset are converted into machine readable format by using feature extraction technique in the above-mentioned step. The feature extraction serves as an input for making a vector of the URL using feature vectorizer technique. The types of feature vectorizer techniques are:

1. TF-IDF Vectorizer: TF-IDF stands for “word frequency-inverse text frequency,” which means that the weight assigned to each token is determined not just by its frequency in a text, but also by how persistent the term is across all corpora.

2. Hash vectorizer: This one is designed to use as little memory as possible. The vectorizer uses the hashing trick to encode the tokens as numerical indices instead of storing them as strings. The disadvantage of this approach is that once vectorized, the names of the functions cannot be retrieved.

3. Count vectorizer: The most basic, it counts the number of times a token appears in the text and uses this value as its weight.

VI. EXPERIMENTAL ANALYSIS

The training of the model was done one in split ratio of 1:1, 1:4 and 10:1 using Logistic Regression machine learning algorithm which was imported using sklearn model from python sci-kit library (from sklearn. linear_model import Logistic Regression).

The results obtained by the model in the form of a confusion matrix are:

	Predicted as positive	Predicted as negative
Labelled as positive	True Positive (TP)= 68915	False Negative (FP) = 6
Labelled as negative	False Positive (FP) =406	True Negative (TN) = 20709

The classification report obtained is:

	Precision	Recall	F1-Score	Support
Safe	1.00	1.00	1.00	68921
Malicious	0.98	0.98	0.99	21115
Accuracy			1.00	90036
Macro avg	0.99	0.99	0.99	90036
Weighted avg	1.00	1.00	1.00	90036

IX. CONCLUSION

Detection of malicious using machine learning plays a vital role in cyber security applications. Across the world many users are tricked into opening malicious URL everyday which inflicts great personal and professional damage. Using this model, we are able to predict the nature of the URL beforehand which is advantageous and loss can be avoided. The method proposed in this paper using Logistic Regression algorithm and had shown successful results on large experimental datasets, by using Flask and Heroku we were also able to deploy it on the web. Detection of malicious URL using machine learning has shown better results as compared to traditional detection methods. In future, for development and increasing the accuracy of the model more training and testing is to performed on the model using more feature extraction methods so that the model can handle more difficult and upcoming challenges.

X. REFERENCES

[1]. Zhang Y, Hong J I, Cranor L F. Cantina: a content-based approach to detecting phishing web sites ［C ］ / /16th International World Wide Web Conference. Banff, Alberta,Canada, 2007: 639-648 .

[2]. Chunlin Liu, State Key Lab of Software Development Environment School of Computer Science and Engineering, Beihang University Beijing, China

[3]. Lidong Wang, National Computer Network Emergency ResponseTechnical Team/Coordination Center of China Senior Engineer Beijing, China

[4]. Bo Lang, State Key Lab of Software Development Environment School of Computer Science and Engineering, Beihang University Beijing, China

[5]. R. Naresh**,** Associate Professor, Dept. of CSE, SRMIST, Chennai, India, , International Journal of Advanced Research in Engineering and Technology (IJARET) Volume 11, Issue 4, April 2020

[6]. Ayon Gupta**,** Dept. of CSE, SRM IST, Chennai, India, International Journal of Advanced Research in Engineering and Technology (IJARET) Volume 11, Issue 4, April 2020

[7]. Sanghamitra Giri , Dept. of CSE, SRM IST, Chennai, India, International Journal of Advanced Research in Engineering and Technology (IJARET) Volume 11, Issue 4, April 2020

[8]. Vanitha Anandkumar Dr. N.G.P. Arts and Science College, **Article** in International Journal of Engineering Business Management · December 2019

[9]. Verma R. & Das A. (2017). Whats in a URL: Fast feature extraction and malicious URL detection. In 3rd International Workshop on Security and Privacy Analytics, pp. 55–63.

[10]. Zuhair, H., Selamat, A., & Salleh, M. (2015). Selection of robust feature subsets for phish webpage prediction using maximum relevance and minimum redundancy criterion. Journal of Theoretical and Applied Information Technology, 8*l*(2), 188–205.