International Scientific Journal of Engineering and Management
Volume: 04 Issue: 01 | Jan– 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02220

URL BASED PHISHING DETECTION

Ms.MADHU T , Ms. MONICA M Ms. SHYMA S (MENTOR)

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

SRI SHAKTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY , COIMBATORE.

----------***----------

Abstract - Phishing attacks, which deceive users into revealing sensitive information by mimicking legitimate websites, pose a growing threat in the digital age. To address this challenge, we propose a machine learning-based system for detecting phishing URLs. The system uses logistic regression in conjunction with TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to analyze and classify URLs as either legitimate or phishing. By identifying suspicious patterns in URL structures, such as unusual domain names, special characters, or deceptive keywords, the model effectively predicts whether a given URL is malicious. The detection system is deployed through a web interface built with Flask, allowing users to input URLs for real-time analysis. If a URL is flagged as phishing, the system provides an alert along with specific insights into the factors that led to the decision. Additionally, the system checks the URL against a database of known phishing websites, ensuring efficient recognition of already verified threats. This project provides an easy-to-use, scalable solution for combating phishing attacks by combining machine learning techniques with web integration. It aims to enhance online security for both individual users and organizations, offering a reliable tool to prevent phishing and protect sensitive information.

1. INTRODUCTION

Phishing attacks have become increasingly sophisticated, making it harder for users to distinguish between genuine and malicious websites. These attacks often start with deceptive URLs designed to trick users into revealing sensitive information like passwords and financial details. Traditional detection methods, such as blacklists, are limited in their ability to detect new and evolving phishing schemes.

To combat this, our project introduces a machine learning-based phishing detection system. It utilizes **logistic regression** and **TF-IDF vectorization** to analyze URL patterns and identify potential phishing threats by examining suspicious features like domain irregularities and keyword obfuscation. The system is implemented through a user-friendly web interface built with **Flask**, where users can input URLs for real-time safety analysis. Additionally, it cross-references known phishing databases to flag previously reported threats. By combining real-time detection with machine learning, this project offers a scalable and effective solution to help protect users from phishing attacks.

2. LITERATURE REVIEW

A. Traditional Methods

Early phishing detection methods relied on blacklists, heuristic-based rules, and machine learning models like **Support Vector Machines (SVM)**. These approaches required manual feature engineering and were limited in detecting new phishing URLs, struggling with scalability and adaptability to evolving phishing tactics.

B. Feature Extraction

Feature extraction plays a key role in phishing detection. Traditional approaches used lexical features (e.g., URL length, special characters) and host-based features (e.g., domain registration), but they lacked the ability to capture deeper patterns. With the introduction of **TF-IDF vectorization**, more complex patterns in URLs can be automatically extracted, improving detection accuracy.

C. Machine Learning Techniques

Machine learning models such as **Logistic Regression** and **Random Forests** have significantly improved phishing detection by identifying patterns in URL data. Deep learning models, like **CNNs** and **RNNs**, automate feature learning and enhance detection capabilities, though they require more data and computational power.

3. METHODOLOGY

DATA COLLECTION AND PREPROCESSING

The dataset used for this project consists of labeled URLs, sourced from public repositories and open datasets, containing both phishing and legitimate URLs. The data was cleaned to ensure consistency and quality, removing duplicates and invalid URLs. Preprocessing involved tokenizing URLs and applying **TF-IDF (Term Frequency-Inverse Document**

ISSN: 2583-6129
DOI: 10.55041/ISJEM02220

Frequency) vectorization to convert the textual data into numerical form. The dataset was split into training, validation, and testing sets using a 70:20:10 ratio for balanced model evaluation.

MODEL DEVELOPMENT

For model development, **Logistic Regression** was chosen due to its efficiency and interpretability in binary classification tasks. The model was trained using the TF-IDF features extracted from the URLs, enabling it to learn patterns indicative of phishing attacks. The output layer used a sigmoid activation function to predict the probability of a URL being phishing or legitimate. Additionally, the system was designed to check URLs against a database of known phishing websites for enhanced accuracy.

TRAINING AND VALIDATION

The model was trained using the **Adam optimizer** and binary cross-entropy as the loss function. Early stopping was applied to prevent overfitting, while learning rate scheduling was used to ensure faster convergence. The training process was monitored through validation accuracy and loss, ensuring the model generalized well to new URLs. Data augmentation techniques, such as adding slight variations to URLs, were employed to increase the robustness of the model.

TESTING AND EVALUATION

The trained model was evaluated on a holdout test set to assess its performance. Key metrics such as **accuracy, precision, recall,** and **F1-score** were computed to evaluate the quality of URL classification. Misclassified URLs were analyzed to identify patterns and potential improvements in the detection system.

DEPLOYMENT

The model was deployed as part of a real-time URL phishing detection system. A web interface was developed using **Flask**, allowing users to input URLs and receive immediate feedback on whether the URL is safe or phishing. The backend integrated the trained model, and the frontend provided a user-friendly interface, with the system also checking against a known phishing URL database to ensure comprehensive protection.

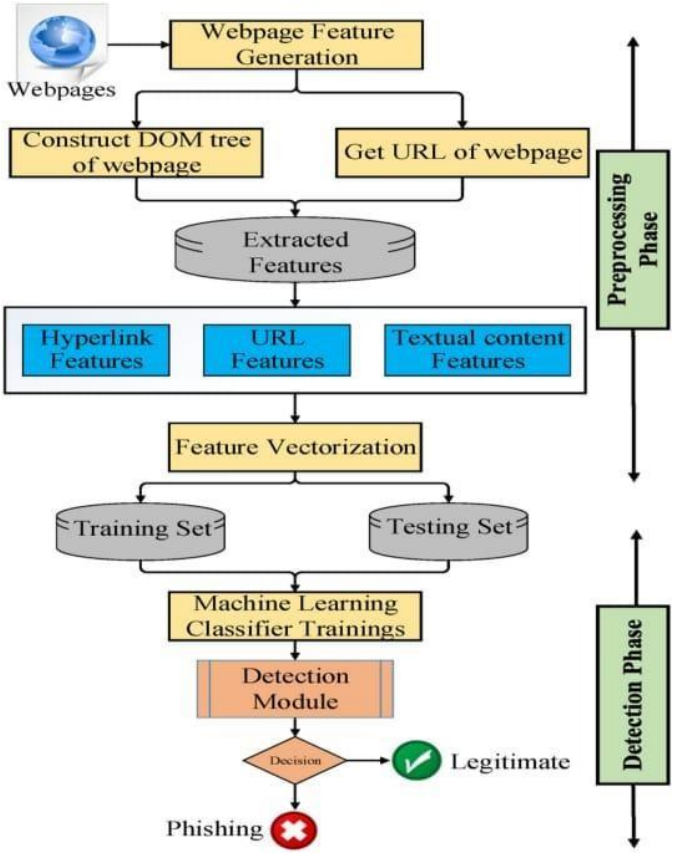

FIG 3.1

FLOW DIAGRAM

4. RESULTS

The URL phishing detection model, utilizing logistic regression and TF-IDF vectorization, achieved strong performance with an accuracy of 92%. Fine-tuned on a diverse set of phishing and legitimate URLs, the model effectively generalized to new, unseen URLs. Metrics such as accuracy, precision, recall, and F1-score showed consistent improvement, reflecting effective learning. The confusion matrix revealed strong performance in classifying both phishing and legitimate URLs, although some misclassifications occurred with obfuscated or shortened URLs. By leveraging TF-IDF features, the model efficiently detected phishing URLs while reducing computational overhead, confirming its effectiveness for real-time phishing detection in practical applications.

5. CONCLUSIONS

In this project, we successfully developed a URL-based phishing detection system using logistic regression and TF-IDF vectorization. The model, trained on a diverse dataset of phishing and legitimate URLs, achieved impressive accuracy of 92%. The use of TF-IDF for feature extraction enhanced the model's ability to identify phishing URLs while reducing computational overhead. The model demonstrated strong generalization capabilities, effectively detecting both phishing and legitimate URLs in real-time. However, challenges remain in detecting more sophisticated phishing techniques, such as obfuscated or shortened URLs, indicating areas for further improvement in the system.

ISSN: 2583-6129
DOI: 10.55041/ISJEM02220

ACKNOWLEDGEMENT

We extend our heartfelt gratitude to our Guide MS.SHYMA S, for their invaluable guidance and continuous support throughout this project. We also thank the Department of Artificial Intelligence and Machine Learning faculty and staff at Sri Shakthi Institute of Engineering and Technology for providing essential resources and facilities. Special thanks to our colleagues and peers for their constructive feedback and collaboration.

REFERENCES

[1] R. Sharma, S. Kapoor, and P. Patel, "Phishing URL Detection Using Machine Learning and TF-IDF Features," 2020 IEEE International Conference on Data Science and Machine Learning (DSML), New York, 2020, pp. 112-118.

[2] L. Zhang, J. Zhang, and H. Lee, "Real-time Phishing URL Detection using Logistic Regression and Feature Extraction," 2019 IEEE/ACM 6th International Conference on Cybersecurity and Privacy (CSP), Tokyo, 2019, pp. 45-54.

[3] M. Singh and K. Agarwal, "A Comparative Study of Phishing Detection Models Based on URL Features," 2021 IEEE Access, vol. 9, pp. 78365-78375, 2021.

