Phishing Detection

Gokul R
B.Sc. Information Technology Student
Department of CS&IT, JAIN (Deemed-To-BeUniversity)
gr2k14@gmail.com

Dr. Felix M Philip
Assistant Professor
Department of CS&IT, JAIN (Deemed-To-Be University)
m.felix@jainuniversity.ac.in

Abstract

The phishing email is one of the significant threats in the world today and has caused tremendous financial losses. Phishing is a type of social engineering attack often used to steal user data,
including login credentials and credit card numbers. It occurs when an attacker, masquerading as a trusted entity, dupes a victim into opening an email, instant message, or text message.
Although the methods of confrontation are continually being updated, the results of those methods are not very satisfactory at present. Moreover, phishing emails are growing at an alarming rate in recent years. Therefore, more effective phishing detection technology is needed to curb the threat of phishing emails. So There are many ways to detect these phishing mails nowadays using Machine Learning. so using the phishing mail detector where these links could be tested and then predicted and to detect whether it is a spam or not.

Keywords: Phishing mail, Social engineering, Machine learning

1. Introduction

This project focus on creating a phishing detection tool using Machine Learning. Phishing is a type of social engineering where an attacker sends a fraudulent (e.g., spoofed, fake, or otherwise deceptive) message designed to trick a person into revealing sensitive information.

The tool is trained with Logistic Regression and multinomial algorithm, where we train the detection model with a collection data fed into the dataset and is split into two for training and testing.

After training the model successfully predicts whether the URL is safe (malicious) or not.

The main objective of the project is to provide safety and security for the users by preventing them from spammers and hackers. As there are a lot of spammers around the globe they could send spam links and when users visit these links or sites their banking credentials and personal information could be stolen by spammers and which could lead to lots of problems such as cybercrime.

2. Literature Review

2.1. Phishing detection:

A recent intelligent machine learning comparison based on models content and features

Several fake websites have been developed on the World Wide Web in the last decade to mimic trusted websites, with the goal of stealing financial assets from users and organisations. Phishing is a type of online attack that has cost the online community and various stakeholders hundreds of millions of dollars. As a result, effective countermeasures that can detect phishing are required. Machine learning (ML) is a popular data analysis tool that has recently shown promising results in combating phishing when compared to traditional anti-phishing approaches such as awareness workshops, visualisation, and legal solutions. This article investigates the applicability of machine learning techniques to detect phishing attacks and discusses their advantages and disadvantages. Various types of ML techniques, in particular, have been investigated to reveal the suitable options that can serve as anti-phishing tools. More importantly, we experimentally compare large numbers of ML techniques on real phishing datasets and with respect to different metrics. The purpose of the comparison is to reveal the advantages and disadvantages of ML predictive models and to show their actual performance when it comes to phishing attacks. The experimental results show that Covering approach models are more appropriate as anti-phishing solutions, especially for novice users,
because of their simple yet effective knowledge bases in addition to their good phishing detection rate.

2.2. Intelligent cyber-phishing detection for online

Phishing attacks are becoming more common, causing financial loss and the theft of sensitive information from online services and users. Blacklist-based anti-phishing approaches that use manually verified Unified Resource Locators (URLs) or content-based methods that use heuristics-based machine learning (ML) classifiers have received the most attention. Online deception, on the other hand, is still on the rise. In this paper, we present a novel methodology that combines blacklist-based, web content-based, and heuristic-based approaches, as well as ML algorithms with comprehensive features, to detect phishing attacks more accurately. Extensive evaluation was carried out using evaluation methods (metrics) to measure the proposed method performance based on Adaptive neuro-fuzzy inference system (ANFIS), Nave Bayes (NB), PART, J48, and JRip with features. All of the classifiers achieved greater than 99.33 percent accuracy. PART achieved 99.33% accuracy with 0.006 seconds (secs) speed, which is the best performance. We experimentally demonstrate that the proposed methodology can detect phishing websites with a high accuracy in real-time and generalise well to new phishing attacks. The proposed approach has the best performance compared to related approaches in the field.

2.3. Web phishing detection techniques:

a survey on the state-of-the-art, taxonomy and future directions

More than half of the world's population has been drawn into the cyber world by the Internet. Unfortunately, as the number of internet transactions grows, so does the number of cybercrimes. With the internet's anonymous structure, attackers attempt to deceive end users through various methods such as phishing, malware, SQL injection, man-in-the-middle, domain name system tunnelling, ransomware, web trojan, and so on. Among these, phishing is the most deceptive attack, exploiting end-user vulnerabilities. Phishing is frequently carried out via emails and malicious websites in order to entice the user by posing as a trusted entity. Many anti-phishing techniques have been proposed by security experts. There is currently no single solution capable of mitigating all vulnerabilities. A comprehensive examination of current trends in web phishing detection techniques.

2.4. A predictive model for phishing detection

Many anti-phishing systems are currently being developed to detect phishing content in online communication systems. Despite the availability of numerous anti-phishing systems, phishing persists due to insufficient detection of a zero-day attack, excessive computational overhead, and high false rates. Although Machine Learning approaches have demonstrated promising accuracy rates, the selection and performance of the feature vector limit their effectiveness in detection. An improved machine learning-based predictive model is proposed in this paper to improve the efficiency of anti-phishing schemes. The predictive model includes a Feature Selection Module, which is used to build an effective feature vector. The incremental component-based system extracts these features from the URL, webpage properties, and webpage behavior and presents the resulting feature vector to the predictive model.

3. Material and Methods

3.1. Development

The phishing mail detection is developed using Python programming language and uses several Python libraries such as Numpy, Pandas, Seaborn, Selenium etc. This project was completely made using the Logistic Regression model in machine learning and the testings and predictions were made possible using the train-test split method. Front-End was done using Tailwind Css. Back-End using Django and a dataset of 5 million arrays are used.

3.2. Testing

After developing the website, a trail run has been run to ensure the working of the website. Testing helps to find some bugs and h
ave fixed the bugs for the smooth working of the website.

3.3. Methods

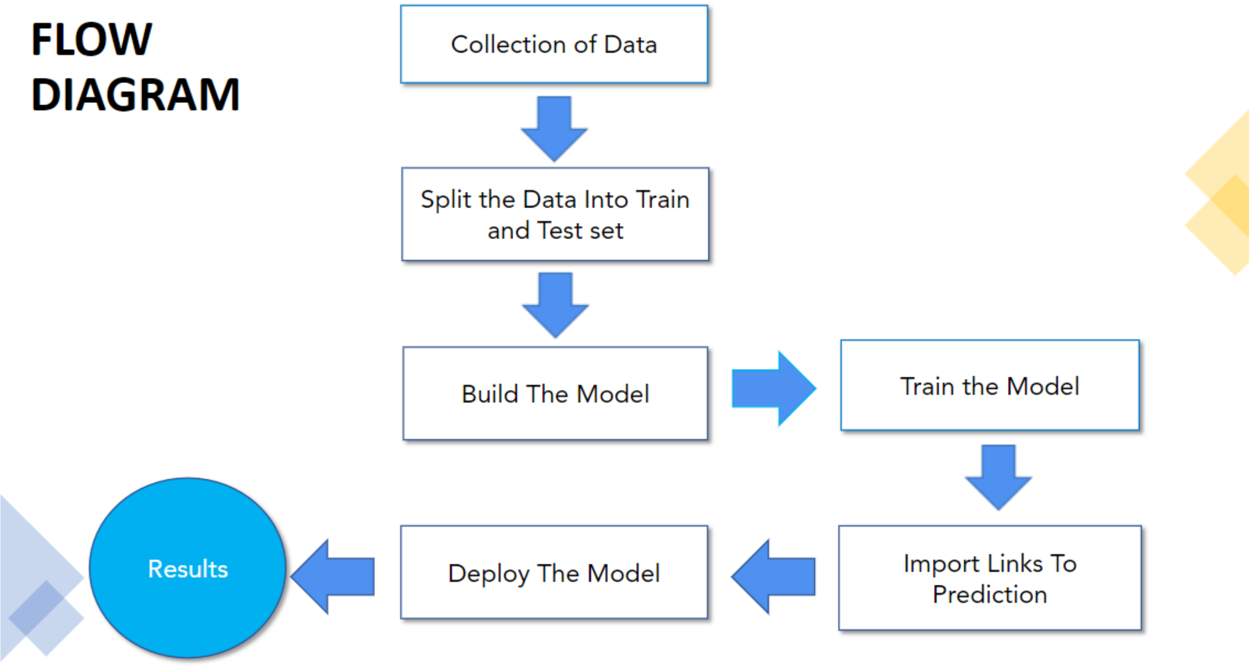
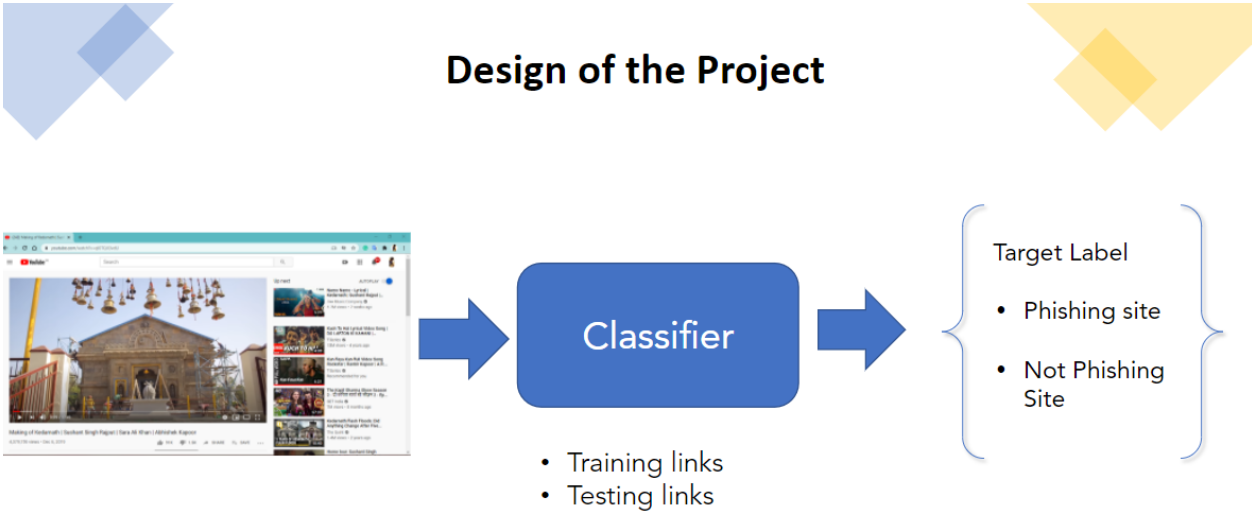


Fig 1. This is the actual working model of the phishing detector



Design of the Project

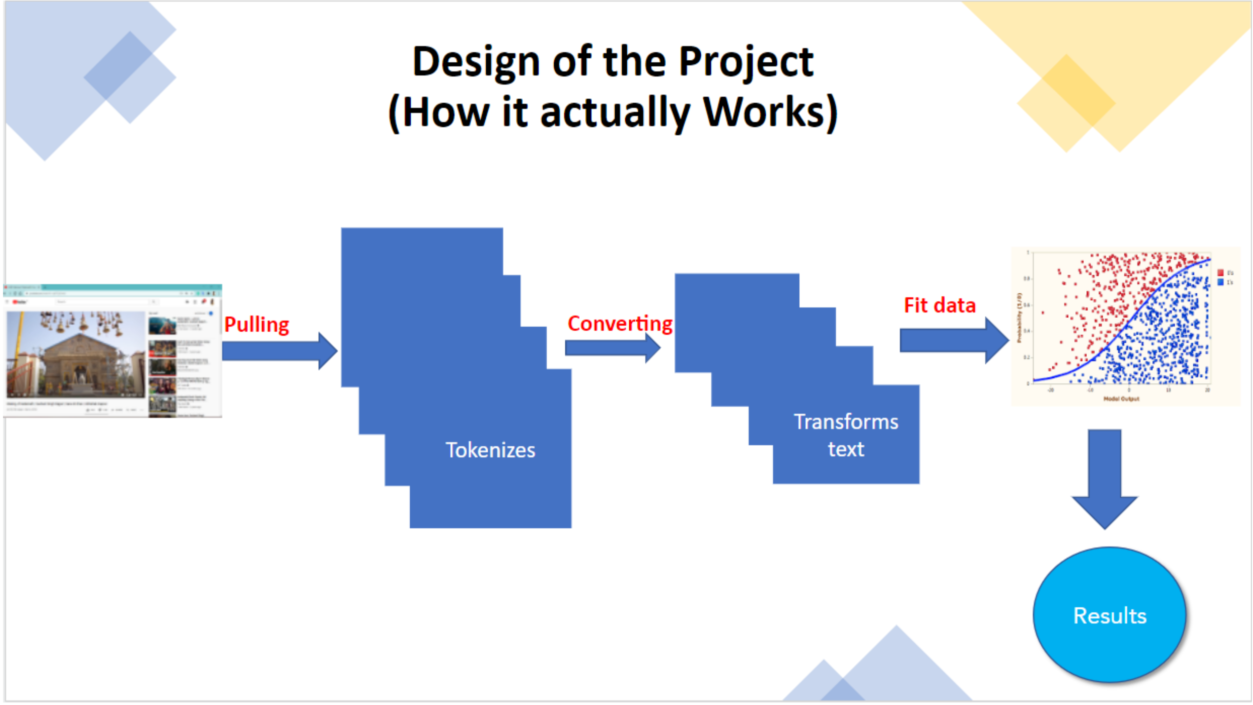
Classifier

- Training links
- Testing links

Target Label

- Phishing site
- Not Phishing Site

Fig 2. This is how the links are tested and predicted using the train-test split method



Design of the Project (How it actually Works)

Pulling

Tokenizes

Converting

Transforms text

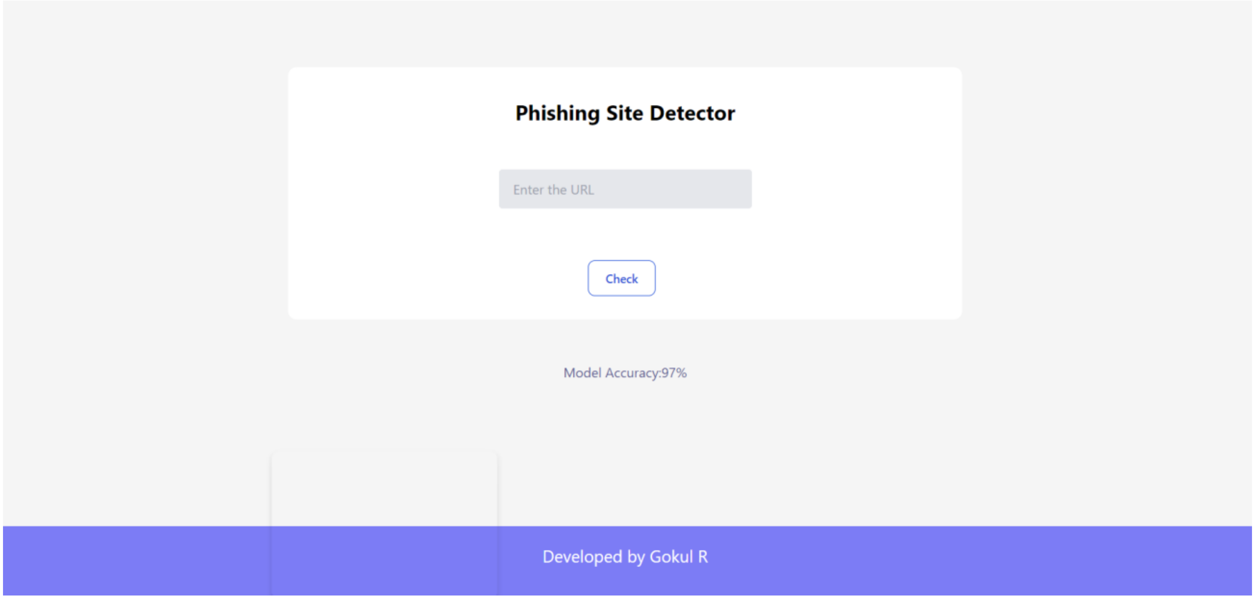
Fit data

Results

Fig 3. This is how the links are pulled and tokenized and tested if it fits the data from the dataset.

4. Result and Discussion

We aim at ensuring the safety and security of the system and user by detecting malicious links, spams and scams using the phishing detector

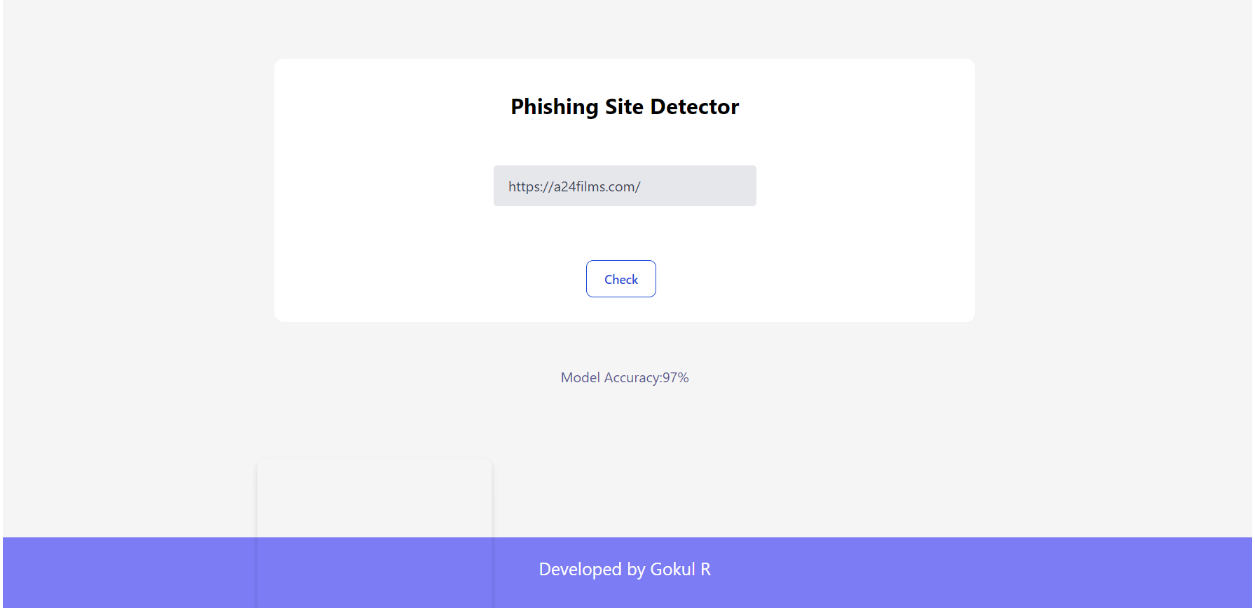


Phishing Site Detector

Model Accuracy:97%

Developed by Gokul R

Fig 1. Home Page



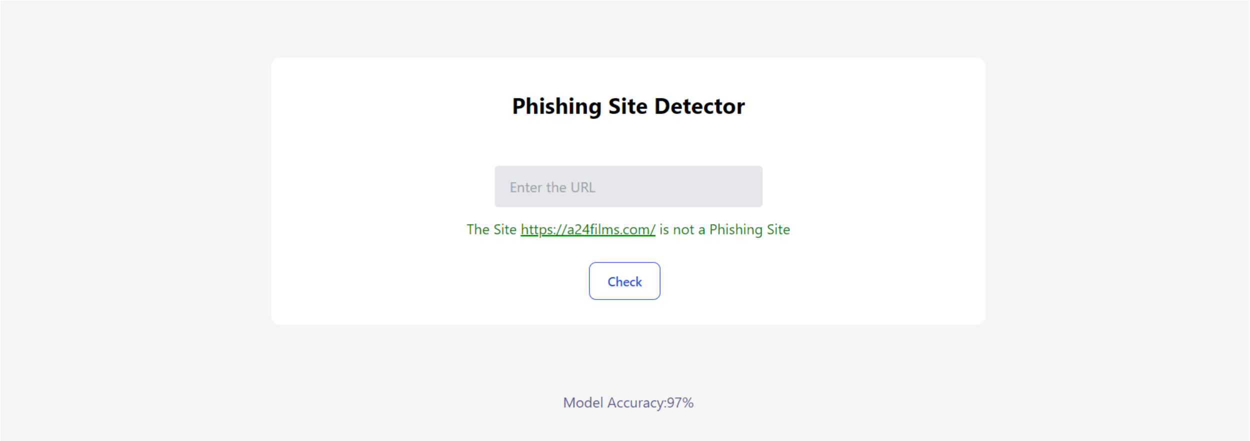
Phishing Site Detector

https://a24films.com/

Model Accuracy:97%

Developed by Gokul R

Fig 2. Checking whether the link is a spam or not



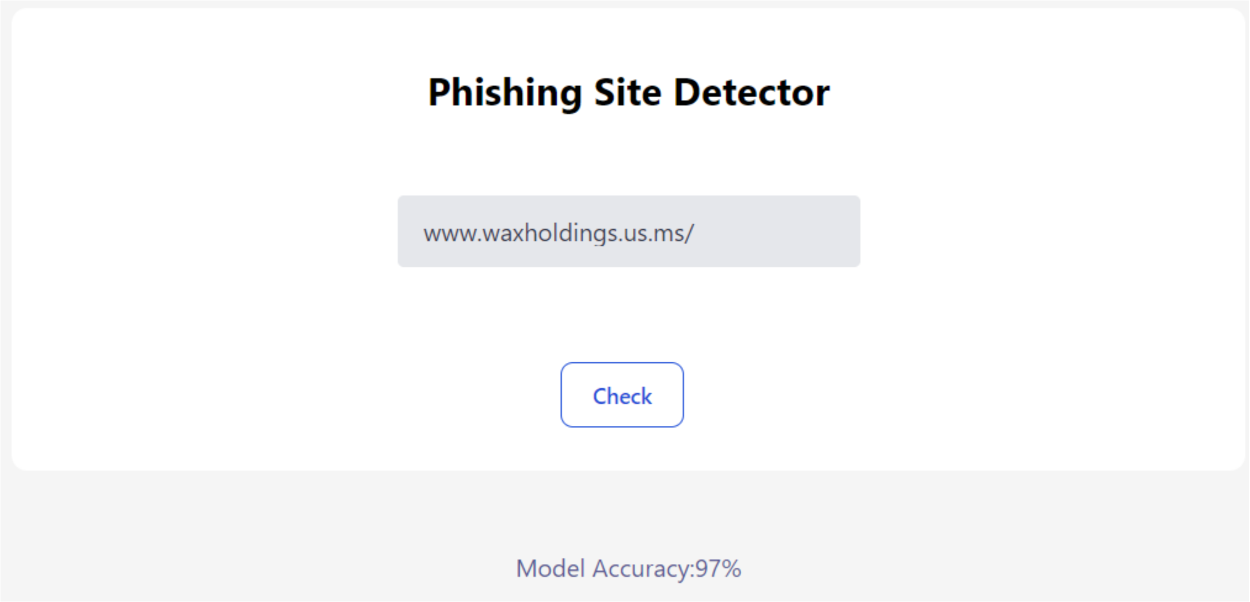
Phishing Site Detector

Enter the URL

The Site https://a24films.com/ is not a Phishing Site

Model Accuracy:97%

Fig 3. The link is tested and predicted and it is not a spam or phishing link and it is shown in green

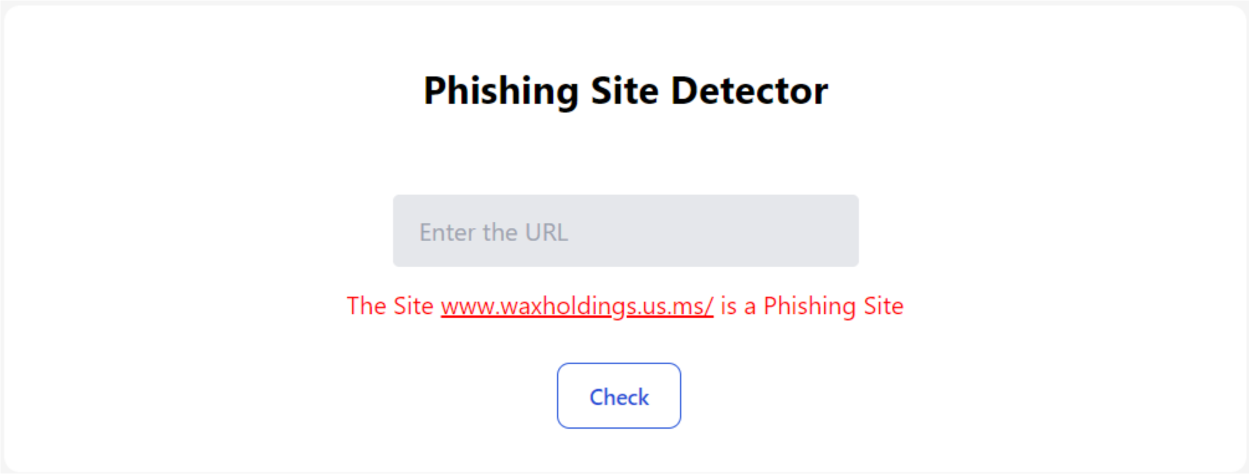


Phishing Site Detector

www.waxholdings.us.ms/

Model Accuracy:97%

Fig 4. Another link is submitted to find whether it is phishing link or not


Phishing Site Detector

The Site www.waxholdings.us.ms/ is a Phishing Site

Fig 5. The link is checked and it is a phishing link hence it is shown in red

5. Conclusion

This project aims to enhance detection method to detect phishing websites using machine learning technology.

We achieved 97.14% detection accuracy using Logistic Regression Algorithm with lowest false positive rate. Also result shows that classifiers give better performance when we used more data as training data.

In future hybrid technology will be implemented to detect phishing websites more accurately, for which Logistic Regression Algorithm of Machine Learning technology and Train-Test Split method will be used.

6. Reference

[1]. Abdelhamid, N., Thabtah, F., & Abdel-jaber, H. (2017, July). Phishing detection: A recent intelligent machine learning comparison based on models content and features. In 2017 IEEE international conference on intelligence and security informatics (ISI) (pp. 72-77). IEEE.

[2]. Barraclough, P. A., Fehringer, G., & Woodward, J. (2021). Intelligent cyber-phishing detection for online. Computers & Security, 104, 102123.

[3]. Vijayalakshmi, M., Mercy Shalinie, S., Yang, M. H., & U, R. M. (2020). Web phishing detection techniques: a survey on the state-of-the-art, taxonomy and future directions. Iet Networks, 9(5), 235-246.

[4]. Orunsolu, A. A., Sodiya, A. S., & Akinwale, A. T. (2019). A predictive model for phishing detection. Journal of King Saud University-Computer and Information Science