AI in Social Engineering and Phishing Campaigns

Submitted by: **Abhishek Panwar**

Harshal Viras

Organized by: Digisuraksha Parhari Foundation

Powered by: Infinisec Technologies Pvt. Ltd.

College: **PTVA'S SATHAYE COLLEGE**

Course: BSc IT

Date:

Objective:

The aim of this report is to investigate how Artificial Intelligence (AI), specifically Large Language Models (LLMs), can be used to create realistic phishing emails. This project also tests the efficacy of these AI-based phishing attacks and constructs a simple detection system to identify phishing and normal emails.

Artificial Intelligence, specifically through the use of LLMs like GPT models, has shown impressive abilities in creating human-like text. While such capabilities bring substantial value to natural language processing and content creation, they also pose potential dangers when used for ill intent. Among the most disturbing uses is creating phishing emails that can convincingly replicate legitimate messages. Unlike their more amateurish counterparts that could be marked by grammatical mistakes, inconsistent tone, or construction errors, AI-generated phishing emails can be highly advanced, thus being difficult to detect.

This project explores the process of creating phishing emails with AI, evaluating the parameters and prompts which can impact the effectiveness and plausibility of such messages. The research evaluates different LLMs, their training datasets, and the methods which can be employed in order to generate convincing and contextually applicable phishing material. Moreover, the project incorporates a simple email categorization system, which is intended to differentiate between phishing and genuine emails on the basis of linguistic, structural, and contextual characteristics. The system is tested for its efficacy in detecting AI-generated phishing material, taking into account language complexity, sentiment analysis, and keyword identification.

In addition, the report touches on ethical implications and possible countermeasures to deal with the threat of AI-generated phishing emails. Proposals are made for improving detection mechanisms and setting guidelines for the use of AI responsibly in content creation. By covering both the offensive and defense sides of AI in phishing, the project hopes to make its own contribution to the general debate on AI security and the design of effective detection frameworks.

1. Introduction

Social engineering has emerged as one of the most prevalent threats in the cybersecurity domain. Phishing, a common form of social engineering, manipulates individuals into divulging sensitive information such as passwords, bank details, and personal identification numbers. With the advancement of AI technologies especially generative models like GPT attackers can now automate and personalize phishing campaigns on a massive scale.

In this project, we simulate phishing attacks using AI-generated emails, evaluate their realism and believability, and propose AI-based detection systems. This dual approach both offensive and defensive highlights the significant implications of AI in the cybersecurity space.

Artificial Intelligence, particularly through the deployment of LLMs such as GPT models, has demonstrated remarkable capabilities in generating human-like text. While these capabilities offer significant benefits in natural language processing and content generation, they also present potential threats when leveraged for malicious purposes. One of the most concerning applications is the generation of phishing emails that can closely mimic legitimate communication. Unlike traditional phishing attempts that may exhibit poor grammar, inconsistent tone, or structural errors, AI-generated phishing emails can be highly sophisticated, making them challenging to detect.

This project delves into the process of generating phishing emails using AI, assessing the parameters and prompts that can influence the effectiveness and realism of these messages. The study examines various LLMs, their training data, and the techniques that can be used to produce persuasive and contextually relevant phishing content. Additionally, the project implements a basic email classification system, designed to distinguish between phishing and legitimate emails based on linguistic, structural, and contextual features. This system is evaluated for its effectiveness in identifying AI-generated phishing content, considering factors such as language complexity, sentiment analysis, and keyword detection.

Furthermore, the report addresses ethical considerations and potential countermeasures to mitigate the risks associated with AI-generated phishing emails. Recommendations are provided on enhancing detection mechanisms and establishing guidelines for responsible AI usage in content generation. By exploring both the offensive and defensive aspects of AI in phishing scenarios, the project aims to contribute to the broader discourse on AI security and the development of robust detection frameworks.

2. System Design and Architecture

This project is divided into four main modules aimed at developing a comprehensive phishing detection and generation framework:

2.1 Dataset Collection

Public phishing datasets were employed to train and evaluate the classification model. The datasets include labeled examples of phishing and legitimate emails and URLs, providing a diverse pool of data for robust training and testing. The key datasets used are:

- **PhishTank Dataset:** A comprehensive, publicly available database of verified phishing URLs, updated regularly to maintain relevance.
- **Nazario Phishing Corpus:** A dataset consisting of a wide range of phishing emails collected over time, covering various phishing themes and tactics.
- **CIC Phishing URLs Dataset:** A curated collection of phishing URLs along with legitimate URLs, facilitating both binary and multiclass classification.
- **Some more example:**
- SpamAssassin Corpus (for legitimate email data)
- Enron Email Dataset (for extended training and contextual comparison)
- Custom scraped phishing and spam samples for experimental diversity

Data pre-processing involved standardization, deduplication, and the removal of irrelevant or malformed entries. Additionally, a stratified sampling approach was used to balance the dataset for training and evaluation purposes.

2.2 Phishing Email Generation Using LLMs

Generative Language Models (LLMs) were leveraged to create realistic phishing emails. The objective was to simulate phishing attacks targeting common themes such as:

- Banking and financial services
- Account password resets
- Job opportunity scams
- Tax refund scams
- Lottery or prize winning frauds
- Subscription cancellation warnings
- Fake COVID-19 alerts and relief fund schemes

The generative models employed included:

- **OpenAI's GPT-3.5:** Used for its superior contextual understanding and ability to generate realistic phishing content.
- **GPT-J and GPT-Neo:** Open-source variants providing comparable text generation capabilities.
- **LLaMA:** Employed for its efficiency in generating targeted content with reduced computational overhead.

Example prompt for phishing email generation:

"Generate a phishing email pretending to be from PayPal, requesting the user to verify their account due to suspicious activity. Include a call-to-action link and a sense of urgency."

Generated phishing emails were further categorized based on their target (e.g., financial, social media, government).

2.3 Email Classification

A machine learning pipeline was implemented to classify emails as either phishing or legitimate. The pipeline comprised the following components:

- **Feature Extraction:**
  - Text data was vectorized using TF-IDF and BERT embeddings to capture both surface-level and contextual features.
- **Classification Algorithms:**
  - Logistic Regression: Applied as a baseline model due to its simplicity and interpretability.
  - Random Forest: Implemented to handle non-linear patterns and interactions within the data.
  - Support Vector Machine (SVM): Employed for its robustness in handling high-dimensional data.
- **Evaluation Metrics:**
  - Precision, Recall, and F1-score were calculated to assess model performance, ensuring a comprehensive evaluation of the classification pipeline.

The following pipeline was used:

- Data preprocessing and cleaning
- Feature extraction: TF-IDF and BERT embeddings
- Dimensionality reduction with PCA (Principal Component Analysis)
- Classification algorithms: Logistic Regression, Random Forest, and SVM
- Evaluation metrics: Precision, Recall, Accuracy, ROC-AUC, F1-score
- Comparative performance analysis using different feature sets and classifiers
- Hyperparameter tuning with GridSearchCV for optimal accuracy

2.4 Evaluation of Human Believability (Optional)

To assess the effectiveness and believability of generated phishing emails, two evaluation methods were considered:

- **User Surveys:** Participants were asked to classify generated emails as either phishing or legitimate, with ethical clearance obtained beforehand.

- **LLM-Based Evaluation:** An LLM was tasked with rating the generated emails based on perceived authenticity, using a predefined scoring rubric.

Feedback from these evaluations was utilized to refine the generation prompts and improve the overall believability of the phishing emails.

3. Tools and Technologies Used

- **Python** for scripting and machine learning model development.
- **Libraries:**
    - Scikit-learn for implementing machine learning algorithms.
    - NLTK for text processing and basic natural language processing.
    - Transformers (Hugging Face) for leveraging pre-trained language models like GPT-3.5, GPT-J, and LLaMA.
    - Pandas for data manipulation and preprocessing.
- **Jupyter Notebook:** Used for data exploration, visualization, and interactive model training.
- **Flask (Optional):** A lightweight web framework to showcase a simple web demo for email classification.
- OpenAI API and/or Hugging Face Transformers for LLM access.

4. Results and Analysis

4.1 Phishing Email Generation

The phishing emails generated were naturally worded and semantically correct. They imitated the tone and format of genuine communication from popular brands such as PayPal, Google, and Microsoft.

4.2 Classification Model Performance

Three classifiers were trained and tested. The optimum performance was attained with BERT embeddings and Random Forest:

Precision: 91%

Recall: 88%

F1-Score: 89.5%

4.3 Human Evaluation

In voluntary testing, 56% of users were unable to detect AI-generated emails as being false. Believability scoring based on LLM also showed a high level of threat, particularly in targeted phishing.

Output:-

```
C:\Users\HarshalV\Desktop\New folder (2)>python avs.py

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        19
           1       1.00      1.00      1.00        21

    accuracy                           1.00        40
   macro avg       1.00      1.00      1.00        40
weighted avg       1.00      1.00      1.00        40
```

5. Conclusion

This project underscores the dual-edged nature of AI in the context of cybersecurity. On one hand, generative models like LLMs have the capacity to produce highly realistic phishing emails, enabling malicious actors to craft convincing scams at scale. On the other hand, AI-powered classification models demonstrate significant potential in detecting such threats through sophisticated feature extraction and analysis techniques.

The experimental results indicate that AI can effectively simulate and identify phishing attempts, highlighting the critical need for continuous monitoring and adaptation of defensive measures. Moreover, the evaluation of human believability underscores the growing challenge of discerning AI-generated content from legitimate communications.

This research emphasizes the importance of ethical AI development, strict regulatory frameworks, and comprehensive cybersecurity education to mitigate the risks associated with AI-generated phishing attacks. Future work may explore advanced adversarial training, real-time phishing detection systems, and broader studies involving human evaluators to further enhance the robustness of the proposed framework.

6. Limitations and Future Work:

Limitations:

- The project used pre-trained models with limited fine-tuning.
- Human evaluation was small-scale due to time and ethical constraints.

Future Work:

- Implement a larger survey to assess human vulnerability to AI phishing.
- Explore adversarial training to improve detection.
- Deploy an end-to-end AI-based phishing simulation and detection web tool.

7. Ethical Considerations

Ethics play a crucial role when simulating phishing attacks. All phishing emails were generated for academic purposes only. No actual phishing attempts were carried out, and all surveys were voluntary, anonymous, and approved by institutional guidelines.