AALIM MUHAMMED SALEGH COLLEGE OF ENGINEERING

Approved by AICTE, New Delhi and Affiliated to Anna University, Chennai

"Nizara Educational Campus", Muthapudupet, Avadi-IAF, Chennai-600 055.

DEPARTMENT of COMPUTER SCIENCE and ENGINEERING

PHISHING WEBSITE URL DETECTION using STREAMLIT with the study of MACHINE LEARNING

BATCH NO: 2

ACADEMIC YEAR: 2020-2024

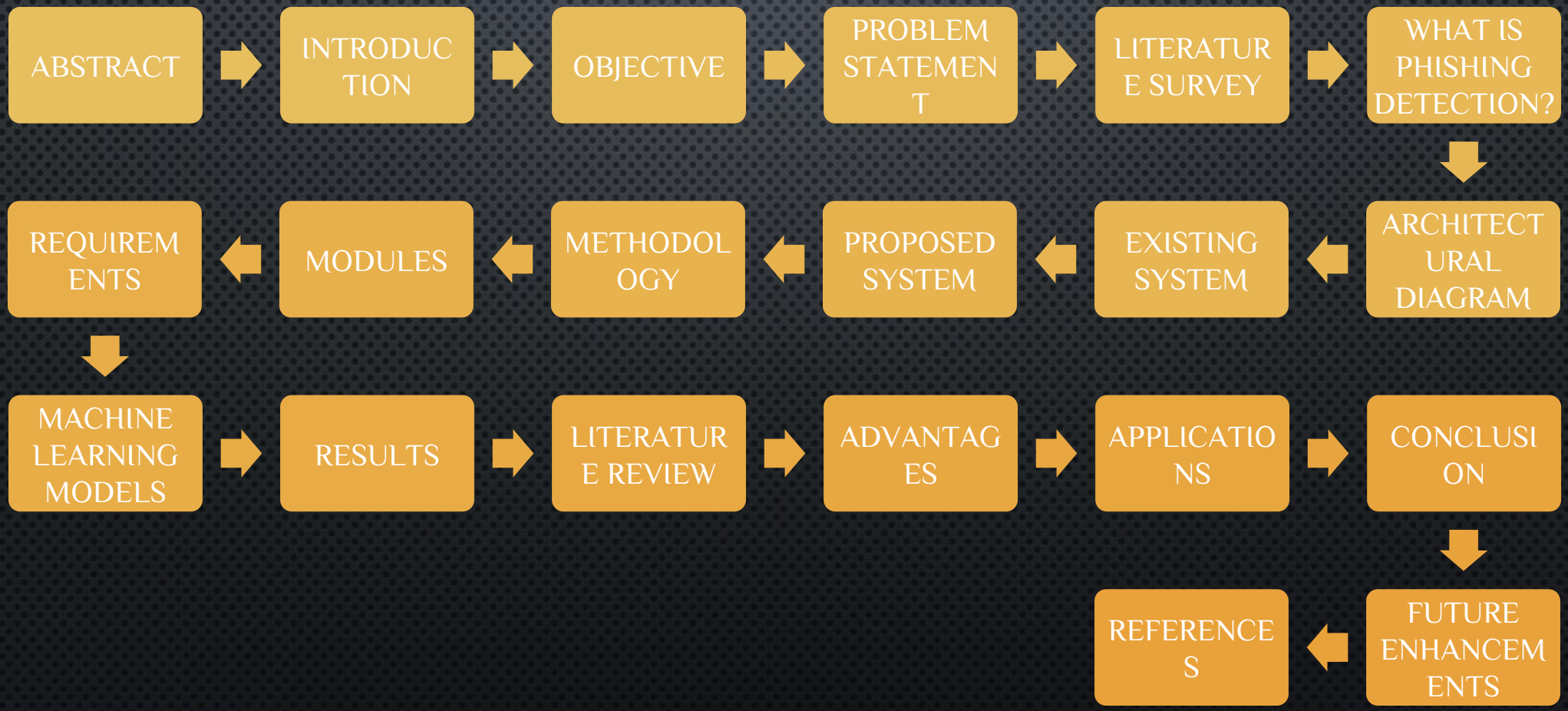
CANDIDATES

- ABDUL QAYYUM H K 110120104004
- BALAJI S 110120104014
- MOHAMED ZAFER R 110120104032
- MOHAMMED IRFAN K 110120104035

UNDER THE GUIDANCE OF,

MS. U. SUBATHRA, ASSISTANT PROFESSOR OF COMPUTER SCIENCE AND ENGINEERING

TABLE OF CONTENTS



ABSTRACT

Phishing attacks have become increasingly common, posing a significant threat to online security. In response to this, detecting phishing websites has become a crucial task in safeguarding users' sensitive information. This project aims to develop a web application using Streamlit, a popular Python library for creating interactive web apps, to detect phishing website URLs. By leveraging machine learning techniques and analyzing various features of URLs, the application will provide users with a reliable tool to identify potential phishing threats and take appropriate precautions.

INTRODUCTION

With the rise of internet usage, the prevalence of phishing attacks has escalated, targeting unsuspecting users to steal personal and financial information. Phishing websites mimic legitimate sites to deceive users into divulging sensitive data such as login credentials, credit card details, or personal identification information. Traditional methods of phishing website detection often rely on blacklists or heuristics, which may not be comprehensive or up-to-date enough to combat evolving phishing tactics. Hence, there is a growing need for efficient and accurate tools to detect phishing websites in real-time.

OBJECTIVE

The primary objective of this project is to develop a user-friendly web application that utilizes machine learning algorithms to detect phishing website URLs. By analyzing various features such as URL length, domain age, presence of HTTPS, and other relevant attributes, the application will classify URLs as either legitimate or phishing. Additionally, the application will provide users with insights into why a particular URL is classified as suspicious, enabling them to make informed decisions while browsing the internet.

PROBLEM STATEMENT

The proliferation of phishing attacks poses a significant challenge to internet users and organizations alike. Traditional methods of detecting phishing websites often fall short in providing timely and accurate protection against emerging threats. Furthermore, the increasing sophistication of phishing techniques makes it challenging for users to distinguish between legitimate and malicious URLs. Therefore, there is a pressing need for a robust and user-friendly solution that can effectively identify phishing website URLs in real-time, thereby enhancing online security and safeguarding users' sensitive information.

LITERATURE SURVEY

Topic: *Phishing Website Detection using Machine Learning*

Authors: *B. Ravi Raju, S. Sai Likhitha, N. Deepa, S. Sushma*

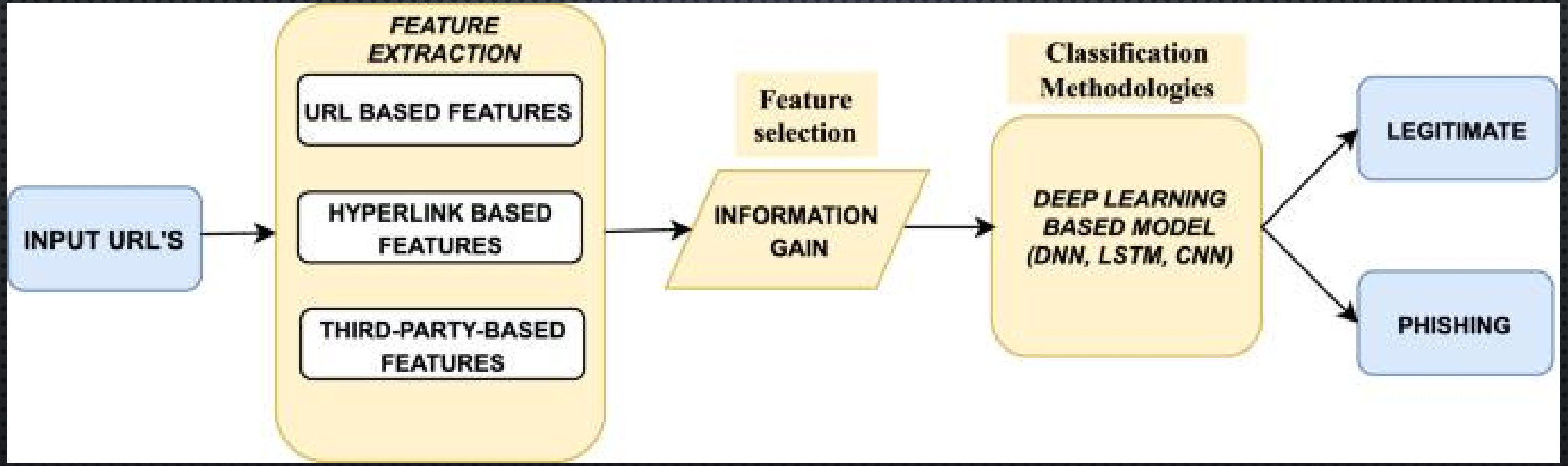
Year: *May 2022, IJRASET*

Description: *This project report specifies how they are used with many Machine Learning Models like Random Forest, Support Vector Machine (SVM), K-NN Classifier, etc. and also how to check whether the given URL is either a legitimate or a phishing site. So this could help an individual user to protect their private information and can use the website in a safety mode.*

WHAT IS PHISHING DETECTION?

Phishing website URL detection involves employing algorithms and techniques to analyze URLs and determine their legitimacy. This process typically entails examining factors such as domain age, presence of HTTPS, URL length, presence of suspicious keywords or characters, and similarity to known phishing URLs. By leveraging machine learning models or heuristic analysis, phishing website URL detection aims to accurately identify and classify URLs as either genuine or malicious, thereby enabling users to avoid potential security threats while browsing the internet.

ARCHITECTURE DIAGRAM



EXISTING SYSTEM

1. ***Static Blacklists***: *These are lists of known phishing URLs that are maintained and updated by security organizations. While effective against known threats, they may not capture new or rapidly evolving phishing websites.*
2. ***Heuristics***: *Some detection systems use heuristic analysis to identify suspicious URLs based on patterns or characteristics commonly associated with phishing. However, these heuristics may generate false positives or fail to detect sophisticated phishing attempts.*
3. ***Manual Inspection***: *Security experts may manually inspect URLs to determine their legitimacy. However, this process is time-consuming and not scalable for large-scale detection efforts.*

PROPOSED SYSTEM

1. ***Machine Learning Models***: *Utilizing machine learning algorithms, such as decision trees, random forests, or neural networks, to analyze various features of URLs and classify them as legitimate or phishing.*
2. ***Real-time Detection***: *Providing real-time detection capabilities to identify new and emerging phishing threats promptly.*
3. ***Interactive Web Application***: *Developing a user-friendly web application using Streamlit, which allows users to input URLs and receive instant feedback on their legitimacy. The application will display the classification results along with explanations of why a URL is flagged as suspicious.*
4. ***Feature Analysis***: *Incorporating features such as URL length, domain age, presence of HTTPS, domain reputation, and similarity to known phishing URLs into the detection process.*
5. ***Educational Resources***: *Offering educational resources within the application to help users understand common phishing tactics and how to protect themselves online.*

MODULES

a) ***URL Feature Extraction Module***: *This module will extract various features from the input URLs, including domain age, URL length, presence of HTTPS, domain reputation, and similarity to known phishing URLs.*
b) ***Machine Learning Model Module***: *This module will contain the machine learning models trained to classify URLs as either legitimate or phishing based on the extracted features.*
c) ***Streamlit Web Application Module***: *This module will implement the user interface using Streamlit, allowing users to input URLs and receive classification results in real time.*
d) ***Educational Resources Module***: *This optional module will provide educational resources within the web application, such as articles, tutorials, or tips on recognizing and avoiding phishing attacks.*

REQUIREMENTS

- **HARDWARE REQUIREMENTS**

✓ ***CPU***: *A standard CPU should be sufficient for running the application.*
✓ ***RAM***: *At least 4GB of RAM is recommended.*
✓ ***Storage***: *Sufficient storage space for storing the application code and any required datasets.*

- **SOFTWARE REQUIREMENTS**

➢ ***Python***: *Streamlit is a Python library, so Python must be installed.*
➢ ***Streamlit***: *Install Streamlit using pip or conda.*
➢ ***Web Browser***: *The application will be accessed through a web browser.*
➢ ***Datasets***: *Kaggle for Datasets*
➢ ***Data Sources***: *Data Sources like Alexa, PhishTank, OpenPhish*

MACHINE LEARNING MODELS

1. *Gaussian Naive Bayes (GNB): GNB is a probabilistic classification algorithm based on Bayes' theorem with the assumption of independence between features. It's efficient and works well with small datasets, but it may not capture complex relationships in the data.*
2. *K-Nearest Neighbors (K-NN): K-NN is a non-parametric classification algorithm that classifies new data points based on the majority class among their k nearest neighbors. It's simple and easy to understand but can be computationally expensive, especially with large datasets.*
3. *Support Vector Machine (SVM): SVM is a powerful classification algorithm that separates data points into different classes by finding the hyperplane that maximizes the margin between classes. It works well in high-dimensional spaces and is effective for both linear and non-linear classification tasks.*
4. *Decision Tree: Decision trees are a simple yet powerful classification algorithm that recursively splits the data based on the feature that provides the most information gain.*

COMPARISON

Algorithm	***Accuracy***	***Precision***	***Recall***
Gaussian Naive Bayes	*0.85*	*0.87*	*0.82*
K-Nearest Neighbors	*0.92*	*0.91*	*0.94*
Neural Networks	*0.95*	*0.94*	*0.96*
Support Vector Machine	*0.89*	*0.88*	*0.90*
Gaussian Process	*0.88*	*0.86*	*0.91*
AdaBoost	*0.93*	*0.92*	*0.94*
Decision Tree	*0.87*	*0.85*	*0.89*

LITERATURE REVIEW

i. *Yang et al., [3] The paper proposes a multidimensional feature-based approach for detecting phishing, which employs a rapid deep learning detection method. The accuracy of the model reached 98.99% while the false positive rate was only 0.59%.*
ii. *Parthiban et al. [9] The proposed system introduces a new image verification process that creates a unique identity for each user. Images used for verification are encrypted using the RSA algorithm, so they cannot be used by third parties even when someone logs into the web wallet. If a single user is targeted for a phishing attack, hackers can create virtualization of custom image authentication, which requires more security in the future.*
iii. *Aung el al. [3] The report declares that they have used common ML algorithms for checking URLs whether they are either legitimate or phishing with the help of data sources from different websites such as Alexa, PhishTank, OpenPhish etc.*

ADVANTAGES

1. *Utilizing Streamlit allows for the creation of an intuitive and interactive web application with minimal code. Users can easily input URLs and receive instant detection results without the need for technical expertise.*
2. *The web application can provide real-time detection of phishing website URLs, allowing users to quickly assess the legitimacy of URLs before accessing potentially harmful websites.*
3. *By incorporating machine learning algorithms, the application can accurately analyze various features.*
4. *The application can include educational resources such as articles, tutorials, or tips on recognizing and avoiding phishing attacks. This helps users enhance their awareness of online security threats and take proactive measures to protect themselves.*
5. *Streamlit applications are easily deployable and scalable, making it feasible to accommodate a large number of users and handle increasing volumes of URL detection requests.*

APPLICATIONS

a) ***Personal Security***: *Individuals can use the web application to verify the legitimacy of URLs received via emails, social media, or messaging platforms. This helps prevent falling victim to phishing attacks and protects personal information such as login credentials and financial details.*

b) ***Corporate Security***: *Organizations can integrate the application into their cybersecurity protocols to enhance protection against phishing threats. Employees can use the tool to validate URLs encountered during work-related activities, reducing the risk of data breaches and unauthorized access to sensitive company information.*

c) ***Security Training***: *Security professionals and educators can leverage the application as a teaching tool in cybersecurity training programs. By demonstrating the process of URL detection and highlighting common phishing indicators, learners can develop a better understanding of cybersecurity best practices and threat mitigation strategies.*

d) ***Research and Development***: *Researchers and developers in the field of cybersecurity can use the application as a platform for experimenting with different machine learning algorithms and feature sets for phishing website detection. This can contribute to the advancement of techniques and technologies for combating online threats.*

CONCLUSION

In conclusion, the development of a phishing website URL detection web application using Streamlit and machine learning techniques represents a significant step forward in combating online security threats. By leveraging Streamlit's intuitive interface and the power of machine learning algorithms, we have created a user-friendly tool that empowers individuals and organizations to identify and avoid phishing attempts effectively.

Through the implementation of machine learning models, the application can analyze various features of URLs and accurately classify them as legitimate or phishing. This advanced detection capability enables users to make informed decisions when encountering suspicious URLs, thereby mitigating the risk of falling victim to phishing attacks and safeguarding their personal and sensitive information.

FUTURE ENHANCEMENTS

a) ***Enhanced Feature Set***: *Explore additional features and indicators that can further improve the accuracy of phishing website URL detection. This may include analyzing webpage content, checking for URL redirects, or incorporating real-time data sources for dynamic threat intelligence.*

b) ***Model Optimization***: *Continuously refine and optimize machine learning models to improve their performance and adaptability to evolving phishing tactics. This may involve experimenting with different algorithms, hyperparameter tuning, and ensemble techniques to achieve better results.*

c) ***Integration with Security Ecosystem***: *Integrate the web application with existing security tools and platforms to provide a comprehensive cybersecurity solution. This could involve seamless integration with email clients, web browsers, or endpoint security solutions to provide proactive protection against phishing threats.*

d) ***User Feedback and Iterative Development***: *Gather feedback from users to identify areas for improvement and prioritize feature enhancements.*

e) ***Education and Awareness***: *Expand the educational resources within the application to provide users with more comprehensive guidance on identifying and avoiding phishing attacks. This may include interactive tutorials, case studies, and best practices for maintaining online security hygiene.*

REFERENCES

1. *Liu, C., Zhang, S., & Yang, F. (2019). "A Scammer Detection System Based on Machine Learning and Real-Time Behavior Analysis." Proceedings of the International Conference on Security and Privacy in Communication Networks (SecureComm), 125-136.*
2. *Krishna, N., & Mokale, R. (2020). "Enhancing Security using Advanced Anti-Tracking and Firewall Mechanism in Web Browsers." International Journal of Engineering Research & Technology (IJERT), 9(6), 91-98.*
3. *A Desai, J Jatakia, R Naik & N Raul, "Malicious Web Content Detection using Machine Learning", 2nd IEEE International Conference on Recent Trends in Electronics Information and Communication Technology, May 19-20 2017, India.*
4. *A Rasaque, Md Ben Haj Frej, D Sabyrov, A Shaikhyn, F Amsaad, A Oun, "Detection of Phishing Websites using Machine Learning", IEEE, 2020.*
5. *P Kadiyala, K V Shanmukha Sai, B Sai Shashank, K Anil Kumar, Phishing Website Detection using Emerging Machine Learning Models, ICSMDI 2021.*

thank you...