BVRIT HYDERABAD College of Engineering for Women

Department of Information Technology

Detection of Malicious URLs

Under the Guidance of:

Guide Name: Dr.SL.Aruna Rao

Designation: Professor & HOD

Team- 3

Anuhya Javvaji (18WH1A1240)

A.Meher Gayatri Devi (18WH1A1206)

T. G.S.N Sai Chandana(18WH1A1207)

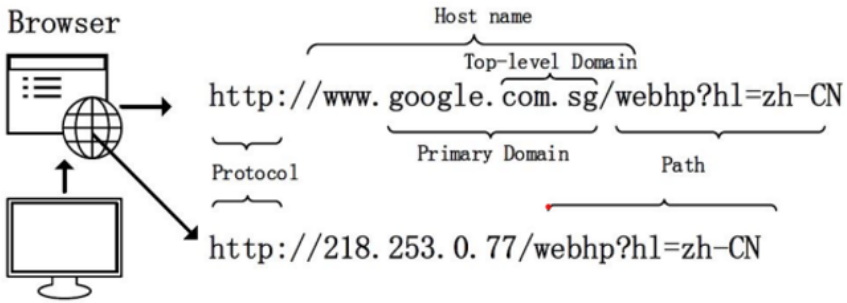
Contents

Abstract

While the World Wide Web has become a killer application on the Internet, it has also brought in an immense risk of cyber-attacks. The more we grew in technology over internet the more we get exposed to malicious data which led to cyber-attacks. With the rise in the underground Internet economy, automated malicious programs popularly known as malwares have become a major threat to computers and information systems connected to the internet. Malicious URLs have been widely used to mount various cyber-attacks including spamming, phishing and malware. We train a machine learning model that takes a dataset, then distinguish benign, phishing, malware and defacement urls.

Introduction

- Currently, the web is a widely used platform that supports an increasing number of daily activities. However, the security of web applications is, unfortunately, becoming a serious problem as web-based services become increasingly prevalent.
- URL is the abbreviation of Uniform Resource Locator, which is the global address of documents and other resources on the World Wide Web.
- Comparing to the traditional detection approaches, machine learning is a novel and flexible method to detect intrusions in the network, it is applicable to any network structure.



Problem Statement

Many web applications suffer from various web attacks due to the lack of awareness concerning security. Malicious URLs host unsolicited content (spam, phishing, drive-by downloads, etc.) and lure unsuspecting users to become victims of scams (monetary loss, theft of private information, and malware installation), and cause losses of billions of dollars every year. It is imperative to detect and act on such threats in a timely manner. Therefore, it is necessary to improve the reliability of web applications by accurately detecting malicious URLs.

Objective

To address the detection of malicious URLs and evaluate the performance of several well-known machine learning classifiers. We adopt a public dataset from Kaggle comprising of URLs to train the model.

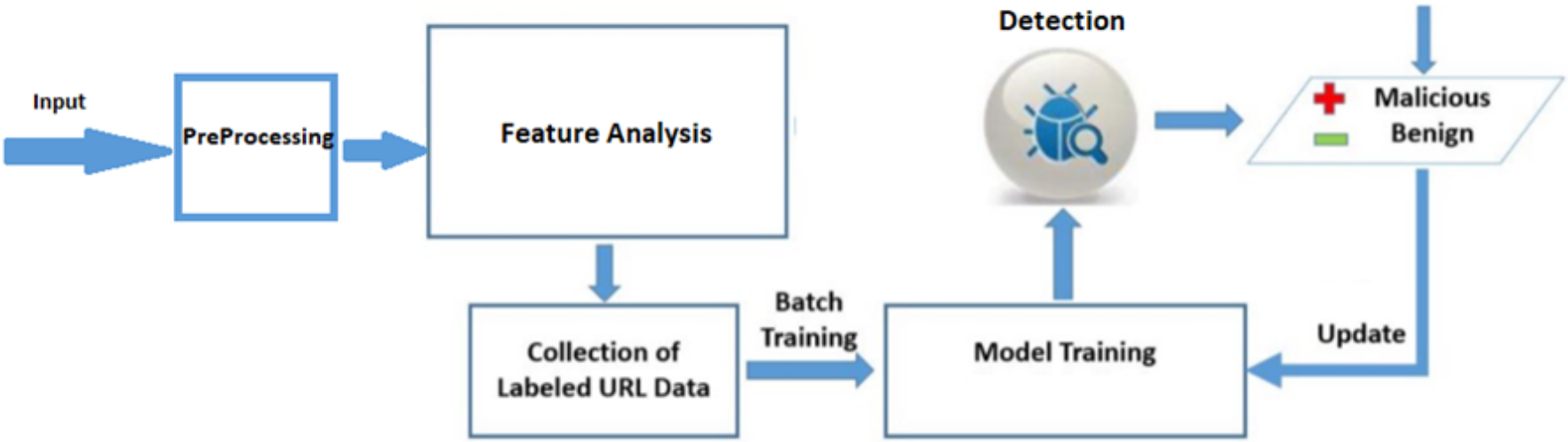
Literature Survey

S.No	Author	Article Title	Journal Name	Year	Observations
1.	J. K. Chahal, V. Gandhi.	KAS-IDS: A Machine Learning based Intrusion Detection System	2021 6th International Conference. IEEE	2021	This paper proposes a technique called KAS-IDS, i.e. K-Means and Adaptive SVM based Intrusion Detection System and discuss the accuracy of these methods.
2.	R. Chiramdasu, G. Srivastava, S	Malicious URL Detection using Logistic Regression	IEEE Conference Publication	2021	An ML model is implemented using Logistic Regression to detect malicious URLs and evaluated against traditional malicious URL models.
3.	Jatin Acarya, Anshul Chaudhary	Detecting Malware, Malicious URLs and Virus Using Machine Learning and Signature Matching.	IEEE Conference Publication	2021	Using Tokenizer domains and sub domains are categorized and converts the text into the weighted URL. These values are used to predict if the URL is safe to visit or not.
4.	Tariro Manyumwa, Phillip Francis Chapita	Towards Fighting Cybercrime: Malicious URL Attack Type Detection using Multiclass Classification	International Conference Publication.	2020	This paper discuss the performance of the models by training the algorithms on 126 983 URLs from benchmark datasets and all four learners XGBoost, AdaBoost, LightGBM, CatBoost returned an overall accuracy above 0.95.

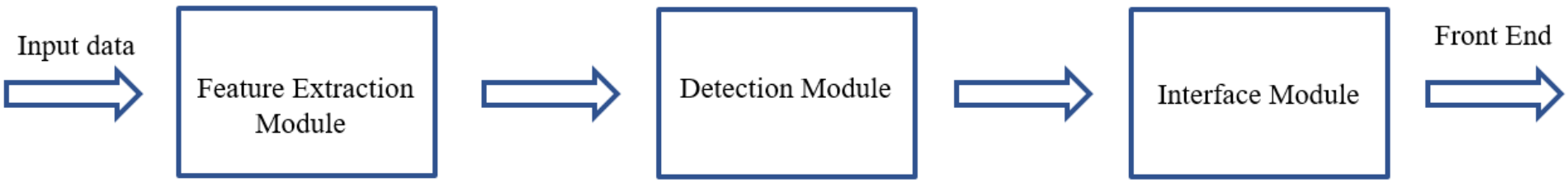
S.No	Author	Article Title	Journal Name	Year	Observation
5.	Baojiang Cui , Shanshan He , Xi Yao and Peilin Shi	Malicious URL detection with feature extraction based on machine learning	International Journal of high performance Computing and Networking.	2018	As keyword matching technique is not adaptive, statistical analyses based on gradient learning and feature extraction are discussed in this paper.
6.	R. Kumar, X. Zhang, H. A. Tariq and R. U. Khan	Malicious URL detection using multi-layer filtering model	2017 14th International Computer Conference	2017	This paper propsed a multi-layer model for detecting malicious URL. The filter can directly determine the URL by training the threshold of each layer filter when it reaches the threshold
7.	Adrian Stefan Popescu, Dumitru Bogdan prelicean	A Study on Techniques for Proactively Identifying Malicious URLs	IEEE Conference Publication	2016	This paper focus on the usage of different machine learning techniques and unsupervised learning methods for detecting malicious URLs with respect to memory footprint.

S.No	Author	Article Title	Journal Name	Year	Observation
8.	M. N. Feroz and S. Mengel	Phishing URL Detection Using URL Ranking	IEEE Conference Publication	2016	Clustering technique is performed on the entire dataset. The paper discusses about classifying URLs automatically based on their lexical and host based features.
9.	S. B. Rathod and T. M. Pattewar	A comparative performance evaluation of content based spam and malicious URL detection in E-mail	IEEE Conference Publication	2016	Using data mining approach like supervised classification which improves the systems accuracy and detects more amount of spam and malicious URLs.

Work Flow


Input
PreProcessing
Feature Analysis
Detection
Malicious
Benign
Collection of
Labeled URL Data
Batch
Training
Model Training
Update

Modules


Input data
Feature Extraction Module
Detection Module
Interface Module
Front End

Dataset

- **Malicious phish.Csv – 641119 Records**
- DataSet Categorized into:
  - Benign
  - Phishing
  - Defacement
  - Malware
- Number of Columns - 2
- Classification of records in dataset:
  - Benign - 66%
  - Defacement - 15%
  - Other(126631) - 19%

References

- https://ieeexplore.ieee.org/document/5675808
- https://ieeexplore.ieee.org/document/9609402
- https://ieeexplore.ieee.org/document/9610045
- https://arxiv.org/abs/2105.13435
- https://ieeexplore.ieee.org/document/9456440
- https://ieeexplore.ieee.org/document/9396014
- https://medium.com/cuelogic-technologies/evaluation-of-machine-learning-algorithms-for-intrusion-d


Thank
you