

Palestine Technical University- Kadoorie

faculty of applied science

Department Applied Computing

**developing a Phishing Detection Tool**

**Prepared by:**

Ahmad Saleh-202012191

Nagham Maraheen-202012658

Muhannad Tomeh-202011744

**Supervisor:**

Dr..Eman Droubi

**2022/2023**

**Abstract**

With the wide spread of the Internet and its entry into all areas of life, until our personal information and bank accounts became stored on the Internet, and with research and statistics, it was found that the greater the use of the Internet, the greater the penetration rate around the world. The Internet has become a place where thieves and intruders aspire to intrude on others and break into their privacy or cause damage or theft.

Therefore, there has become an urgent need for us to preserve our privacy and personal information and protect ourselves from exposure to danger and penetration.

Therefore, in this project, we decided to develop a mechanism, which is a website. This site provides a service that can detect dangerous sites and protect them from the danger they cause from breaching the privacy of the user or the browser on the Internet in general.

In this project, we will build a machine learning model for prediction

if website is phishing or not start fetch data then preprocess data by applying feature scaling then create neural network (mlp multilayer perceptron) and set number of layers and each layer has specific number of neurons

And the last thing that determines the number of repetitions.

Therefore, this project is very necessary for everyone who uses the Internet, especially if he has little experience in the Internet, so it is easy to penetrate, and this service can be used by anyone in the world, as it is restricted to a specific region.

Therefore, this project will help many people, official and unofficial institutions and companies to ensure the safety of links to websites, with an accuracy rate of more than 90% to protect them from hacking.

By making it a reference for them in case they encounter any external links that may harm them.

**Contents:**

1. **Introduction** 
   1. Introduction
   2. Problem statement
   3. Project objectives
   4. Project benefits
   5. Project methodology
   6. Research plan (working plan)
   7. Project constraints (research limitations)
   8. Summary and recommendations
2. **System Analysis** 
   1. Introduction
   2. A brief history of the organization
   3. Project implementation options
   4. The proposed system
   5. System requirements
   6. Feasibility study
      1. Economic feasibility
      2. Technical feasibility
      3. Legal feasibility
      4. Schedule feasibility
   7. Project added values
   8. Project management
   9. Summary and recommendations
3. **System Functional Requirements** 
   1. Introduction
   2. Context or Use-Case Diagrams
   3. Functional requirements description
   4. Non-functional requirements description
   5. Summary and recommendations
4. **System Design and Development**
   1. Introduction
   2. Class diagram
   3. Sequence diagrams
   4. Entity Relationship Diagram (ER-model)
   5. Activity diagrams
   6. System interface (input/ output design)
   7. Summary and recommendations
5. **Coding and Implementation** 
   1. Introduction
   2. Coding programming languages
   3. Database system
   4. Establishment of Development Environment
   5. Stored procedures
   6. Summary and recommendations
6. **System Testing** 
   1. Introduction
   2. Coding programming languages
   3. Database system
   4. Establishment of Development Environment
   5. Stored procedures
   6. Summary and recommendations
7. **Conclusions and Future Works** 
   1. Conclusions
   2. Future works
   3. Closing remarks

**List of figures:**

Figure 1.1. Cyber Phishing websites

Figure 3.2.1 context diagram.

Figure 3.2.2 use case.

Figure 4.2 class diagram

Figure 4.3 Sequence diagrams

Figure 4.4 ER-Model.

Figure 4.5 Activity diagram.

Figure 4.6.1 main page

Figure 4.6.2 tips page.

Figure 4.6.3 check page

Figure 4.6.3.1 check page (not phishing)

Figure 4.6.3.2 check page(phishing)

**List of Tables:**

Table 2.5.1 Functional System requirement

Table 2.5.2 Non-Functional System requirement

Table 2.6.1 Technical Feasibility (Hardware)

Table 2.6.2 Technical Feasibility(software)

Table 2.6.4 Schedule feasibility

Table 3.3 Functional Requirements

**Chapter** **1**

**1.1 Introduction:**

Nowadays, with the spread of the Internet and websites, and the almost daily use of websites, some people are taking advantage of the spread of links in a large way and take advantage of people's inexperience with security on the Internet! They are creating malicious fake links and deceiving people in them, aiming to seize the victim's personal information for personal reasons, either for extortion, stealing money, or just sabotage.

A number of subscribers and users of mobile phones are exposed to an increasing number of fraud attempts in recent times, due to the penetration of their devices as a result of entering malicious or fake links that arrive through WhatsApp messages, text messages (sms) or other electronic programs in which they ask you for an amount of money or some assistance. Then you are surprised that that person's phone device has been hacked and that the person who sent that message is the hacker or the hacked organization of the phone or device.

In this following figures , examples of these malicious link :



Figure 1.1. Cyber Phishing websites [1]

**1.2 Problem Statement:**

Phishing is one of the easiest types of cyberattacks for hackers to do and one of the easiest types of cyberattacks to fall for too! This is often done to steal user data such as login details, credit card numbers, and other personal information. The attacker appears as a trusted facility to trick the person into clicking on a specific link or opening an email or WhatsApp message. And some links come short so that you cannot know their source.

Malware may be installed if the recipient clicks on the link, the system may be frozen (a type of ransomware attack), or sensitive information may be exposed due to a security vulnerability in the operating system.

Statistics reveal that more than 4.1 billion people are email users and about 3 million emails are sent per second which means that more than 50% of the world’s population uses email and the number is expected to rise to 4.5 billion users by 2024. These The high numbers encouraged criminals to focus their attacks on email rather than other sources. And with COVID-19, many organizations have ordered their employees to work remotely for their safety, and this has led to many people relying on email communication a lot, and with employees receiving a large number of messages daily, this makes them more vulnerable to cyberattacks. And when it comes to phishing, all it takes is one click to lose it all.

Phishing attacks, in their most common form, are email messages that urge mail recipients to take action, usually to achieve one of two goals:

- Trick the user into sharing their personal information

- Tricking the user into downloading malicious software

Once you give the software access, hackers can access your bank account, steal your identity, or make fraudulent purchases in your name.

A common example of this deception is that you receive in your e-mail

A fake letter in the name of your bank containing a request

The bank requires you to log in to your account and update your personal information

Clicking on the link in the message, you will be directed to a site

a different electronic mail, the appearance of which is identical to the website of that bank,

And when you enter the username and password as requested

From you, this data is sent to the hacker's address, not to the bank.

Over the past few years, email scams have increased by over 400%. The growth and success of email phishing scams has also led to the frenzy of these scams. We will discuss more about it below:

SMiShing

As the name suggests, SMiShing is similar to phishing email, but it tricks users via text messages. All in all, many people are aware of email phishing, but the level of awareness about SMS fraud is lower than the aforementioned pattern, which increases the possibility of people falling victim to a phishing scam.

Spear Phishing

Spear phishing uses the same methods of tricks that we mentioned above, but the difference is that it targets a specific person. You might see a series of emails designed to entice you to take action. Phishing attacks can also target multiple messaging platforms.

**1.3 Project objectives:**

The following points summarize the main objectives of this project:

\*Investigating the nature of cyberattacks content and its types.

\*Developing a websites for detecting cyberattacks and phishing systems

\* Conducting extensive experiments for evaluating the proposed model.

**1.4 Project benefits:**

**- For users:**

Building a tool for Detecting cyberattacks in websites which may be added to Facebook officially as adds-on may play a role in:

1-Making phishing email or websites detection easy for users.

2-Making internet safer and reducing attacks.

3-Protect people from attackers and save they especially data

**- For developers:**

Enhancing and expanding general knowledge about cyberattacks and improving the experience of developing useful data mining-based programs.

Learning about the many algorithms that are used to solve a problem and which are the building blocks of the advanced digital world we see today. It is an important concept that must be understood, because in machine learning, learning algorithms - not computer programmers - create the rules.

**1.5 Project methodology:**

The software development methodologies are very important which are mostly used for various software development projects. Moreover, all these methodologies work well in certain projects depending upon the nature of the project. None of these methodologies are foolproof as each has its pros and cons. The basic purpose of these methodologies is to provide smooth software development according to the project requirements [3].

Below are some of the methodologies that are primarily used software development methodologies.

In our project, We will adopt **the waterfall model** because :

1 - requirements will be very understood so we don’t need to send to customer proto type for each small update

2 – It is simple and easy to understand and use.

3 - The project status is more easily measured based on a complete schedule and resource plan.

4 - Risk is zero or minimum

**1.6 Research Plan :**

In this project we are going to build machine learning model to predict

if website is phishing or not starting with fetching data and then preprocess the data by apply feature scaling then create neural network (mlp multilayer perceptron) and set number of layers and for each layer set number of neurones  
and last thing set number of iterations .

**1.7 Project constraints:**

There are some big challenge for any machine learning project

and in this project we need to mention some of them :

**1 – Accuracy Of Model :**

In any machine learning model all the time we need to optimize model

to get high accuracy

So we need to maximum accuracy percentile.

**2 – Model Efficiency:**

If Model take very long Time to predict weather website is phishing or not, what do we benefit?

So in project we the performance and effectiveness of the algorithm used have been taken into account .

**1.9 Summary and Recommendation:**

In this chapter, we have discussed the introduction to this project, the problem statement of the project and its solution, and we have mentioned the objectives of the project and how it will benefit both developers and users. We have mentioned some of the mythologies and discussed how the work plan will be done. Finally, we talked about project constraints and limitations.

**Chapter 2**

**{System Analysis}**

**2.1 Introduction :**

In this chapter, we will cover the important topics. will explain what the system analysis is in general mention and discus previous researches that the same our project , explain the project Implementation options and specify the one that we will choose , explain what it is the system requirements , then finally, we will show the feasibility study of this project.

**2.2 Related work:**

According to APWG phishing attack trends reports [13, 14], the number of phishing attacks observed by APWG and its members grew through 2020, doubling over the course of the year. Phishing are spread via e-mail, SMS, instant messaging, social networking etc., but e-mail is a popular way to carry out this attack. The phishing email can lead to financial loss. Attacker always sending email tends to make user believe that they are communicating with trusted entity and deceive them into providing personal credentials in order to access service, such as credit card numbers, account login credential or identity information. In 2019, 293.6 billion emails were sent and received daily. This includes billions of promotional emails sent by merchants every day.

In Sect. 1 of this paper, we applied machine learning on three different data sets where the first two datasets depend on multi features and the third one depends on text feature only. Section 2 we review the Related Work of classifiers used in detecting phishing emails, in Sect. 3 we mentioned the targeted victims in phishing. The methodology that has been followed to do this research has been introduced in Sect. 4. Section 5 presents the experiments for classifying Phishing Email Using Machine Learning, Finally, the work is concluded in Sect.

**2.3 Project implementation options :**

There are several project implementations options, like Desktop Applications, Web Applications, and Mobile application.

**- Desktop Applications**

Desktop applications are stand-alone application which runs on system and laptops.

**- Web Applications**

This type of software application is used through the internet via a web browser. A web browser allows you to access the app and its content and also runs all the scripts responsible for its features. What differentiates a simple static web page from the web application is interactivity. They often allow you to create, edit, or manipulate data and content

**- Mobile application**

It’s a type of application designed to run on a mobile device, which can be a smartphone or tablet computer.

In our project, we are using Python to develop a Web application. Python provides many useful features which make it popular and valuable from the other programming languages. It supports object-oriented programming, procedural programming approaches Additionally many libraries that support data science and machine learning. Web applications are easy to use for anyone and much enjoyable than mobile application.

**2.4 The proposed system :**

The main objective of the project is to discover phishing sites that help the general public, companies and institutions and protect them from those trap sites that, if they deal with them, cause them harm and expose them to theft or damage to their sites .

In this project we are going to implement a website for phishing detection . the tool will have information about phishing that admin give it as dataset from Kaggle contain alot of website and its result phishing or not , and tool be trained about phishing websites then will be ready to act a new website which checked in future .

and tool use many algorithims like Logistic.Regression and SVM Support Vector Machine ,Naive Bayes

K Nearest Neighbors , Neural Network(MLP), and Multilayer Precptron that tool depend on its to take decesion .

And the tool will have edit text which allow the user to enter a URL .

Then based on a trained ML model it will check if the URL is phishing or not .

**2.5 System requirements:**

The Functional requirements explain how the system must work, while the nonfunctional requirements explain how the system should perform.

**The Functional requirement of the system:**

Table 2.5.1 Functional System requirement

|  |
| --- |
| Data Preprocessing 🡪 Clean Data And Convert Sentence To Vectors |
| Rate Model 🡪 Rate Model From 5 And Write Feedback |
| Import Data 🡪 Extract Data From Kaggle |
| Classification 🡪 Classification Weather URL Phishing OR Not |
| Retrain Model 🡪 Train Model From Imported Data To Improve Results |
| Show Tips🡪 Show Tips For User For Increase Awareness For User . |
| Invite Friend🡪Invited Friend Via Facebook , Linkedin , … |
| Show Results🡪Show Precision , Recall And F1 Score For Model |
| Show Rates & Feedback 🡪 Show People Rates For Model |

**Non Functional Requirements:**

Table 2.5.2 Non Functional System requirement

|  |  |  |
| --- | --- | --- |
| **Description** | **Nonfunctional** | # |
| Clear, simple and easy that can be understood and used by anyone who does not have experience in the field | Understandable &Usability | 1 |
| Be available to all users at any time | Availability | 2 |
| Able to give high F1 Score result | Correctness | 3 |
| Ability to perform speedily and execute efficiently | Performance | 4 |
| Ability to cope with errors during execution and cope with erroneous input | Robustness | 5 |

**2.6 Feasibility study:**

**2.6.1 Technical Feasibility :**

**-Hardware:**

Table 2.6.1 Technical Feasibility (Hardware)

|  |  |
| --- | --- |
| **Components** | **Description** |
| HP Laptop | Intel(R) Core(TM) i5-1035G1 CPU @ 1.00GHz 1.19 GHz |
| HP Laptop | Intel(R) Core(TM) i3-6006U CPU @ 2.00GHz 2.00 GHz  RAM : 8.00 GB |
| ASUS | Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz 1.99 GHz  RAM: 20.0 GB |

**-Software:**

Table 2.6.2 Technical Feasibility(software)

|  |  |
| --- | --- |
| **البرامجComponents** | **استخدام البرنامج Description** |
| PyCharm | To build the Apps |
| Sql Server | To create the data base |
| EdrawMax, VP | To draw the diagrams |
| Internet Connectivity | Available at the client phone |

**2.6.3 Legal feasibility :**

The proposed project ensures that it is legally accept and conform the legal and ethical requirements.

**2.6.4 Schedule feasibility :**

Table 2.6.4 Schedule feasibility

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Task** | **First month** | | | | **Second month** | | | | **Third month** | | | | | **Fourth month** | | | |
| **Week 1** | **Week 2** | **Week 3** | **Week 4** | **Week 5** | **Week 6** | **Week 7** | **Week 8** | **Week 9** | | **Week 10** | **Week 11** | **Week 12** | **Week 13** | **Week 14** | **Week 15** | **Week 16** |
| **Find my group** |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |  |
| **Find supervisor** |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |  |
| **Find idea** |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |  |
| **Project introduction** |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |  |
| **System**  **analysis** |  |  |  |  |  | | |  |  | |  |  |  |  |  |  |  |
| **System functional requirements** |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |
| **Documentary** |  |  |  |  |  | | | | | | | | | | | | |

**2.7 Project added values :**

Educate the users of the program about the dangerous and trap links, and increase people's knowledge of them and their dangers, so that knowledge of them spreads by providing them with simple advice for knowledge to be sufficient.

**2.8 Project management :**

- Supervision : Dr. Iman Droubi

The team:

1- Ahmad Saleh : leader / task : backend

2- Nagham Maraheen : member / task : interfaces

3- Muhannad Tomeh : member / task : interfaces

**2.9 Summary and recommendations :**

In this chapter, we talked about system analysis staring from a brief history of organization considering previous project and how they did work and how did their performance evolve over the time, project implementation options and proposed system that illustrate where and how the system is applied, we referred to functional and nonfunctional System requirements, discussed Feasibility study, Project added values and ending with mentioning project management.

**Chapter 3**

**{System Functional Requirements}**

**3.1 Introduction :**

Requirements analysis is critical to the success or failure of a systems or software project. The requirements should be documented, actionable, measurable, testable, traceable, related to identified business needs or opportunities, and defined to a level of detail sufficient for system design. In this chapter, we will use the "use case diagram" to show how the user interact s with our application, and we will show the functional and non-functional requirements of this application.

**3.2 Context or Use-Case Diagrams :**

**Context Diagram :**

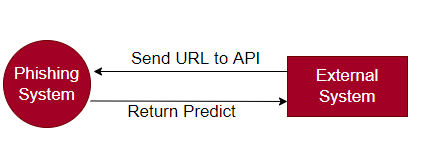
****

Figure 3.2.1 context diagram.

**Use Case Diagram:**

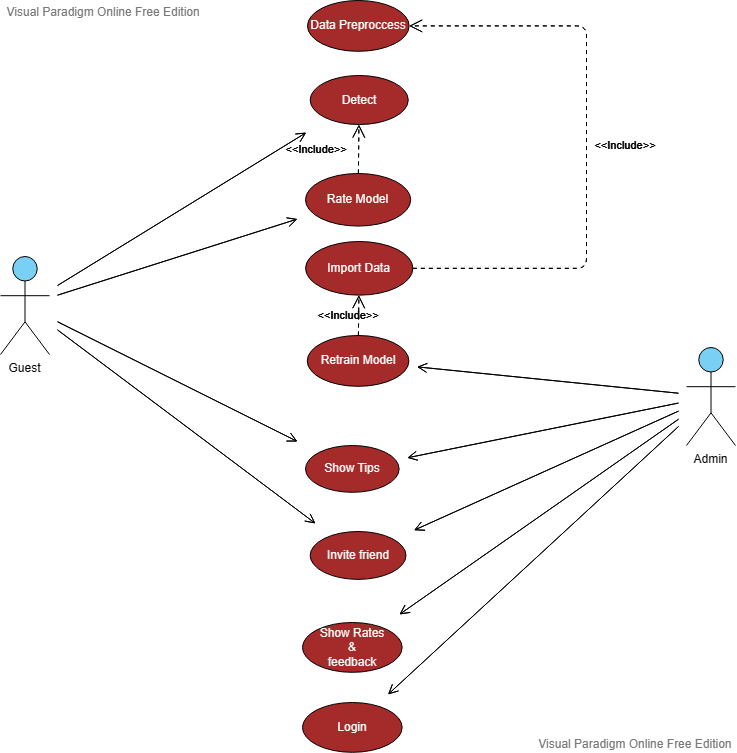


Figure 3.2.2 use case.

**3.3 - Functional Requirements:**

Table 3.3 Functional Requirements

|  |  |
| --- | --- |
| **Functional** | # |
| **Data Preprocessing**  1- Actor🡪Admin.  2- Input🡪Data OR URL.  3- Description🡪 Preprocess Data Like Convert Text To Vector Using Technique Word2Vec And Cleaning Data By Ignoring Missing Values | 1 |
| **Detect**  1- Actor🡪Guest.  2- Input🡪URL.  3- Description🡪 Check Weather URL Is Phishing OR Not  4- Reaction/Output🡪 Alter If Its Phishing OR Its Not Phishing | 2 |
| **Rate Model**  1- Actor🡪Guest.  2- Input🡪Rate From 5 And Feedback.  3- Description🡪Rate Model.  4- Reaction/Output🡪Rated Successfully OR Not | 2 |
| **Import Data**  1- Actor🡪Admin.  2- Input🡪Data  3- Description🡪Import Data From Admin Device To Train Model On It  4- Reaction/Output🡪Imported OR Not | 2 |
| **Retrain Model**  1- Actor🡪Admin.  2- Input🡪Data To Train Model On It  3- Description🡪Retrain Model For Getting Better Results  4- Reaction/Output🡪 Trained Successfully OR Not With New Information About Recall, Precision And F1 Score | 3 |
| **Show Tips**  1- Actor🡪Guest , Admin.  2- Input🡪Press Button Show Tips  3- Description🡪 Show Some Tips For Increase Awareness For User .  4- Reaction/Output🡪 Selection of appropriate and suitable data. | 4 |
| **Invite Friends**  1- Actor🡪Guest, Admin.  2- Input🡪  3- Description🡪Open Small Dialog To Share Website Using Facebook , Linkedin ,..  4- Reaction/Output🡪 | 5 |
| **Show Rates & Feedback**  1- Actor🡪Admin  2- Input🡪 Press Button Show Rates & Feedback  3- Description🡪Show People Ratings And His Feedback  4- Reaction/Output🡪 | 6 |
| **Login**  1- Actor🡪Admin  2- Input🡪 Email And Password  3- Description🡪Login Page For Admin  4- Reaction/Output🡪Login Successfully OR Failed | 7 |

**3.4 Non Functional Requirements:**

Table 3.4 Non Functional Requirements

|  |  |  |
| --- | --- | --- |
| **Description** | **Nonfunctional** | # |
| Clear, simple and easy that can be understood and used by anyone who does not have experience in the field | Understandable &Usability | 1 |
| Be available to all users at any time | Availability | 2 |
| Able to give high F1 Score result | Correctness | 3 |
| Ability to perform speedily and execute efficiently | Performance | 4 |
| Ability to cope with errors during execution and cope with erroneous input | Robustness | 5 |

**3.5 Summary and recommendations:**

In this chapter, we talked about Context , Use-Case Diagrams, Functional requirements description and Non-functional requirements description.

**Chapter 4**

**{System Design and Development}**

**4.1 Introduction:**

In this chapter, we will show several diagrams for our system, which are Class Diagram, Sequence Diagram, ER Diagram and Activity Diagram. And we will show the prototype of our application, which are the expected interfaces.

**4.2 Class diagram:**

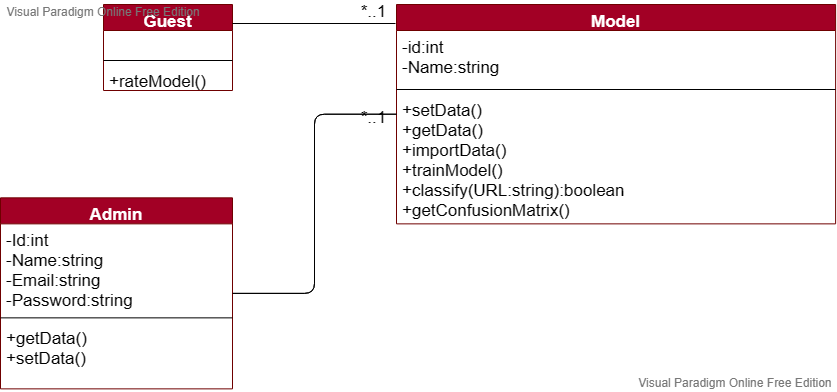
****

Figure 4.2 class diagram

**4.3 Sequence diagrams:**

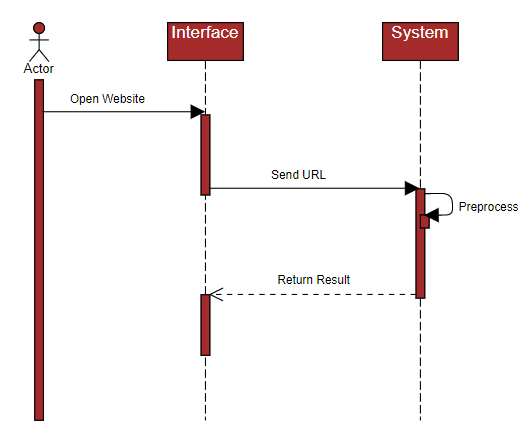


Figure 4.3 Sequence diagrams

**4.4 Entity Relationship Diagram (ER-Model)**

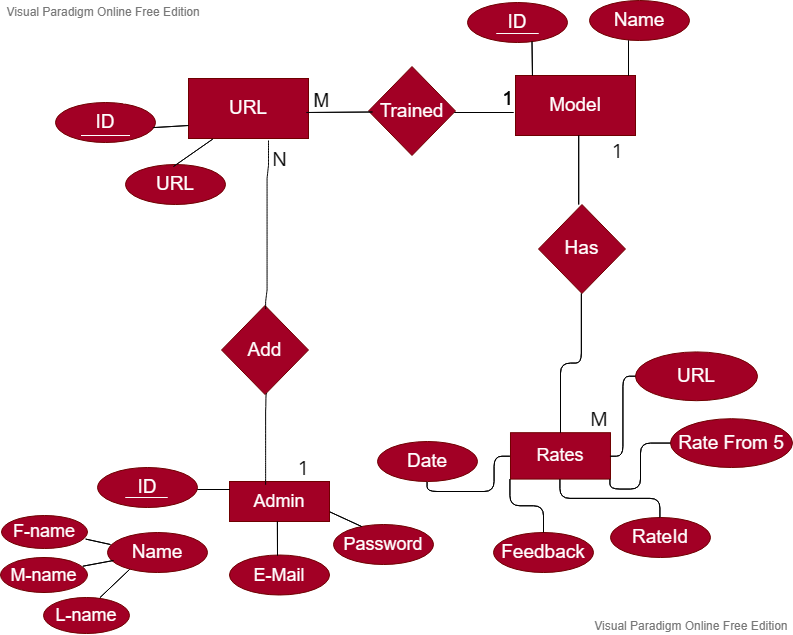
****

Figure 4.4 ER-Model.

**4.5 Activity diagrams:**

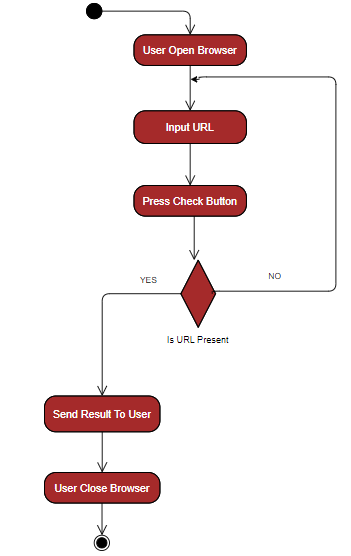
****

Figure 4.5 Activity diagram.

**4.6- System interface (input/ output design) :**

**-Main Page :**

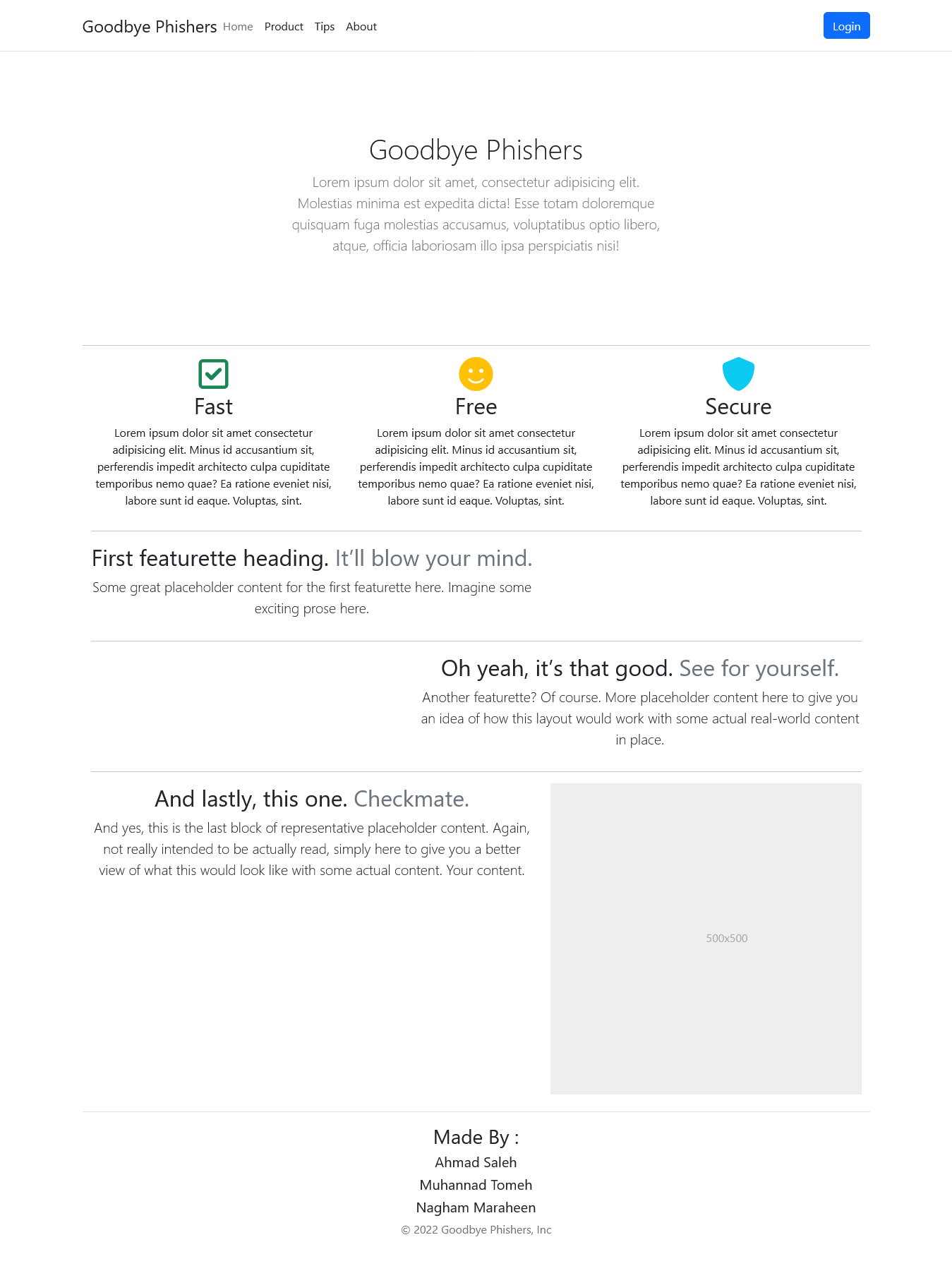
****

Figure 4.6.1 main page

**-Tips Page :**

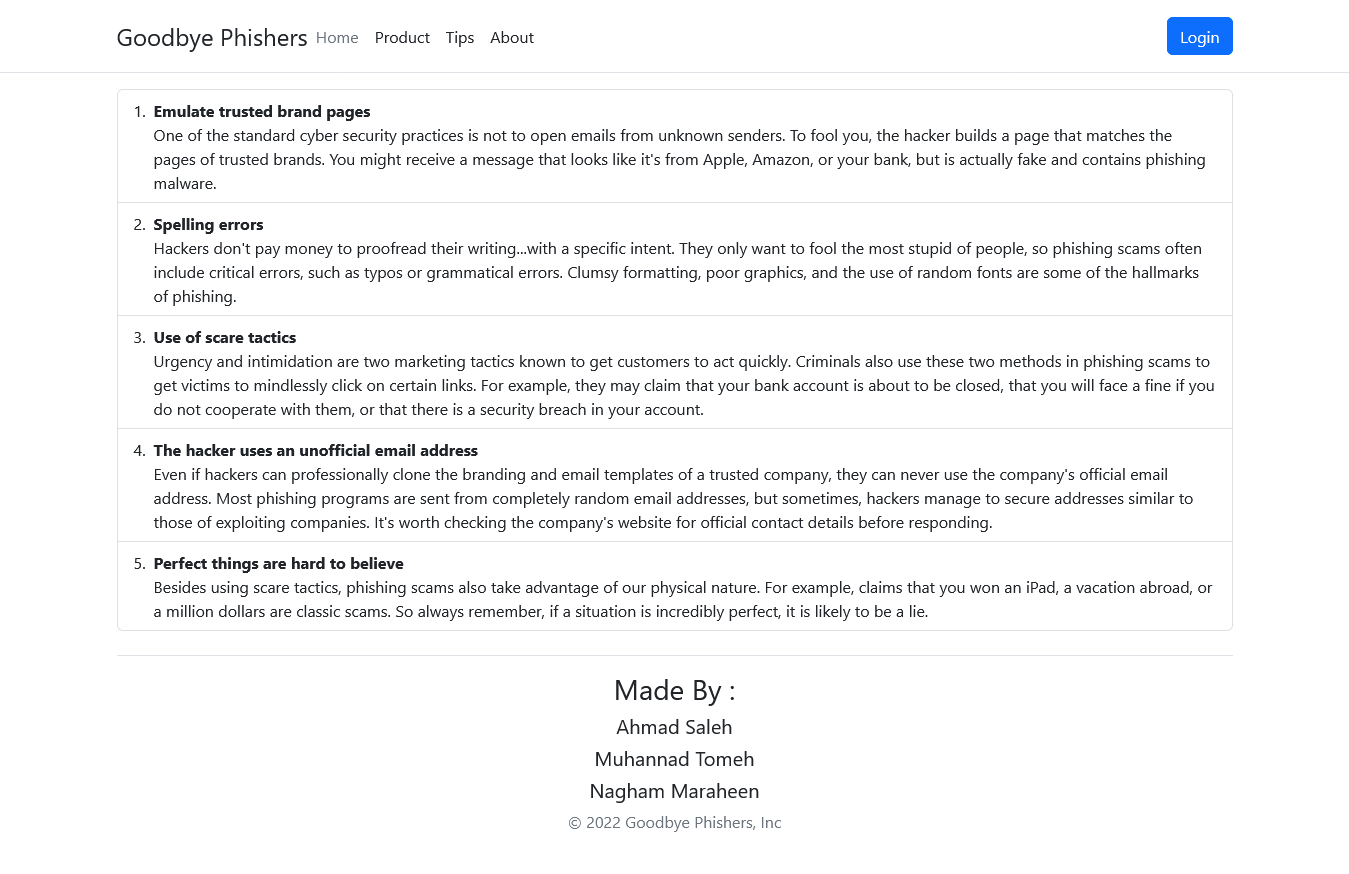


Figure 4.6.2 tips page.

**- Check Page :**

when User Not Test Product At Least One Time  
He Cant Make Rate

So Button Rate Model Is Disabled

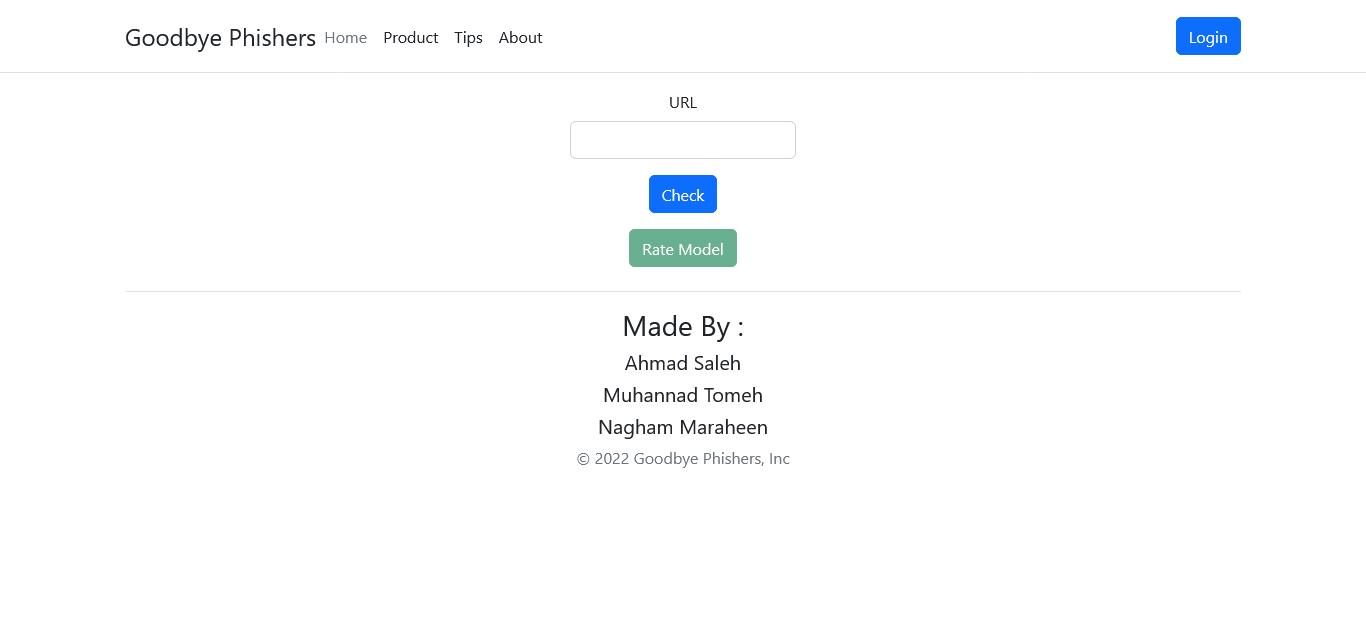


Figure 4.6.3 check page

**-Check Page(URL Is Not Phishing) :**He Can Make Rate Because He Use Model

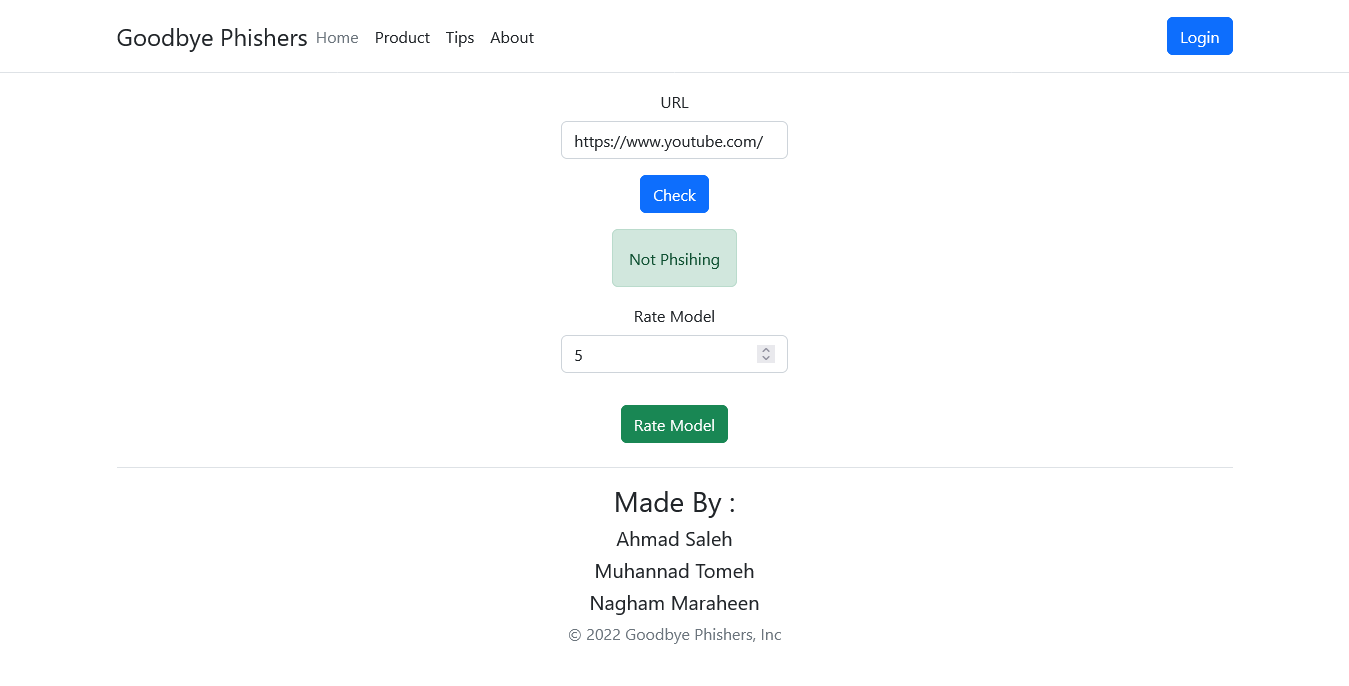
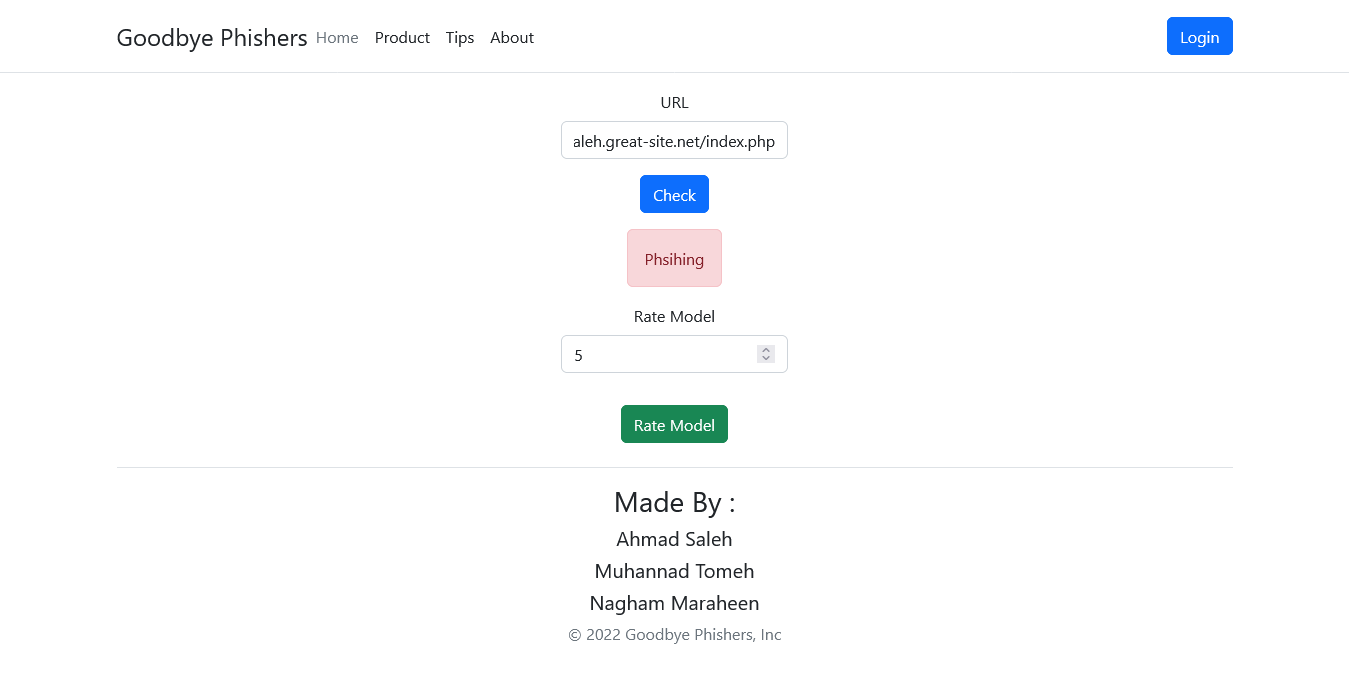
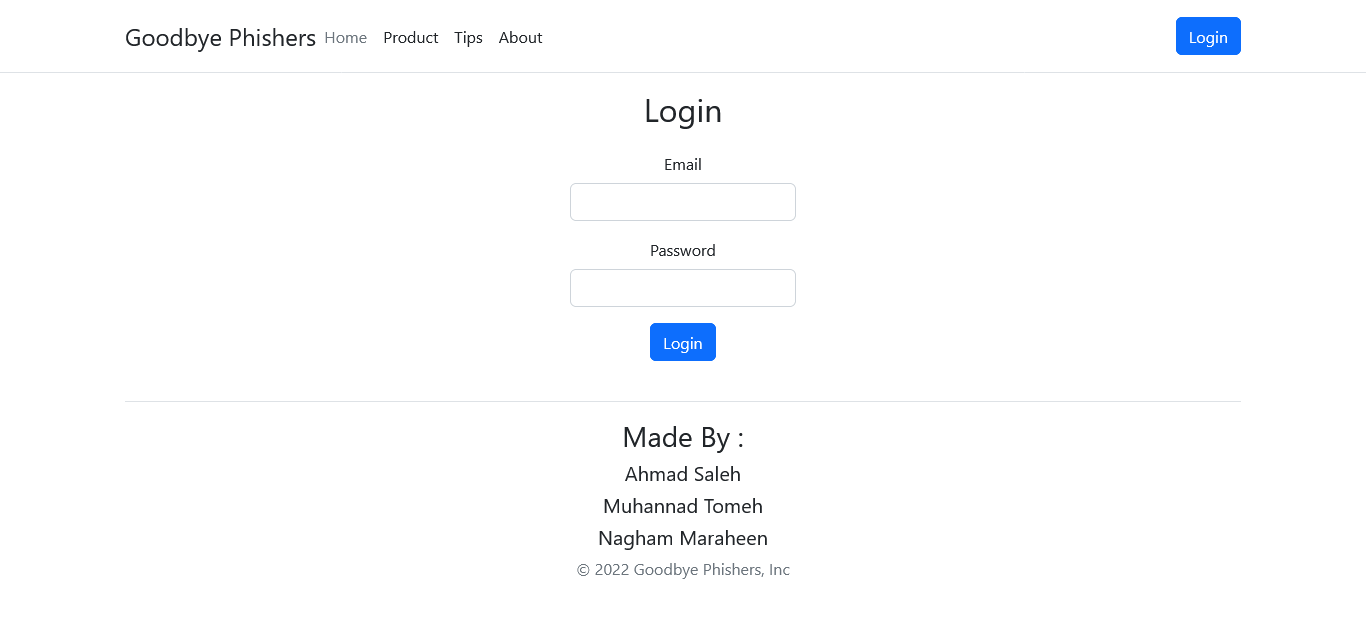


Figure 4.6.3.1 check page (not phishing)

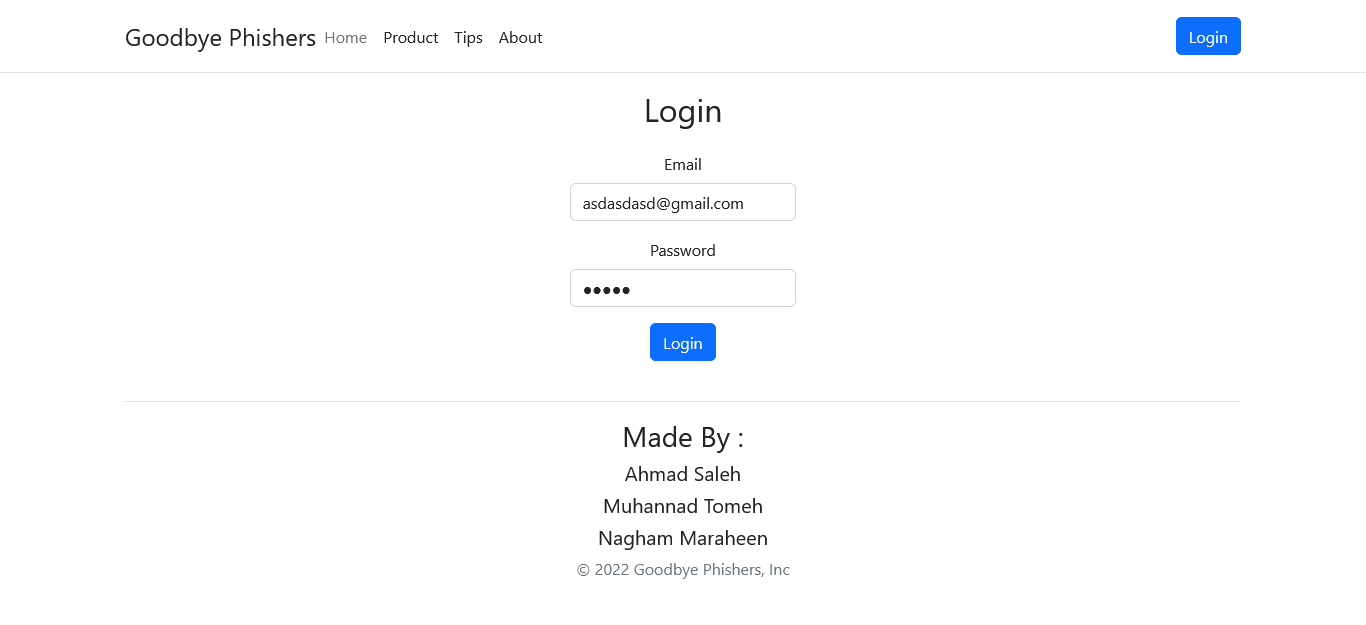
**- Check Page(URL Is Phishing) :**He Can Make Rate Because He Use Model



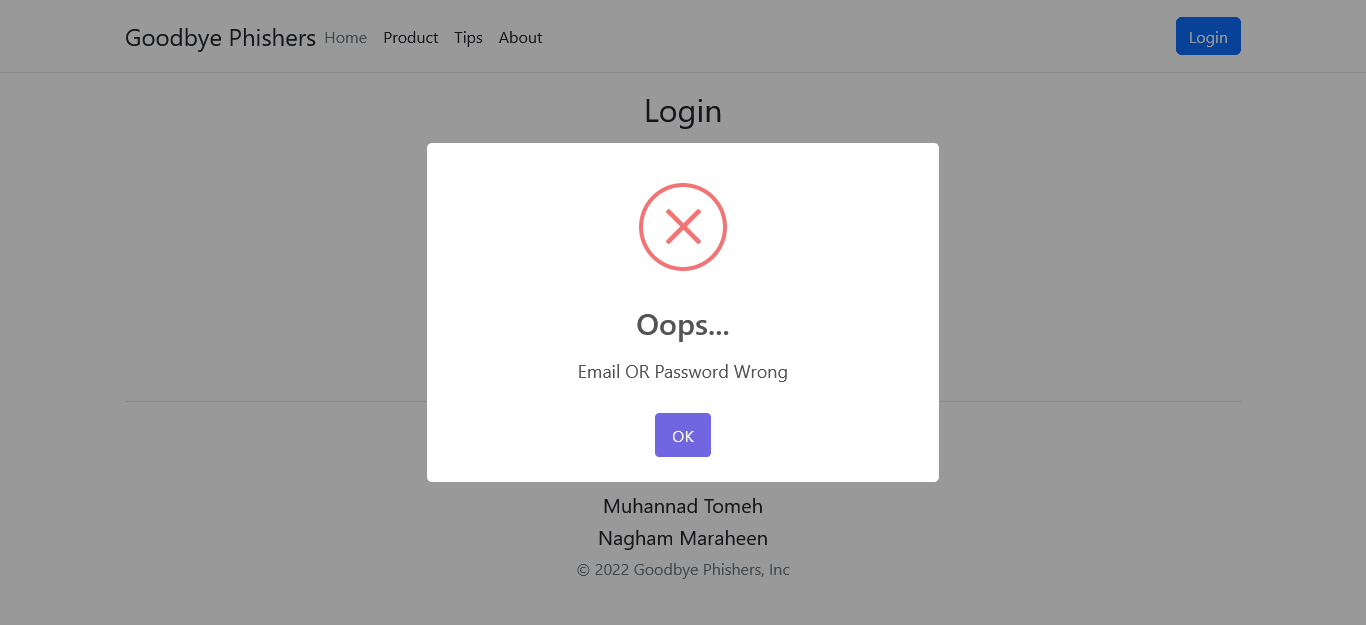
**-Login Page :**



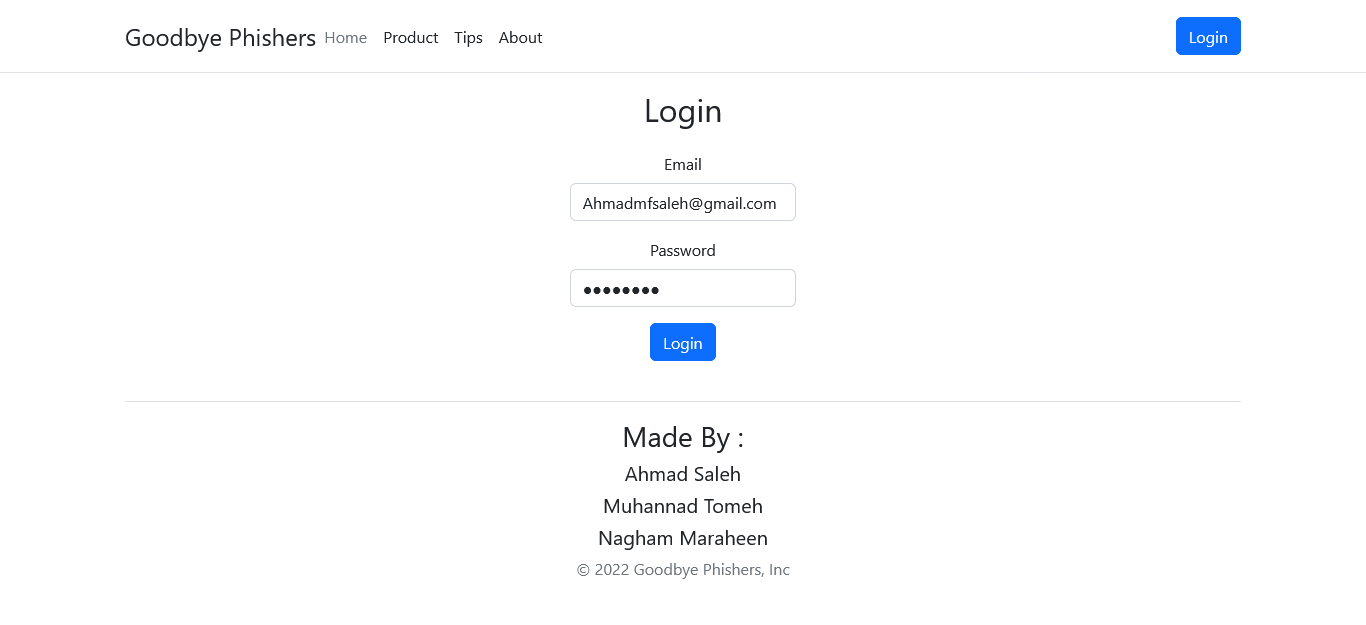
**When Enter Wrong Informations :**

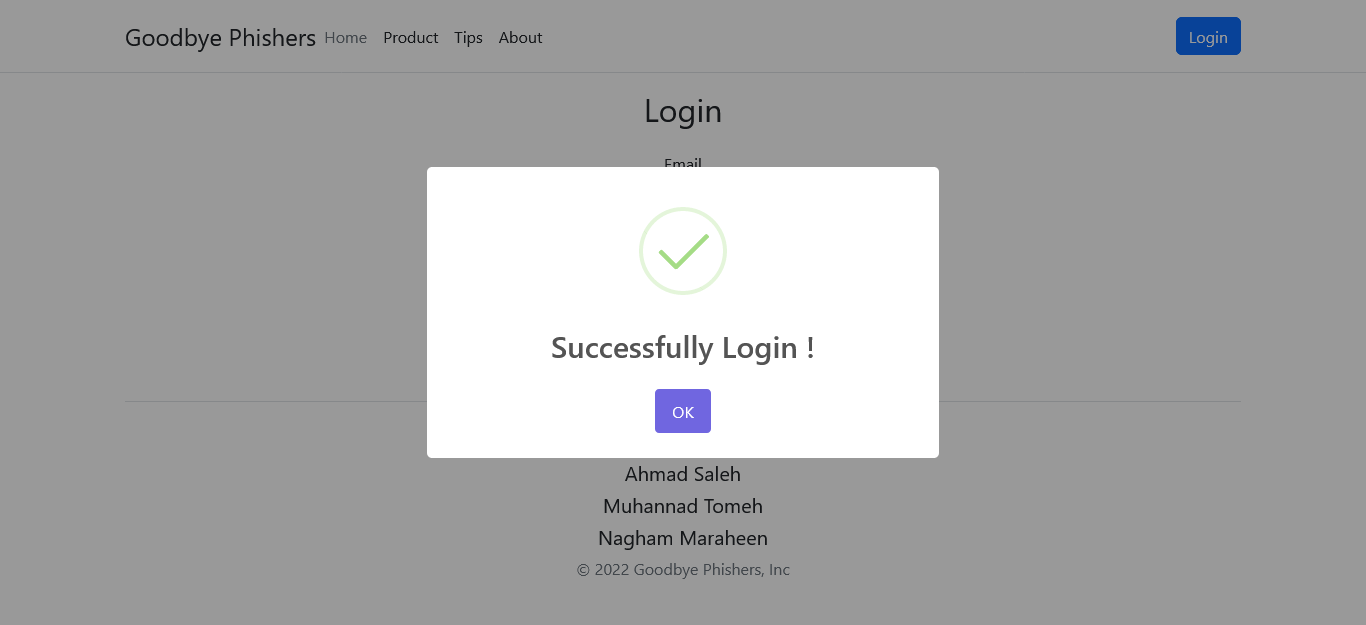


**Reaction :**

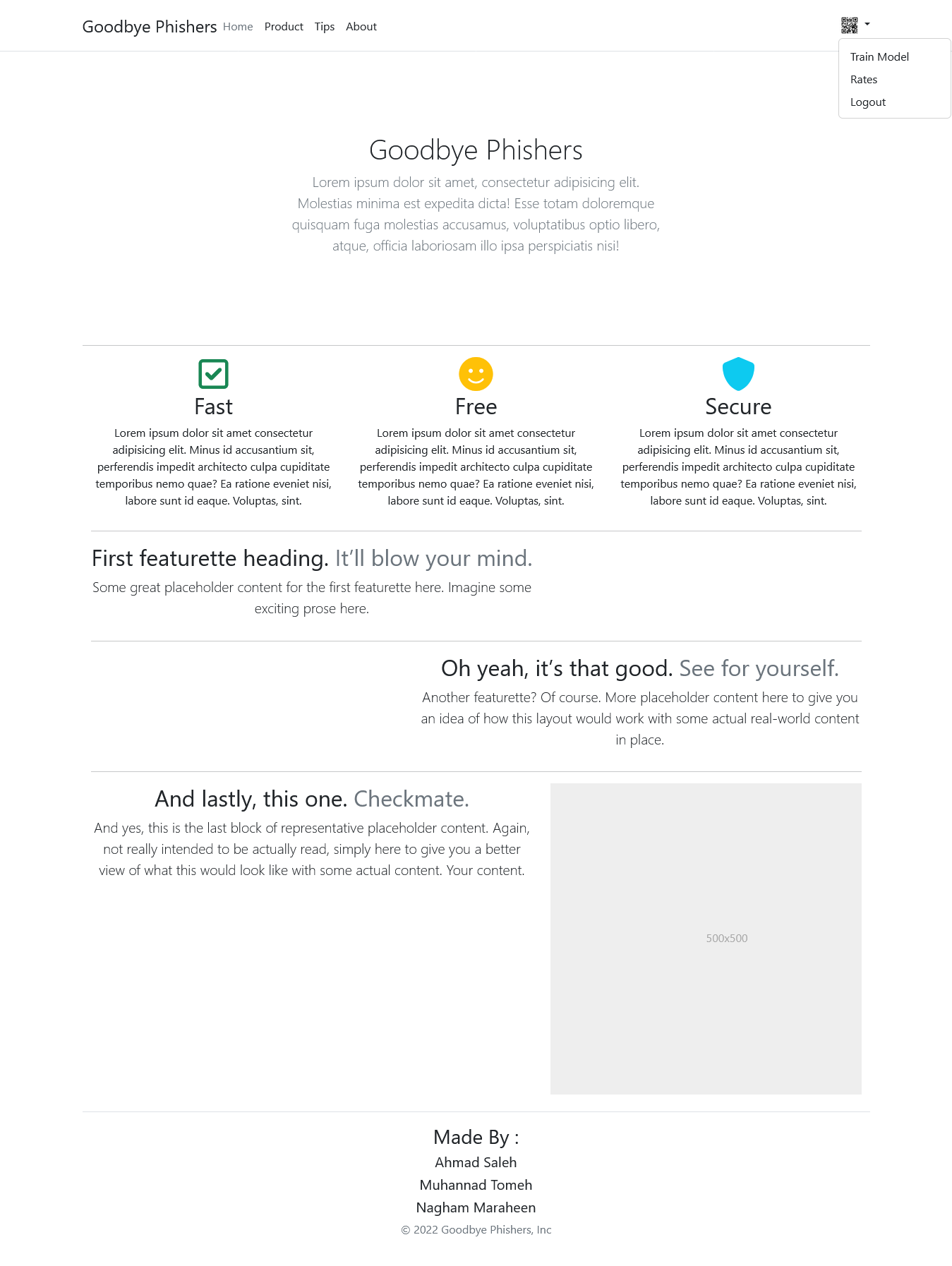


**When Enter Correct Information :**

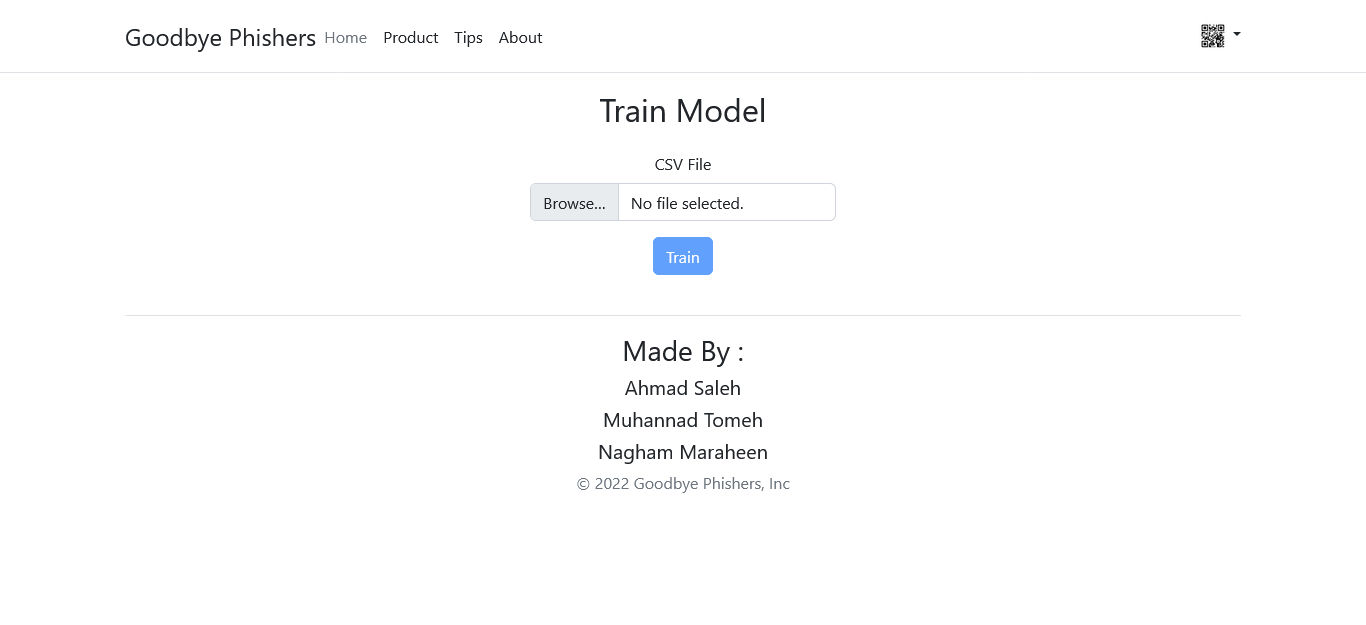


**Reaction :**   


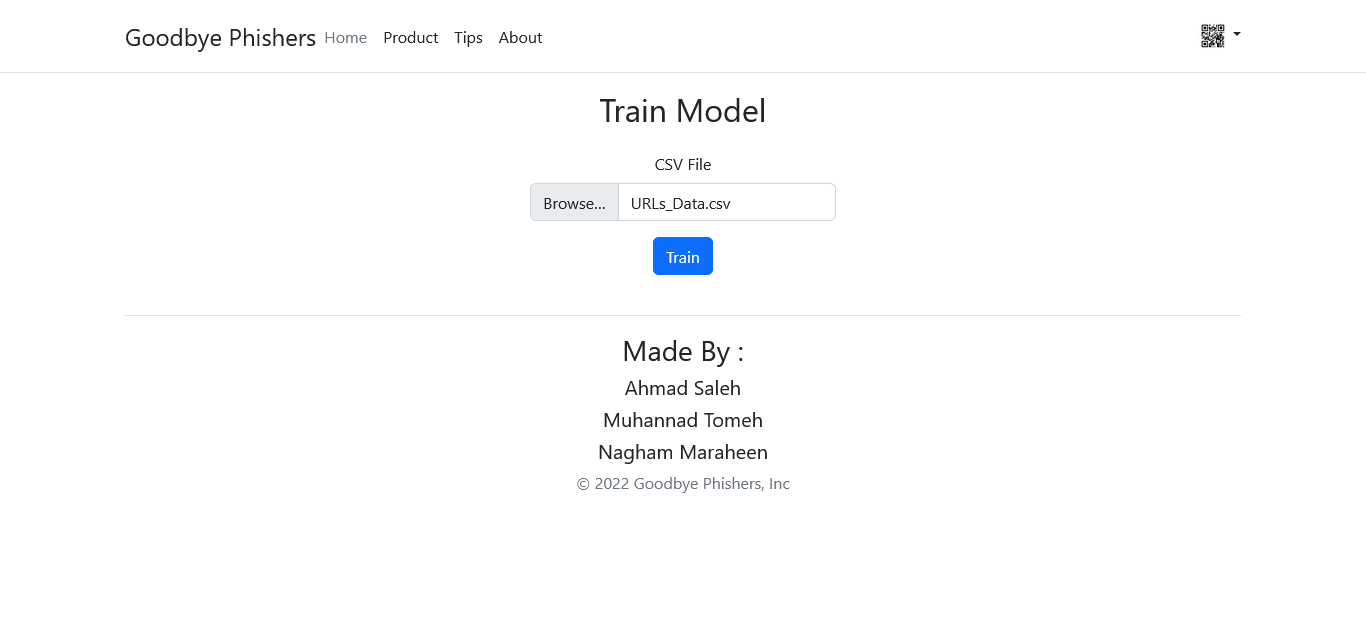
**When Admin Successfully Login :**



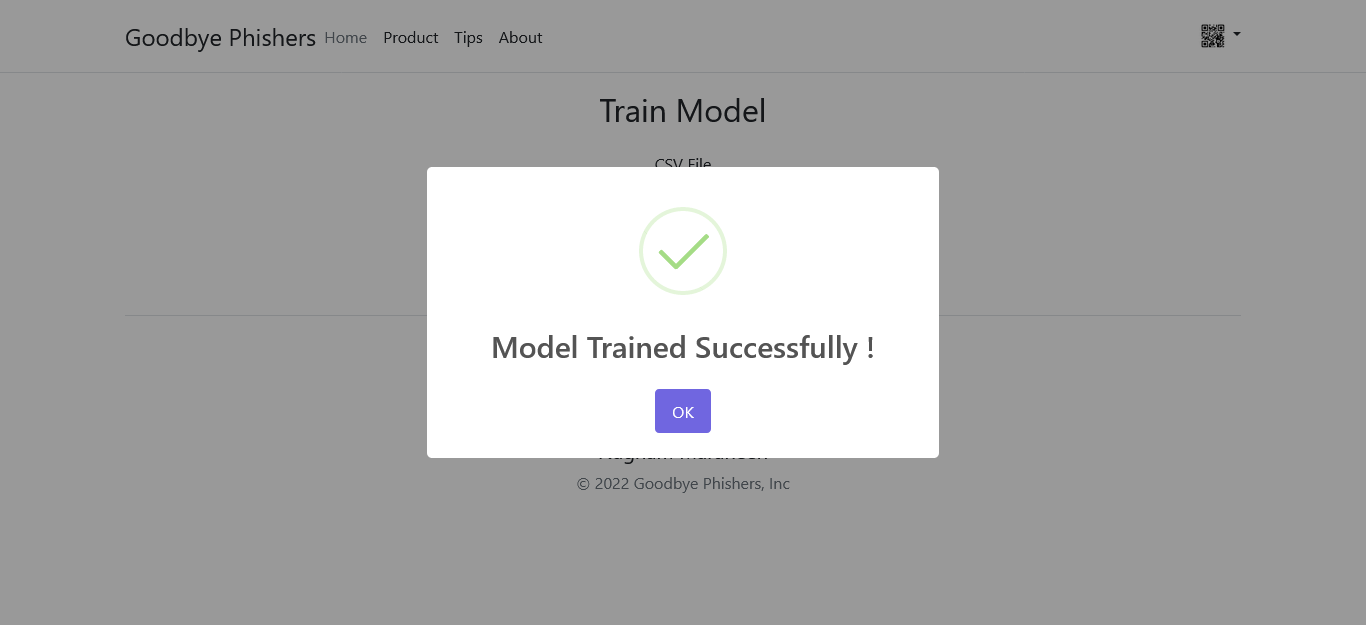
**When Click Train Mode :**



**When Upload Data :**



**Reaction After Train It :**



**When Click Rates :**

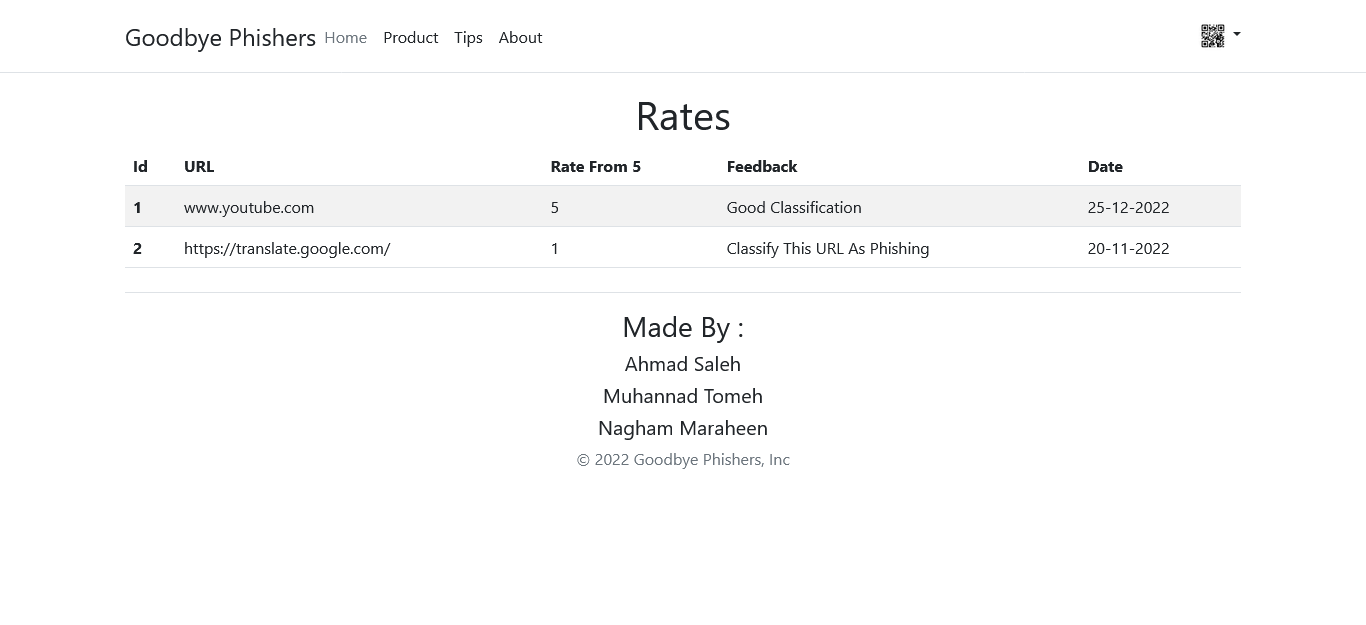


Figure 4.6.3.2 check page(phishing)

**4.7 Summary and recommendations :**

In this chapter, we talked about Class diagram , Sequence diagrams, Entity Relationship Diagram (ER-model) , Activity diagrams and System interface (input/ output design).

**References :**

[ML | Linear Regression - GeeksforGeeks](https://www.geeksforgeeks.org/ml-linear-regression/)

[Introduction to Logistic Regression | by Ayush Pant | Towards Data Science](https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148)

[K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint](https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning)

[Multinomial Naїve Bayes’ For Documents Classification and Natural Language Processing (NLP) | by Arthur V. Ratz | Towards Data Science](https://towardsdatascience.com/multinomial-na%C3%AFve-bayes-for-documents-classification-and-natural-language-processing-nlp-e08cc848ce6)

[Supervised Machine Learning: Regression and Classification - Home | Coursera](https://www.coursera.org/learn/machine-learning/home/week/1)

[Advanced Learning Algorithms - Home | Coursera](https://www.coursera.org/learn/advanced-learning-algorithms/home/week/1)

[Understanding TF-ID: A Simple Introduction (monkeylearn.com)](https://monkeylearn.com/blog/what-is-tf-idf/)

[Understanding TF-IDF for Machine Learning | Capital One](https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/)

[Introduction to Word Embedding and Word2Vec | by Dhruvil Karani | Towards Data Science](https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa)

[Natural Language Processing with Classification and Vector Spaces - Home | Coursera](https://www.coursera.org/learn/classification-vector-spaces-in-nlp/home/welcome)