Artificial Intelligence (CS-351)

Project Report

Naqi Secure URL Scanner

Submitted by: Syed Muhammad Naqi Raza

Registration No: 2022574

Faculty: Cyber Security

On my honor, as student of Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, I have neither given nor received unauthorized assistance on this academic work.

Submitted to: Dr.Usama Janjua

Submission Date: 11/12/2024

Contents

1 Executive Summary

Naqi Secure Scan is a hypothetical URL scanner that categorizes and recognizes threatening sites. To achieve its goal the system has adopted a Multilayer Perceptron Neural Network (MLP), genetically optimized using G**enetic Algorithms (GA)**, to decide whether a specified URL is malicious or safe. This tool becomes a great security weapon that helps to protect users from such internet threats as phishing and other unsafe website attacks. The perfectly programed tool can be embedded in web applications while individuals and organizations can also use it to protect their online space.

2 Introduction

Due to increased incidents of cyber threats, the usage of malicious URLs has evolved and became famous for spreading worms, viruses, Trojan horses, and phishing scams. Naqi Secure Scan tool uses machine learning with an emphasis on algorithms so as to identify such threats based on the structure and attributes of the URLs. In the genetic algorithms, it has been enhanced for better accuracy in scenarios meant for classification.

The web application **Naqi Secure Scan** categorizes the URLs according to the extracted features they contain including domain length, path length presence of IP addresses, and particular characters to mention but a few. This enables a user to easily ascertain the security level of the site and act appropriately to protect himself.

The web application **Naqi Secure Scan** classifies URLs based on extracted features such as domain length, path length, presence of IP addresses, and specific characters in the URL, among others. This allows users to quickly assess the safety of a website and take appropriate action to protect themselves.

3 Problem Statement

Phishing attacks, malware attacks, and other cyber threats delivered through malicious URLs are on the rise and have a high impact probability on both persons and companies. Some of the existing security methods do not detect newly created risky sites because the threats are regularly evolving at the moment. To this end, Naqi Secure Scan represents an effort to develop a scalable and accurate classification of URLs, with integrated machine learning algorithms for URL categorization in real time.

4 Objectives

The purpose of this project is to develop Naqi Secure Scan, a tool for detecting malicious URLs using a machine learning model optimized with genetic algorithms. This project aims to evaluate the effectiveness of machine learning techniques in cybersecurity, specifically in identifying potentially harmful URLs. The primary objective is to enhance awareness of online security risks and provide a tool for users to safeguard themselves against malicious websites. The detailed objectives of this project are as follows:

- To implement a **Malicious URL Detection Model** using a **Multilayer Perceptron Neural Network** optimized through **Genetic Algorithms**.
- To evaluate the efficiency of the model in accurately classifying URLs as safe or malicious, based on extracted features.
- To integrate the model into a **Streamlit web application**, providing an intuitive user interface for real-time URL classification.
- To explore the use of **genetic algorithms** for hyperparameter optimization to improve the accuracy and performance of the model.
- To promote **cybersecurity awareness** by offering a practical tool for users to detect malicious URLs and protect themselves from potential online threats.
- To ensure that the project adheres to **ethical guidelines** and can be used in a controlled, educational, and non-malicious context.

5 Methodology

5.1 Dataset and Feature Extraction

The model is trained on a dataset of both benign and malicious URLs. Various features are extracted from the URLs to aid in classification. Some of the key features include:

- **Hostname Length:** The length of the domain name.
- **Path Length:** The length of the URL path.
- **Special Character Counts:** Counts of special characters such as @, ?, =,
- **Presence of IP Address:** Detection of whether an IP address is used instead of a domain name
- **Number of Directories:** The number of directories in the URL path.
- **HTTP/HTTPS Protocol Count:** Whether the URL contains http or https and its frequency.

These features are used to train the neural network model, which then learns the patterns that distinguish malicious URLs from safe ones.

5.2 Model Architecture

The tool exploits a type of deep learning called Multilayer Perceptron (MLP) as the neural network for classification problems. Most of the model optimization in employs the use of Genetic Algorithms (GA) to get the best out of the model.

5.3 Genetic Algorithm Optimization

Genetic Algorithms are applied to evolve and optimize the hyperparameters of the MLP neural network. The GA helps in selecting the best features and fine-tuning parameters such as the learning rate, number of layers, and neurons, resulting in a more accurate and robust model.

5.4 Model Training and Evaluation

The model is trained using a dataset split into training and testing sets (80:20 ratio). The training process involves 10 epochs with a batch size of 256. The model's performance is evaluated based on accuracy, precision, and recall, ensuring that it effectively detects both benign and malicious URLs.

6 Results and Discussion

6.1 Model Performance

The neural network model achieves high accuracy in classifying URLs. The use of genetic algorithms significantly improved the model's performance by optimizing hyperparameters and ensuring better generalization. The model successfully detects malicious URLs with a high degree of confidence, making it an effective tool for online safety.

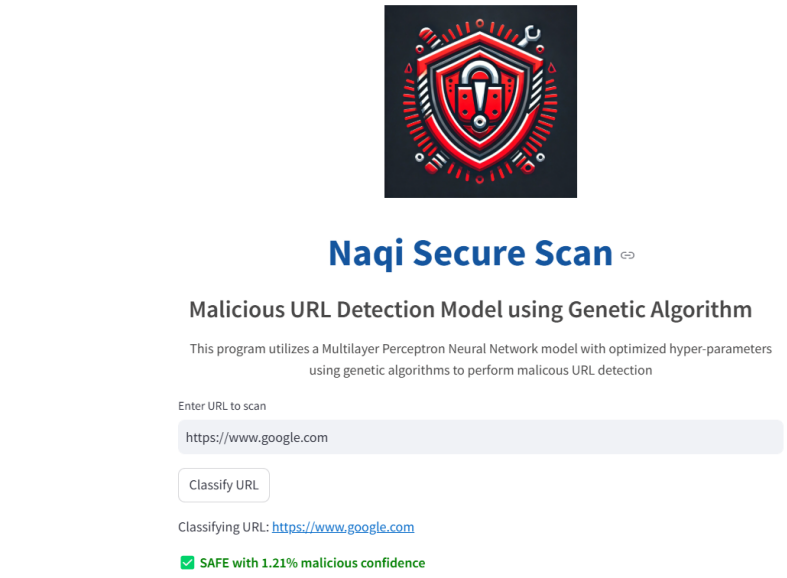
6.2 Real-Time Application

Naqi Secure Scan can be implemented in real-time scenarios, where users can input URLs to get immediate feedback on whether they are safe or malicious. This makes the tool highly valuable for individual users, security systems, and web administrators.

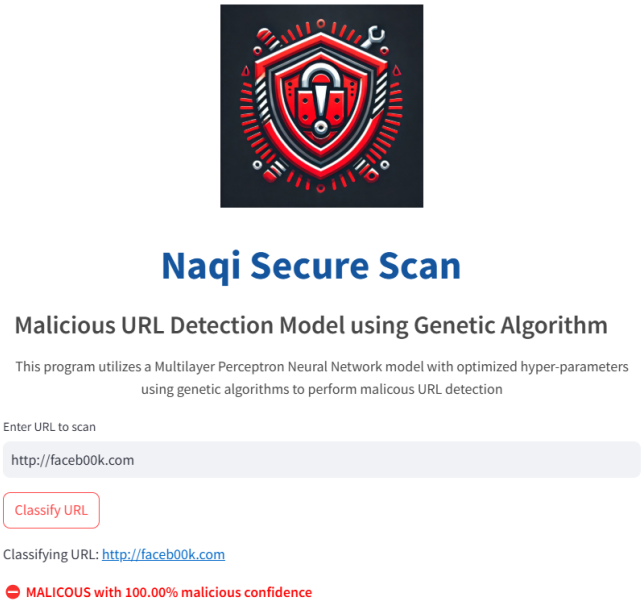
6.3 User Interface

The system is designed with a user-friendly web interface powered by Streamlit. Users can input URLs and instantly receive classification results. The interface is clean and responsive, providing clear and easy-to-understand outputs such as:

- **SAFE:** Indicates that the URL is safe to visit.



- **MALICIOUS:** Indicates that the URL is harmful and may pose a threat.



6.4 Evaluation Metrics

The model is evaluated using metrics like accuracy, precision, recall, and F1-score, ensuring that it not only classifies URLs correctly but also minimizes false positives and false negatives.

```
Accuracy: 1.00
Precision: 1.00
Recall: 1.00
F1-Score: 1.00
```

7 Conclusion:

The Naqi Secure Scan tool represents a significant advancement in malicious URL detection. By combining neural network-based classification with genetic algorithm optimization, the tool is both efficient and accurate. It is designed to protect users from cyber threats by providing a reliable method for identifying dangerous URLs. The system is highly customizable and can be easily integrated into various security applications.

8 Code and Resources

The Naqi Secure Scan tool is implemented using Python and leverages popular libraries such as TensorFlow and Streamlit for web deployment. The full code, including the model training, evaluation, and web application, is available in the GitHub repository.

9 References

1. Angelinatsuboi. (2024). Bio-Cybersecurity: Using Genetic Algorithms to Detect Malicious URLs. Medium Article.
2. Deepesh Mhatre. (2023). Phishing Attack Domain Detection. GitHub.
3. Suryansh S. (2022). GAs and NNs. Towards Data Science.