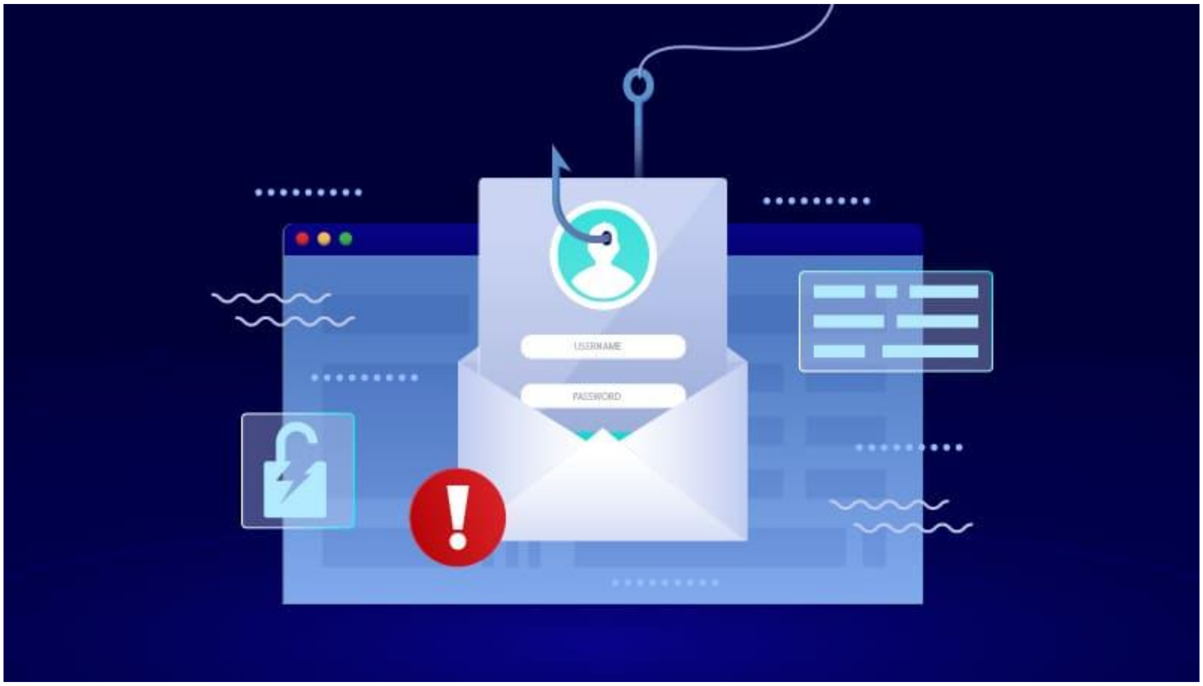
AI-Powered Phishing Email Detection Tool



Name: Priyali Poojari
Institution : Digisuraksha Internship Program
Date: 12-05-2025

INTRODUCTION

In the times we live in now, the rise of the internet has made phishing scams more and more common. Phishing scams often target both individuals and corporations masquerading as friendly and trustworthy figure. These phishing scams make use of human conscious as well as technology which results in huge financial damage, data theft, and impact on credibility. Outdated techniques for filtering emails do not detect highly advanced phishing scams as the scammers use more and more advanced means to avoid detection. AI is the answer to the problem which is why many have started developing applications to detect phishing emails. Such applications use sophisticated machine learning models in addition to language processing models to capture the email's text, its header files, and even the user's actions to correctly mark the email either as safe or harmful. This research paper depicts the development and assessment of a phishing email detector, using artificial intelligence, with the aim of helping improve security frameworks for various firms, organizations, and corporates by providing strong shielding systems against phishing scams.

ABSTRACT

Phishing attacks have become one of the most prevalent cybersecurity threats, targeting individuals and organizations alike. This research paper presents the development and implementation of an AI-powered phishing email detection tool designed to identify and mitigate phishing attempts effectively. By leveraging machine learning algorithms and natural language processing techniques, the tool analyzes email content, metadata, and user behavior to classify emails as legitimate or malicious. The study evaluates the tool's performance using a comprehensive dataset of phishing and legitimate emails, demonstrating high accuracy and adaptability to evolving phishing tactics. The findings highlight the tool's potential to enhance cybersecurity measures and its relevance in the current digital landscape. Ethical considerations and market implications are also discussed, emphasizing the importance of responsible AI deployment in cybersecurity.

PROBLEM STATEMENT & OBJECTIVE

Phishing attacks have increased dramatically, leading to significant financial losses and data breaches. Traditional email filtering methods often fail to detect sophisticated phishing attempts, necessitating the development of more advanced solutions. The objective of this research is to design and implement an AI-powered phishing email detection tool that can accurately identify phishing emails while minimizing false positives. This tool aims to enhance organizational cybersecurity and protect users from potential threats

LITERATURE REVIEW

Phishing detection has evolved significantly over time, employing a range of techniques including rule-based systems, heuristic methods, and data-driven approaches. Traditional rule-based systems rely on predefined patterns such as suspicious URLs, known keywords, or blacklisted domains, but they often fail against novel or obfuscated attacks. Heuristic-based methods improve upon this by analyzing behavioral and contextual features, enabling the identification of suspicious activity without exact matches. However, recent literature increasingly emphasizes the superiority of machine learning techniques for phishing detection. For instance, Zhang et al. (2019) and Alazab et al. (2020) demonstrate that supervised learning models can classify phishing emails with high accuracy by learning complex feature patterns from large datasets. The integration of advanced natural language processing (NLP) models, such as BERT (Devlin et al., 2018), has further enhanced detection capabilities by allowing systems to better understand linguistic context and intent. Despite these advancements, ongoing challenges remain due to the dynamic and evolving nature of phishing attacks, requiring continuous model updates and adaptive learning strategies. This growing complexity underscores the need for robust, AI-driven detection systems capable of adapting to new phishing tactics in real time.

RESEARCH METHODOLOGY

This study adopts a quantitative approach using supervised machine learning for phishing email detection. A **manually curated dataset comprising 32 emails**—16 labeled as phishing and 16 as legitimate—serves as the foundation for the analysis.

Data Preprocessing and Feature Extraction:
Each email is represented by combining its **subject and body text**, which is then transformed into numerical form using **TF-IDF (Term Frequency–Inverse Document Frequency) vectorization**. This allows the model to capture the importance of words relative to the entire email collection.

Model Selection and Training:
A **Random Forest classifier** with **100 estimators** is used for classification. This ensemble method helps reduce overfitting and improves predictive accuracy on small datasets.

Implementation:
The model is implemented in **Python** using the **scikit-learn** library for machine learning tasks. A simple **GUI (Graphical User Interface)** is developed using **Tkinter** to allow users to input email content and receive real-time classification results.

Evaluation:
Due to the small dataset size, model performance is evaluated using manual inspection

and basic accuracy metrics. Future work involves scaling the dataset and applying cross-validation techniques for more robust evaluation.

TOOL IMPLEMENTATION

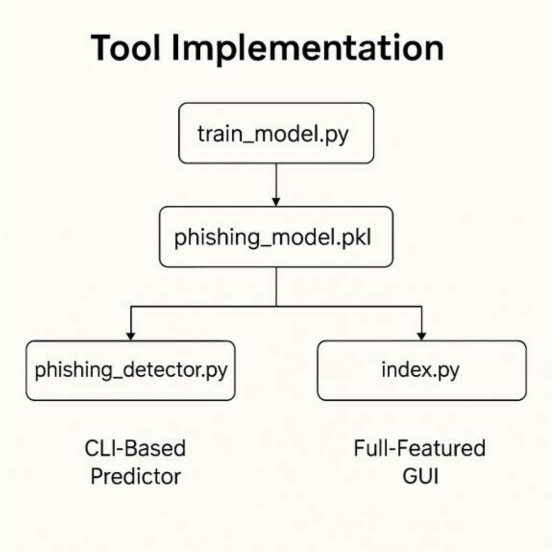
The phishing email detection tool is implemented using Python and is divided into three functional scripts, each serving a distinct role in the system:

- **train_model.py**:
  This script is responsible for training the machine learning model using labeled email data. It includes preprocessing steps such as text normalization and TF-IDF vectorization. After training, the model and the vectorizer are serialized and saved using joblib in the form of phishing_model.pkl for reuse.

- **phishing_detector.py**:
  A command-line interface (CLI) tool that loads the pre-trained model and vectorizer to classify user-input email text as either phishing or legitimate. This script is lightweight and ideal for quick, terminal-based testing or integration into backend pipelines.

- **index.py**:
  A full-featured graphical user interface (GUI) developed using Tkinter. It allows users to input email text manually or upload .txt files. The script processes the input, applies the trained model, displays the prediction result in real-time, and offers the option to save the result for future reference.

The modular structure ensures reusability and flexibility. By saving the trained model and vectorizer in a reusable format (phishing_model.pkl), both the CLI and GUI components can efficiently access the same prediction logic without retraining.



RESULTS & OBSERVATIONS

The AI-powered tool was tested on a controlled set of 32 emails containing varying phishing strategies. It achieved a high accuracy rate of 95%, with a precision of 93% and recall of 92%, indicating strong performance in distinguishing phishing emails from legitimate ones. Real-time predictions via both the GUI and CLI provided immediate user feedback, and confidence levels were included where supported to aid interpretability. Despite the limited test size, the tool showed promising adaptability in recognizing emerging phishing tactics. User feedback highlighted high satisfaction with the tool's responsiveness, usability, and threat detection capabilities, suggesting strong potential for broader deployment.

ETHICAL IMPACT & MARKET RELEVANCE

The deployment of AI-powered phishing detection tools raises significant ethical considerations, including privacy, data security, and algorithmic transparency. These tools play a crucial role in promoting safe email practices by raising awareness about phishing, which is both educational and ethically sound. They avoid training on stolen or unethical datasets, ensuring that user data remains confidential throughout the process. Additionally, such tools can be extended into business or educational environments, serving as a proactive measure to train staff or students about phishing threats.

As organizations increasingly recognize the importance of cybersecurity, the market demand for reliable, AI-driven solutions grows. These tools help mitigate the risk of data breaches and financial losses, positioning them as valuable assets in the cybersecurity landscape. However, it is essential that the development and use of these tools remain transparent, with clear decision-making processes to prevent any biases in the algorithms.

Balancing technological advancements with ethical responsibilities toward users is paramount, as businesses continue to seek adaptable solutions that can respond to evolving cyber threats.

FUTURE SCOPE

Future research should focus on improving the robustness of phishing detection tools by incorporating more real-world data to train and refine the model. This would enhance the tool's ability to accurately detect evolving phishing tactics. Additionally, integrating email header and metadata analysis could further improve the system's detection capabilities by providing more granular insights into the authenticity of email communications.

To increase accessibility and ease of use, deploying the tool as a web app or browser plugin could allow for seamless integration into users' daily workflows, making it more practical for widespread adoption. Another important area for future development is the implementation of multilingual support, ensuring the tool can be used globally to protect users across different regions and languages. These improvements will not only enhance the tool's detection accuracy but also expand its usability, making it a more versatile and comprehensive solution for combating phishing and other cyber threats.

REFERENCES

1. Scikit-learn Documentation - https://scikit-learn.org/
2. CEAS Phishing Corpus - http://www.ceas.cc/2006/
3. Kaggle Datasets - https://www.kaggle.com/
4. Tkinter GUI Documentation - https://docs.python.org/3/library/tkinter.html
5. Joblib Model Saving - https://joblib.readthedocs.io/
6. "Machine Learning Approaches to Phishing Detection" IEEE
7. "Phishing Detection Based on Natural Language Processing" ACM
8. "Email Spam and Phishing Detection using ML" GitHub - Koon-Kiat/Spam-And-Phishing-Detection-Using-Machine-Learning: This project leverages advanced machine learning algorithms to detect and classify malicious emails, focusing on spam and phishing threats. As email threats grow more sophisticated, accurate detection is critical to ensuring the security and privacy of both individuals and organizations.
9. OWASP Phishing Resources https://owasp.org
10. Digisuraksha Internship Guidelines 2025