An Intelligent Phishing Website Detection System Using URL Features, Webpage Content Analysis, and Machine Learning

Module Code and Title

7COM1039-0206-2025 - Advanced Computer Science Masters Project

Name and Student ID

Hyndavi Yasarapy 24090531

Aim of the Project

The goal of this project is to combine URL-based features, webpage content analysis, and machine learning models to create, implement, and assess an intelligent system that can identify phishing websites in real time. The technology will be implemented as a browser plugin to provide practical, real-time phishing detection for end users.
Additionally, this research will examine and contrast three methods of detection:

- Conventional rule-based identification
- Detection using machine learning
- A hybrid system that combines the two methods

Research Question

How does a machine learning–based phishing detection system compare to rule-based and hybrid techniques in terms of detection accuracy, false positives, and real-time performance when analysing URL and webpage content?

Hypothesis:
In terms of phishing detection accuracy and dependability, a hybrid strategy that blends rule-based heuristics with machine learning categorisation of URL and site attributes will perform better than systems that are either rule-based or solely ML-based.

Objectives

The key objectives of this project are:

- To study existing phishing detection techniques and identify their limitations.
- To extract meaningful features from:
  - URLs (length, symbols, domain age, redirections, etc.)
  - Webpage content (HTML structure, forms, scripts, text patterns).
- To design and implement:
  - A rule-based phishing detection system
  - A machine learning classification model
  - A hybrid detection model
- To develop a real-time browser extension that integrates the detection system.
- To evaluate and compare the performance of the three approaches using standard metrics.
- To analyze trade-offs between detection accuracy, speed, and false positive rates.

Short Description

One of the most prevalent types of cybercrime is phishing, in which hostile websites pose as trustworthy ones in order to obtain critical information and user passwords. Many of the detection systems in use today rely on either basic rule-based checks or blacklists, which are frequently useless against recently developed phishing websites.
This project presents an intelligent phishing detection system that uses:

- Analysis of URL features, such as length, special characters, IP-based URLs, and suspicious tokens
- Webpage content examination (HTML tags, JavaScript behaviour, login forms, text similarity)
- Machine learning classification to detect phishing patterns
- A browser plugin to provide real-time protection to users

The study is interesting because it compares rule-based, machine learning-based, and hybrid approaches within the same system and tests them in an actual setting.

Research Methodology

This research focuses on detecting phishing websites, which are fraudulent websites aimed to steal personal information such as passwords or bank data. The idea is to design a system that can identify these risky websites and notify consumers in real time.

The process begins with a review of the literature, which entails researching current techniques for identifying phishing websites. These include simple methods like verifying known blacklisted websites, employing predefined criteria based on suspicious patterns, and more complex ways that use machine learning. This aids in determining what is effective and what needs improvement.

Next, a dataset is generated by gathering website links from public sources. Some links are phishing websites, while others are safe and real. In order for the system to distinguish between these relationships, they are explicitly identified.

The system then looks at the features of the webpage. For example, it looks at factors like how long the website link is, whether it employs weird symbols, whether it has a login form, or whether it contains suspicious code. These characteristics aid in determining whether a website appears secure or dangerous.

The project then develops **three detection systems**:

1. a rule-based system using simple rules,
2. a machine learning system that learns from data, and
3. a hybrid system that combines both approaches.

Finally, a **browser extension** is built that checks websites in real time and warns users if a site is likely to be phishing. The system is evaluated using accuracy and speed to see which method works best.

Citations

1. Mohammad, R. M., Thabtah, F., & McCluskey, L. (2015). Phishing detection using machine learning techniques.
2. Sahingoz, O. K., et al. (2019). Machine learning based phishing detection from URLs.
3. Rao, R. S., & Pais, A. R. (2019). Detection of phishing websites using ML approaches.