PHISHING DETECTION

PROJECT SYNOPSIS

OF MINOR PROJECT

BACHELOR OF TECHNOLOGY

Computer Science and Engineering

SUBMITTED BY

ANMOLPREET KAUR (2203403)

MANPREET KAUR (2203500)

JANUARY 2025

GURU NANAK DEV ENGINEERING COLLEGE, LUDHIANA

Table of contents

Introduction

The widespread use of the internet and email has led to a significant increase in phishing attacks, which are attempts to trick individuals into revealing sensitive information such as passwords, credit card numbers, or personal data. Phishing attacks can have severe consequences, including financial loss, identity theft, and compromised security.

This project aims to develop a machine learning-based phishing detection system that can accurately identify phishing websites and emails. The system will utilize natural language processing (NLP) and machine learning algorithms to analyze the content of websites and emails and detect suspicious patterns.

The project will employ a range of technologies, including Python, scikit-learn, and NLTK, to develop a robust and efficient phishing detection system. The system will be trained on a dataset of labeled phishing and legitimate websites and emails, and its performance will be evaluated using metrics such as accuracy, precision, and recall.

The project is specialized in the field of cybersecurity and artificial intelligence, and it requires knowledge of machine learning, NLP, and web development. Some special technical terms used in this project include phishing, machine learning, NLP, supervised learning, and classification algorithms.

The goal of this project is to contribute to the development of more effective phishing detection systems, which can help protect individuals and organizations from the growing threat of phishing attacks.

Rationale

Phishing attacks have become a significant threat to individuals, businesses, and organizations worldwide. According to recent statistics, phishing attacks account for over 90% of all security breaches, resulting in billions of dollars in losses annually. The increasing sophistication of phishing attacks, combined with the growing reliance on online services, has created a pressing need for effective phishing detection systems. Existing solutions often rely on manual reporting, keyword filtering, or simple machine learning algorithms, which can be easily evaded by attackers.

A more advanced phishing detection system, leveraging machine learning and natural language processing, is urgently needed to combat the evolving threat landscape. By developing a robust and accurate phishing detection system, this project aims to provide a critical layer of protection for individuals and organizations, helping to prevent financial losses, protect sensitive information, and maintain trust in online service.

Objectives

1.To design and develop a phishing detection system.

2.To classify phishing emails using a machine learning based approach.

3.To implement a scalable real based phishing detection system.

Literature Review

Phishing detection has been an active area of research in recent years. Several studies have proposed various techniques to detect phishing attacks. Here, we review five relevant papers:

1. "Phishing Detection using Machine Learning" by R. Basnet et al. (2018) [1] - This paper proposes a machine learning-based approach for phishing detection, using features such as URL length, presence of suspicious keywords, and email content.

2. "A Study on Phishing Email Detection using Natural Language Processing" by S. K. Goyal et al. (2020) [2] - This paper explores the use of natural language processing (NLP) techniques for phishing email detection, including sentiment analysis and topic modeling.

3. "Deep Learning for Phishing Detection" by Y. Li et al. (2019) [3] - This paper proposes a deep learning-based approach for phishing detection, using convolutional neural networks (CNNs) to analyze URL and email content.

4. "Phishing Website Detection using URL Features" by A. K. Singh et al. (2019) [4] - This paper extracts URL features such as length, depth, and presence of suspicious keywords to detect phishing websites.

5. "Ensemble-based Phishing Detection" by H. M. Al-Khateeb et al. (2020) [5] - This paper proposes an ensemble-based approach for phishing detection, combining multiple machine learning algorithms to improve detection accuracy.

References

- Akash, A., et al. (2016). URL-based phishing detection using machine learning. Journal of Cybersecurity, 5(2).

- Chandran, A., et al. (2014). Phishing detection using email header analysis. International Journal of Information Security, 12(3).

- Gao, Y., et al. (2021). Scalable phishing detection using distributed machine learning. Cybersecurity Research Journal, 8(1).

- Haris, M., et al. (2020). Visual-based phishing detection using image processing techniques. International Journal of Computer Vision, 28(4).

- Kang, C., et al. (2019). Ethical considerations in phishing detection technologies. Ethics and Information Technology, 21(2).

- Luo, J., et al. (2017). Hybrid phishing detection systems using multiple machine learning models. Cybersecurity and Privacy Journal, 4(5).

- Mishra, A., et al. (2015). URL-based phishing detection through heuristic techniques. Journal of Cyber Defense, 10(2).

Feasibility study

The proposed project aims to develop a phishing detection system that can accurately identify phishing websites and emails. This feasibility study assesses the project's viability, need, and significance.

Technical Feasibility

The project is technically feasible, as it leverages existing machine learning and natural language processing techniques. The required technologies, such as Python, scikit-learn, and NLTK, are widely available and well-documented.

Economic Feasibility

The project is economically feasible, as it can be developed with minimal infrastructure and personnel costs. The benefits of the project, including improved security and reduced financial losses, outweigh the costs.

Operational Feasibility

The project is operationally feasible, as it can be integrated with existing email and web applications. The system can be easily maintained and updated, ensuring its continued effectiveness.

Need and Significance

The need for a phishing detection system is evident, given the increasing frequency and sophistication of phishing attacks. The proposed system can significantly improve security, reduce financial losses, and enhance user trust in online services.

5. Conclusion

The feasibility of phishing detection is high across multiple dimensions, but it depends on the specific requirements, resources, and scope of the implementation. Technically, phishing detection systems leveraging machine learning, heuristic methods, and AI are effective and scalable, though they require continuous updates and adaptation to new phishing techniques.

From a financial perspective, while initial development and maintenance costs can be significant, the return on investment in terms of preventing financial fraud, data breaches, and reputational damage makes phishing detection a sound investment for most organizations.

Operationally, phishing detection is feasible if proper integration, user education, and support mechanisms are in place. Finally, legal and ethical considerations can be managed by ensuring compliance with data privacy regulations and implementing privacy-conscious detection techniques.

In conclusion, the development and implementation of phishing detection systems are technically, financially, and operationally feasible, with strong potential benefits in terms of security and risk mitigation.

Methodology/Planning of work

1. Data Collection: Collect a dataset of labeled phishing and legitimate websites/emails from publicly available sources.

2. Data Preprocessing: Clean and preprocess the data by removing duplicates, handling missing values, and converting data formats.

3. Feature Extraction: Extract relevant features from the dataset, such as URL parameters, email headers, and content.

4. Model Training: Train a machine learning model using the extracted features and labeled dataset.

5. Model Evaluation: Evaluate the performance of the model using metrics such as accuracy, precision, and recall.

Facilities required for propose work

Software Requirements

1. Programming languages: Python, R, or Java for machine learning model development
2. Machine learning frameworks: TensorFlow, PyTorch, or Scikit-learn for building and training models
3. Data analysis tools: Pandas, NumPy, and Matplotlib for data preprocessing and visualization
4. Web development frameworks: Flask or Django for building the web application
5. Database management systems: MySQL or MongoDB for storing and managing data

Hardware Requirements

1. Computer: A laptop or desktop with a minimum of 8 GB RAM and 256 GB storage
2. Processor: A multi-core processor (at least 2 cores) with a minimum clock speed of 2.5 GHz
3. Operating System: A 64-bit operating system (Windows, macOS, or Linux)
4. Internet connection: A stable internet connection for data collection and testing

Additional Requirements

1. Phishing dataset: A labeled dataset of phishing and legitimate emails for training and testing the machine learning model
2. Web server: A web server for hosting the web application
3. Testing tools: Tools like Selenium or Apache JMeter for testing the web application's functionality and performance.

Expected outcomes

Software Requirements

1. Programming languages: Python, R, or Java for machine learning model development
2. Machine learning frameworks: TensorFlow, PyTorch, or Scikit-learn for building and training models
3. Data analysis tools: Pandas, NumPy, and Matplotlib for data preprocessing and visualization
4. Web development frameworks: Flask or Django for building the web application
5. Database management systems: MySQL or MongoDB for storing and managing data

Hardware Requirements

1. Computer: A laptop or desktop with a minimum of 8 GB RAM and 256 GB storage
2. Processor: A multi-core processor (at least 2 cores) with a minimum clock speed of 2.5 GHz
3. Operating System: A 64-bit operating system (Windows, macOS, or Linux)
4. Internet connection: A stable internet connection for data collection and testing

Additional Requirements

1. Phishing dataset: A labeled dataset of phishing and legitimate emails for training and testing the machine learning model
2. Web server: A web server for hosting the web application
3. Testing tools: Tools like Selenium or Apache JMeter for testing the web application's functionality and performance.

References

[1] S. Garera, N. Provos, M. Chew, and D. D. Song, "A framework for detection and measurement of phishing attacks," in Proceedings of the 2007 ACM Workshop on Recurring Malcode, pp. 1-8.

[2] I. Fette, N. Sadeh, and A. Tomasic, "Learning to detect phishing emails," in Proceedings of the 16th International Conference on World Wide Web, pp. 649-656.

[3] R. B. Rao and S. P. Salai, "Detection of phishing websites using machine learning techniques," in Proceedings of the 2018 International Conference on Computing, Power and Communication Technologies, pp. 1385-1390.

[4] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," in Proceedings of the 2007 ACM Workshop on Artificial Intelligence and Security, pp. 60-69.