Information Security Lab
Project Report

Submitted in Partial Fulfilment of requirements for the Award of

Degree of Bachelor of Technology in Computer Science and Engineering

Submitted by

Meet Patel (21BCP360)

Karan Shah (21BCP377)

Kushagra Darji (21BCP387)

Div-6 (G11)

Submitted To

Dr. Rutvij Jhaveri

Department of Computer Science and Engineering

School of Technology

Pandit Deendayal Energy University, Gandhinagar

September 2023

TABLE OF CONTENTS

Phishing Link Detector

○ Introduction:

- A Phishing Link Detector is a crucial cybersecurity tool designed to identify and thwart phishing attacks, a type of cyber threat where malicious actors attempt to deceive individuals into divulging sensitive information or performing harmful actions by impersonating legitimate entities. Phishing attacks are typically delivered through deceptive links embedded in emails, messages, or websites, making it vital to have an effective detector in place.
- This technology employs various techniques to recognize phishing links. One common method is URL analysis, which scrutinizes the URL's components and checks them against known malicious patterns or blacklists. Another approach involves content analysis, where the detector examines the content of a web page or email for suspicious elements such as forged logos or requests for personal data.
- Machine learning and artificial intelligence play a significant role in enhancing the accuracy of phishing link detection. These systems can learn from vast datasets of known phishing attempts and adapt to new and evolving attack techniques. They can also analyze user behavior to identify anomalies that might indicate a phishing attempt.
- A Phishing Link Detector is a valuable addition to any organization's cybersecurity infrastructure, helping protect sensitive data, financial assets, and privacy. It acts as a first line of defense against the ever-evolving tactics of cybercriminals, reducing the risks associated with falling victim to phishing attacks and bolstering overall cybersecurity posture.

○ Problem Statement:

- The problem of Phishing Link Detection revolves around the need to develop effective mechanisms and algorithms for identifying fraudulent and deceptive URLs that are employed by cybercriminals to trick users into disclosing sensitive information or engaging in harmful actions. Phishing attacks are a pervasive and evolving threat, and the challenge lies in reliably distinguishing between legitimate and malicious links in a timely manner. This problem statement necessitates the development of advanced technology and methodologies, including machine learning, data analysis, and artificial intelligence, to mitigate the risks associated with phishing attacks, safeguard sensitive data, and enhance overall cybersecurity defenses. The objective is to create robust and accurate detection systems that can adapt to emerging phishing tactics, thereby protecting individuals and organizations from falling victim to these deceptive schemes.

○ Project Details:

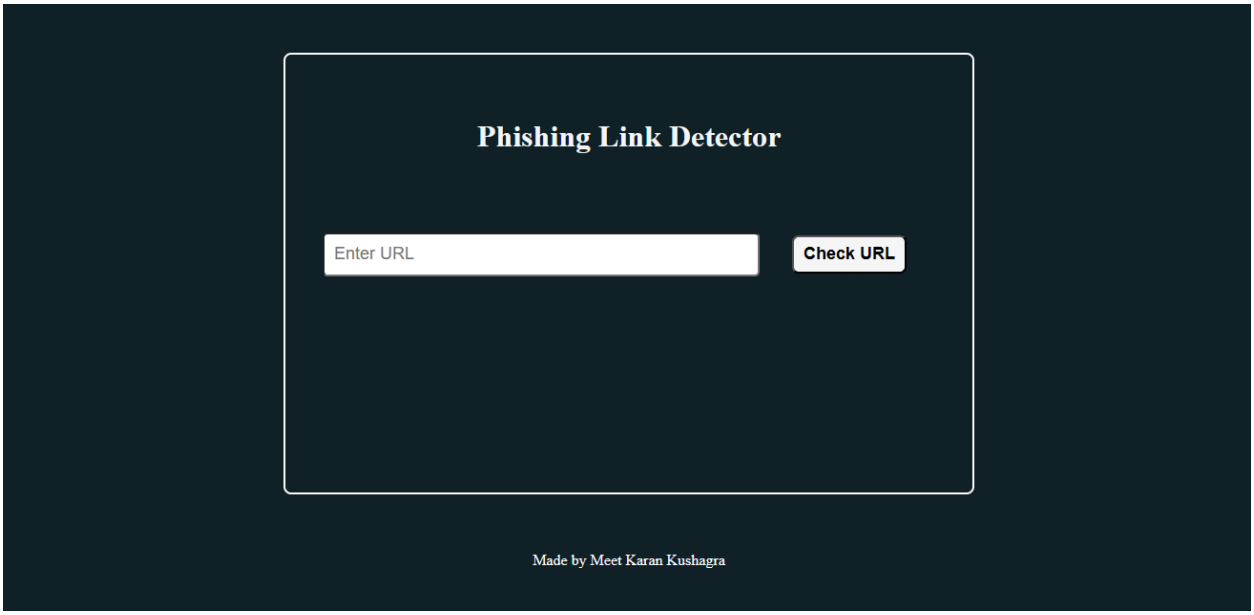

Fig:01 Phishing Link Detector website's homepage

The Phishing Link Detector project aims to develop a robust cybersecurity solution to combat phishing attacks by effectively identifying fraudulent URLs and deceptive content. The project encompasses several key components:-

Problem Statement and Criteria:- The project addresses the pervasive threat of phishing attacks, focusing on the need to assess URLs, content, and user behavior to detect and prevent such attacks.

Machine Learning Integration:- Machine learning and AI techniques will be integrated to enable the system to adapt to evolving phishing tactics and enhance its detection accuracy over time.

Real-time Monitoring:- The system will actively monitor URLs in real-time, ensuring swift detection and protection against emerging phishing threats.

URL and Content Analysis:- The project will develop algorithms to analyze URL components for known malicious patterns and examine web page and email content for fraudulent elements.

User Behavior Analysis:- Algorithms will be created to detect anomalies in user behavior that may indicate phishing attempts, enhancing the system's ability to detect sophisticated attacks.

Machine Learning Models:- Machine learning models will be used to classify and identify phishing links based on historical data and emerging trends.

User Recommendations:- For identified phishing links, the system will provide specific recommendations to users, such as avoiding the link and reporting it.

Testing and Validation:- Rigorous testing and validation will be performed to ensure the system's accuracy, with a focus on minimizing false positives and negatives.

User Interface:- An intuitive and user-friendly interface will be designed to make the tool accessible to a wide range of users.

Documentation and Reporting:- Comprehensive documentation will be provided for the tool's setup, operation, and maintenance. Additionally, a reporting system will be developed to track and analyze phishing attempts and outcomes.

Continuous Improvement:- The project includes plans for ongoing development to stay ahead of emerging phishing techniques and enhance the system's capabilities, providing a proactive defense against phishing attacks.

- **Working Of Our Phishing Link Detector:**

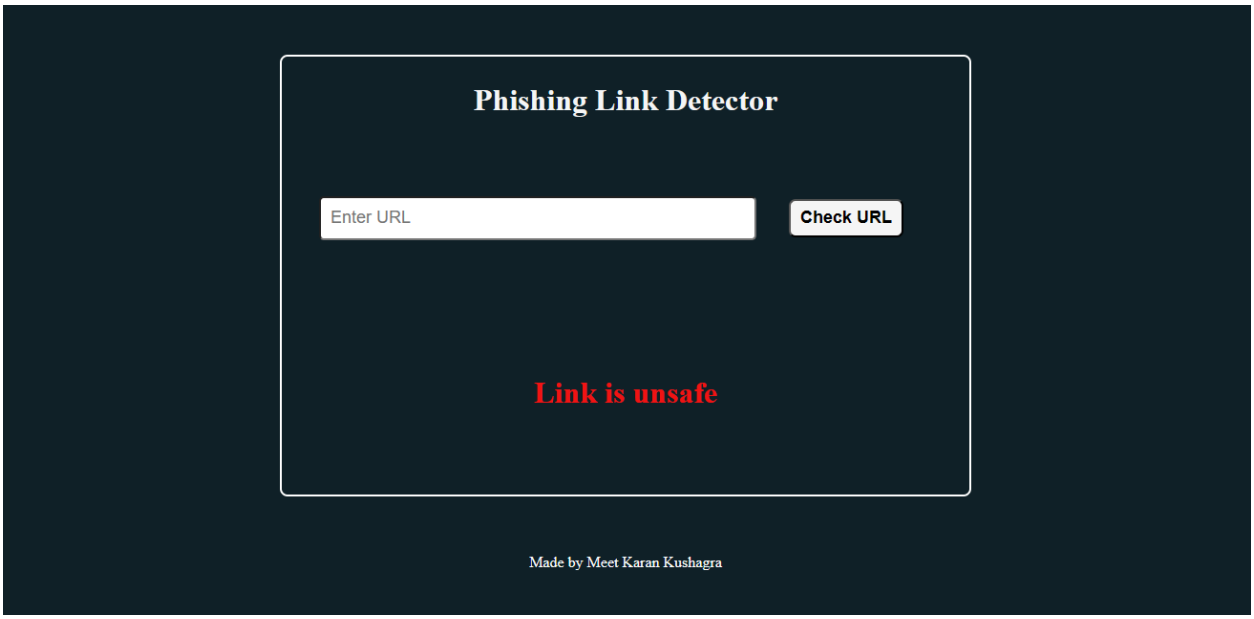

Fig:02 When the given link is Unsafe to use

- Link entered by the user is not safe to process for use.

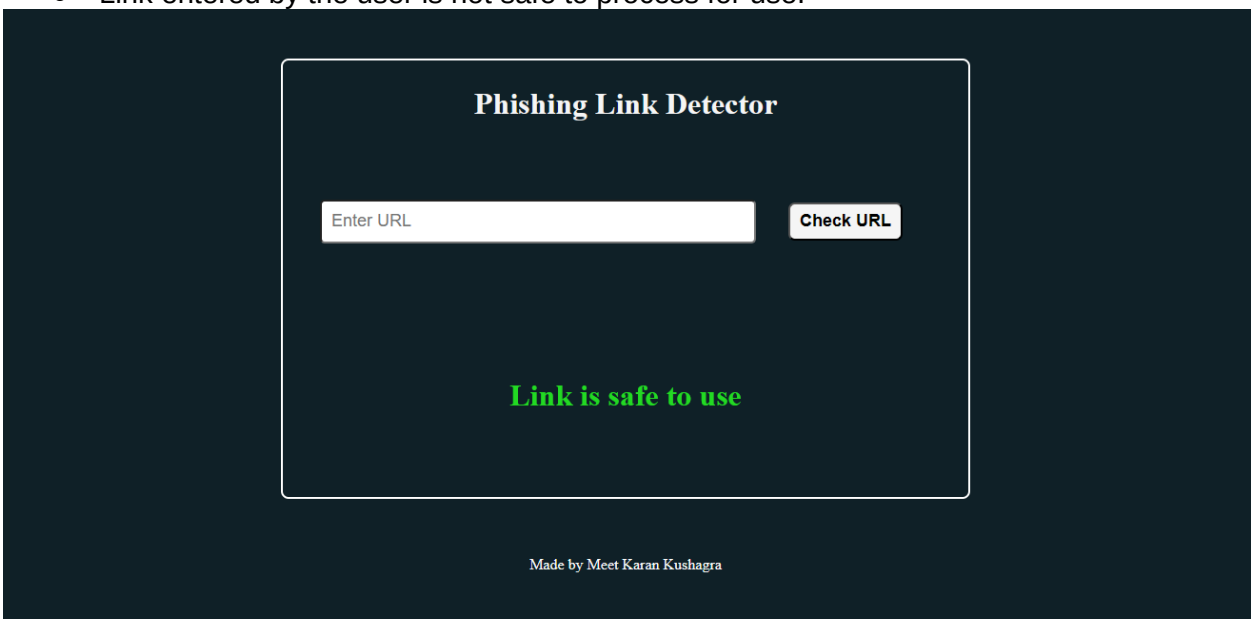

Fig:03 Link entered by user is Safe to use.

- Link entered by user is safe to use.

○ Technologies Used:

1. **Python:** The code is written in Python, a versatile and widely used programming language known for its simplicity and readability.
2. **Pandas:** Pandas is a popular data manipulation library in Python. You use it to read and manipulate the dataset, perform data cleaning, and conduct exploratory data analysis (EDA).
3. **Scikit-Learn (sklearn):** Scikit-Learn is a comprehensive library for machine learning and data analysis. In your code, you use Scikit-Learn for various tasks, including feature extraction and model training.
4. **XGBoost:** XGBoost is an optimized and efficient gradient boosting library. You use it to build a machine learning model for password strength prediction.
5. **Dill:** Dill is a Python library for serializing and deserializing Python objects. You use it to save the trained XGBoost model and the TfidfVectorizer object to disk for later use.
6. **Extracting Features:-** Effective feature extraction is crucial for building accurate and efficient models for phishing link detection. By selecting the right attributes from URLs, content, and user behavior, the system can identify and prevent phishing attacks with high precision.
7. **Flask :** For backend side Python's Flask library has been used.
8. **HTML,CSS:** For frontend HTML and CSS have been used.

○ Model Description:

The most important factor in analyzing the features of the URL using machine learning is collecting a meaningful dataset that contains a huge amount of URLs with classification of phishing and not phishing. The dataset contains phishing and non phishing URLs with 1 and 0 assigned to it respectively.

Here's the code's workflow according to the project:

- Data Loading and Preprocessing: The code starts by loading a dataset from a CSV file ('phishing_data.csv') that contains URL data. It then performs data cleaning, handling any bad lines and missing values in the dataset.
- Feature Extraction: The feature extraction process involves extracting seven parameters from the URL. For every parameter, the URL is classified as phishing or not using values 1 and 0 for phishing and not phishing respectively. These values for all the URLs are stored in another CSV file (‘url_specifications.csv’).
- Training-Testing Split: The dataset is split into training and testing sets using the train_test_split function from Scikit-Learn. This allows for model training and evaluation on different subsets of the data.
- Model Selection and Training: The Logistic Regression algorithm is selected as the model for URL classification. Logistic Regression is an efficient and optimized binary classification algorithm commonly used for classification tasks. An instance of the Logistic Regression classifier (model) is created, and it is trained on the training data (X_train and y_train) using the .fit() method.
- Model Evaluation: The trained Logistic Regresssion model is used to make predictions on the testing data (X_test) using the .predict() method. Performance metrics for classification are computed and printed using Scikit-Learn's functions. These metrics provide insights into how well the model is performing in predicting URL characteristics.
- Predicting URL Strength: A sample URL is provided, and the url_features is used to extract the characteristics of the URL. The Logistic Regression model is then used to predict the URL strength of the sample URL.
- Model Interpretation: The code calculates and displays the predicted class label for the sample URL (model.predict(url_features)) and the predicted class probabilities (model.predict_proba(url_features)). This provides information about the model's confidence in its predictions.

- In summary, the Logistic Regression model is trained to predict the strength of URLs based on their composition and certain features. This machine learning model is designed to assess and classify the strength of URLs, which is essential for security of users in various situations.

  To summarize:

- Data is loaded from a CSV file ('phishing_data.csv') containing URL data which has 80,000 URLs out of which 50,000 URLs are legitimate and 30,000 URLs are phishing.
- Data preprocessing includes loading CSV data and extracting useful information from it.
- Logistic Regression is used to train a machine learning model for URL strength prediction.

Overall, this code represents a typical workflow for building and training a machine learning model to predict URL strength, using Python libraries and machine learning techniques. It aligns with the project's goal of assessing and improving user security in a controlled and responsible manner.

○ Future Scope:

The Phishing Link Detector project also holds substantial promise and future scope in the realm of cybersecurity and information security. As the threat landscape evolves, there are several potential areas of growth and development for this project:-

- **Advanced Machine Learning Techniques:-** Continuously improve the machine learning models used in phishing link detection to adapt to ever-changing attack strategies and enhance detection accuracy.
- **Behavioral Analysis:-** Expand the project's capabilities by incorporating more sophisticated behavioral analysis to identify subtle anomalies in user behavior, which can be indicative of phishing attempts.
- **Big Data Integration:-** Embrace big data analytics to process and analyze vast amounts of data, enabling the detection of broader patterns and trends in phishing attacks.
- **Real-Time Threat Intelligence:-** Integrate with real-time threat intelligence sources to stay updated on the latest phishing threats and rapidly adapt the detection algorithms.
- **Cross-Platform Compatibility:-** Extend the project to work seamlessly across various platforms, including web browsers, email clients, and mobile devices, providing comprehensive protection for users.
- **Collaboration with Industry:-** Collaborate with industry partners, organizations, and cybersecurity experts to share knowledge and insights, contributing to the development of best practices in phishing link detection.

As phishing attacks continue to evolve and grow in sophistication, the Phishing Link Detector project can adapt and expand to provide enhanced protection and security for users and organizations in the face of this dynamic and persistent threat.

○ Conclusion:

- In this project, we have ventured into the domain of phishing link detection, focusing on the development of an advanced cybersecurity tool. Our primary goal has been to identify and thwart phishing attacks by effectively distinguishing fraudulent URLs and deceptive content. We have tackled this issue by implementing a comprehensive system that employs a range of techniques and features.
- The essence of our project lies in the importance of safeguarding individuals and organizations against the ever-evolving threat of phishing attacks. Phishing, being a persistent and adaptable menace, requires proactive measures to counteract it. Our project acknowledges this need and seeks to contribute significantly to the realm of information security.
- We have put extensive efforts into feature extraction, harnessing the power of machine learning, and conducting real-time monitoring. By scrutinizing URL components, content, and user behavior, we have strived to create a sophisticated system that can adapt and evolve with the dynamic phishing landscape.
- As we move forward, the future scope of our Phishing Link Detector project is promising. The project's potential areas of growth encompass advanced machine learning, behavioral analysis, big data integration, real-time threat intelligence, cross-platform compatibility, user education, and collaboration with industry experts. By continuously refining our approach and aligning with regulatory standards, we aim to provide comprehensive and robust protection against phishing attacks.
- In conclusion, our project in the domain of phishing link detection is a vital step towards enhancing cybersecurity. We are committed to the ongoing development of our system and the provision of cutting-edge security measures to safeguard users and organizations against the pervasive threat of phishing attacks.

- ## References:

Phishing Websites Dataset - Mendeley Data