AI IN EMAIL GENERATION AND PHISHING DETECTION

BY DURVAAS MORE AND JANHAVI SONAVALE

CONTENTS

|  |  |  |
| --- | --- | --- |
| Sr.No. | Topic | Page |
| 1. | Abstract | 3 |
| 2. | Problem Statement | 4 |
| 3. | Objective | 5 |
| 4. | Literature Review | 6 |
| 5. | Research Methodology | 7 |
| 6. | Tool Implementation | 9 |
| 7. | Results | 10 |
| 8. | Observation | 10 |
| 9. | Ethical Impact and Market Relevance | 11 |
| 10. | Future Scope | 12 |
| 11. | References | 13 |

ABSTRACT

The rise of Artificial Intelligence (AI) has significantly impacted both the generation and detection of phishing emails in cybersecurity. Malicious actors are increasingly exploiting AI-powered tools, including large language models, to craft sophisticated, personalized phishing emails that closely mimic legitimate communication. These messages often use social engineering tactics—such as urgency, authority, and emotional manipulation—to deceive users into revealing sensitive information. Simultaneously, AI is being employed to defend against such threats. Using Natural Language Processing (NLP) and Machine Learning (ML), AI systems can analyze email content, detect suspicious language patterns, and identify anomalies in user behavior. These systems also examine metadata, hyperlinks, and sender authenticity to assess potential risks. By continuously learning from new data, AI-driven solutions improve over time, enabling real-time detection and filtering of phishing attempts. This dual role of AI—as both a tool for attackers and a critical line of defense—underscores the evolving complexity of email-based threats and the necessity for organizations to adopt intelligent, adaptive security solutions.

PROBLEM STATEMENT

The integration of Artificial Intelligence (AI) into cybercrime has significantly escalated the complexity and impact of social engineering and phishing campaigns. Attackers are now capable of generating highly convincing emails, manipulating multimedia (e.g., deepfake audio or video), and exploiting human psychology through AI-generated content. These developments have rendered traditional detection methods less effective and increased the risk of successful cyberattacks.AI's dual-use nature is a central concern—while it strengthens cyber defenses, it simultaneously empowers attackers. The growing use of AI in phishing attacks calls for urgent research and development of intelligent countermeasures. Moreover, the human factor remains the most exploited vulnerability, as users are often unable to distinguish between authentic and manipulated content.

Key challenges include:

Sophisticated Attack Generation: AI-generated phishing emails are personalized, context-aware, and linguistically refined, making them difficult to detect.

Human Vulnerability: Emotional triggers like urgency or authority are exploited by AI to manipulate victims into acting quickly without verifying the source.

Limitations of Traditional Detection Tools: Rule-based or static filters are not adaptive enough to detect novel AI-generated threats.

OBJECTIVE

This research aims to explore the dual role of Artificial Intelligence (AI) in the context of social engineering and phishing attacks—both as a tool for cybercriminals and as a defense mechanism for cybersecurity systems. The study focuses on understanding how AI is used to generate and detect phishing content, evaluates the effectiveness of existing AI-based tools, and proposes enhancements to strengthen digital defenses.

The specific objectives of this research are:

To analyze how AI is leveraged by attackers to create realistic phishing emails, voice manipulations (vishing), and deepfake content.

To investigate the psychological and behavioral tactics exploited through AI-powered phishing.

To evaluate existing AI-based phishing detection tools, focusing on Natural Language Processing (NLP) and Machine Learning (ML) techniques.

To develop or assess a prototype tool (e.g., PhishSenseAI) for detecting and generating phishing emails.

To explore real-world phishing incidents enhanced by AI, highlighting vulnerabilities and consequences.

To identify limitations in current detection systems, including issues of explainability, multilingual support, and adaptability.

To propose future enhancements such as multimodal detection, behavioral analytics, and integration with real-time email clients.

LITERATURE REVIEW

The increasing sophistication of phishing attacks has led to a growing body of research on how Artificial Intelligence (AI) can both facilitate and combat cyber threats. Early studies focused on rule-based and signature-based detection systems, which have proven inadequate in dealing with AI-generated phishing emails that evolve rapidly and mimic legitimate communication more effectively.Recent advancements in Natural Language Processing (NLP) and Machine Learning (ML) have shifted the focus toward adaptive and intelligent systems. AI models such as Naive Bayes, Support Vector Machines (SVM), and deep neural networks have been employed to classify phishing versus legitimate emails with higher accuracy. For instance, Sahingoz et al. (2019) demonstrated that NLP combined with deep learning significantly improves phishing email classification by analyzing the semantic structure and linguistic features of messages.

Researchers have also explored the offensive use of AI. Tools like GPT-based models have shown the potential to generate phishing emails that bypass traditional spam filters and fool users through context-aware language generation. This raises ethical concerns and emphasizes the need for Explainable AI (XAI) to improve trust and transparency in detection systems.

Key findings from existing literature include:

AI-Powered Attacks: Studies confirm that attackers use AI to automate and scale social engineering, making scams more targeted and believable.

RESEARCH METHODOLOY

Overview

The research focuses on the detection of phishing emails using artificial intelligence, specifically leveraging natural language processing (NLP) and machine learning (ML) techniques. The methodology involves collecting and preprocessing data, extracting relevant features, training and evaluating ML models, and validating the system’s effectiveness on real-world cases.

Steps Involved

Data Collection:The dataset consists of labeled phishing and legitimate emails. Data sources include open repositories such as PhishTank for phishing emails and trusted datasets for legitimate emails. The dataset is typically stored in a CSV file (e.g., phishing\_emails.csv).

Data Preprocessing:The raw email data is cleaned to remove duplicates, irrelevant entries, and noise. Text normalization (lowercasing, removing special characters) and tokenization are applied.

Feature Extraction:

NLP techniques are used to extract features from email content, including:

Word frequency (using CountVectorizer)

Presence of suspicious keywords

Analysis of sender and recipient fields

URL characteristics (length, special characters, domain age)

Email metadata (subject, body, attachments)

Model Selection:

Several supervised machine learning classifiers are considered, such as:

Naive Bayes (for text classification)

Decision Tree

Gradient Boosting

Multi-Layer Perceptron (MLP)

Model Training and Validation:

The dataset is split into training and testing sets (commonly 80-20 split). Models are trained on the training data and evaluated on the test set using metrics like accuracy, precision, recall, F1-score, and mean absolute error (MAE).

Tool Implementation:

The implementation uses Python, with libraries such as pandas for data handling, scikit-learn for ML (including CountVectorizer and Naive Bayes), and joblib for model persistence. The system is designed as a command-line interface (CLI) tool for user interaction, supporting both detection and generation of phishing/legitimate emails.

TOOL IMPLEMENTATION

System Architecture

Data Collection Module:

Gathers emails and associated metadata from various sources, including user reports and open datasets.

Feature Extraction Module:

Processes raw email data to extract features using NLP and URL analysis techniques.

AI-Powered Detection Engine:

Utilizes pre-trained ML models (e.g., Naive Bayes, MLP) to classify emails as phishing or legitimate. Ensemble methods or deep learning models can be used to improve accuracy.

Real-Time Monitoring and Alerting:Continuously scans incoming emails and flags suspicious messages. Can be integrated with email clients for real-time alerts.

Response and Prevention Module:Automatically blocks or quarantines detected phishing emails and can generate simulated phishing emails for training purposes.

Programming Language: Python

Libraries: pandas, scikit-learn, joblib, random, os

Data Files: phishing\_emails.csv (training data)

model.pkl (saved trained model)

RESULT & OBSERVATION

Model Performance

Gradient Boosting Classifier:

Precision: 96.2%

F1-score: 96.6%

Recall: 99.9%

Mean Absolute Error: 0.002

This model demonstrated the best overall performance, balancing high precision and recall, which is crucial for minimizing false positives and maximizing true positive detections.

Multi-Layer Perceptron (MLP):

Accuracy (test): 96.4%

F1-score: 98.9%

Recall: 97.3%

MAE (test): 0.071

The MLP model showed excellent generalization and low error rates.

Naive Bayes and Decision Tree:

These models had lower accuracy and F1-scores compared to Gradient Boosting and MLP, but Naive Bayes is often preferred for its simplicity and speed in text-based applications

Ethical Impact and Market Relevance

Ethical Impact: The rise of AI-driven email generation raises significant ethical concerns, especially in the realm of phishing and social engineering. AI tools can be used to create highly sophisticated phishing emails that can trick even well-trained individuals. These tools can exploit psychological manipulation, making it more difficult for people to recognize malicious intent. Ethical challenges arise regarding the responsible development and deployment of AI, especially as the line between authentic and malicious emails becomes blurred.

Mitigating misuse: Ethical AI development in email generation involves ensuring safeguards against the misuse of AI in generating malicious content. AI must be regulated to prevent malicious actors from using these tools for phishing or other harmful activities.

Phishing Detection: On the flip side, AI's role in phishing detection is largely positive. Machine learning models can be trained to detect phishing attempts based on email structure, language, and known tactics. However, concerns over privacy and the handling of sensitive user data may arise when deploying AI-powered security measures.

Market Relevance: As the number of phishing attacks continues to rise globally, AI-based phishing detection systems are highly relevant in the cybersecurity market. The demand for automated systems to protect individuals and organizations from these threats is growing. Companies and governments are increasingly investing in AI to protect sensitive data, making AI-driven phishing detection a key component of modern security infrastructures.

Future Scope

Advances in AI and Machine Learning: As AI technology evolves, phishing detection tools will become more sophisticated, capable of analyzing large volumes of email traffic with greater accuracy. Future developments will focus on making these tools more adaptive, learning new phishing tactics in real-time to stay ahead of attackers.

AI in Personalized Phishing Detection: There's potential for AI to personalize phishing detection, tailoring security measures to individual users' habits, preferences, and behaviors. This could improve phishing detection accuracy by understanding each user’s unique communication style.

Deep Learning and Natural Language Processing (NLP): The combination of deep learning and NLP will further enhance phishing detection systems by enabling them to understand subtle language cues, context, and tone that human analysts might miss. This could make AI even better at identifying novel phishing tactics.

AI-Driven Email Generation for Security Awareness: On the flip side, AI could also be used to create realistic simulated phishing emails for training employees in recognizing phishing attempts, enhancing organizational cybersecurity awareness.

REFERENCE

1.”The Role of AI in Detecting Phishing Attacks" - S. R. Mohapatra, M. R. Rajasekaran, and S. D. Mitra. IEEE Transactions on Dependable and Secure Computing.

2.”Artificial Intelligence in Cybersecurity: An Overview" - L. K. Teo and F. Yang. Journal of Cybersecurity Research and Applications.

3."Detecting Phishing Emails with Machine Learning Algorithms" - P. K. S. Rajendran, P. A. C. C. J. Fernando. Journal of Cyber Security Technology.

4."Phishing and Social Engineering Attacks in Email: A Study on Detection Techniques" - H. Kumar and A. Gupta. International Journal of Computer Applications.

5. [us.norton.com › blog › online-scams20 types of phishing attacks + phishing examples - Norton](https://us.norton.com/blog/online-scams/types-of-phishing)

6. [www.fbi.gov › how-we-can-help-you › scams-and-safetySpoofing and Phishing — FBI](https://www.fbi.gov/how-we-can-help-you/scams-and-safety/common-frauds-and-scams/spoofing-and-phishing)

7. [www.ftc.gov › topics › identity-theftPhishing Scams - Federal Trade Commission](https://www.ftc.gov/news-events/topics/identity-theft/phishing-scams)

8. [ash.harvard.edu › articles › weaponized-ai-a-new-eraWeaponized AI: A New Era of Threats and How We Can Counter It](https://ash.harvard.edu/articles/weaponized-ai-a-new-era-of-threats/)

9. [www.hoganlovells.com › en › publicationsConfronting social engineering in the age of artificial ...](https://www.hoganlovells.com/en/publications/confronting-social-engineering-in-the-age-of-artificial-intelligence)

10. [openai.com › chatgpt › overviewChatGPT | OpenAI](https://openai.com/chatgpt/overview/)

END