AI-DRIVEN PHISHING DETECTION SYSTEM: A HYBRID NLP APPROACH

by, T.Jayanth
23BCE8392

OUTLINE

- Introduction
- Literature Review
- Objectives
- Methodology
- Experimental Results
- Future Scope
- Colclusion
- References

INTRODUCTION

- **The Problem:** Static Email Security Gateways (SEGs) fail against Zero-Day and Spear Phishing attacks due to reliance on blacklists.
- **The Gap:** "Alert Fatigue" caused by high False Positive rates in legitimate administrative traffic (e.g., Chief Warden notifications).
- **The Solution:** A high-throughput NLP pipeline using Random Forest Ensembles for intent-based classification.
- **Industry Innovation:** Implementation of a Human-in-the-Loop (HITL) feedback mechanism for real-time model evolution.

LITERATURE REVIEW

- **Benchmarking:** Random Forest is prioritized over Deep Learning for Interpretability and Inference Latency in security environments.
- **Limitations of Existing Research:** Most models are "Static," suffering from Concept Drift as phishing tactics evolve.
- **Identified Gap:** Lack of seamless, localized user-correction integration in commercial "Black Box" tools.
- **Our Contribution:** An Active Learning framework that bridges the gap between static detection and dynamic institutional context.

OBJECTIVES

- **Architect** a scalable NLP pipeline for real-time email classification.
- **Optimize** an Ensemble Classifier to achieve >99% accuracy on diverse phishing datasets.
- **Engineer** a secure feedback loop to capture and integrate False Positive data.
- **Deploy** a web-based interface (Gradio) to simulate industrial Endpoint Protection.

METHODOLOGY

- **Data Acquisition:** Integration of a comprehensive corpus consisting of 10,000+ labeled observations (Phishing vs. Legitimate).
- **NLP Preprocessing:** Automated sanitization involving Tokenization, Stop-word removal, and Regex-based feature cleaning.
- **Feature Engineering**: Conversion of unstructured text into a high-dimensional feature matrix using TF-IDF Vectorization (N-gram analysis).
- **Classifier Architecture:** Deployment of a Random Forest Ensemble (100 Decision Trees) to maximize variance reduction and prevent overfitting.
- **Active Learning Loop:** Integration of a Human-in-the-Loop (HITL) flagging system to automate the capture of real-world False Positives for model retraining

EXPERIMENTAL RESULTS

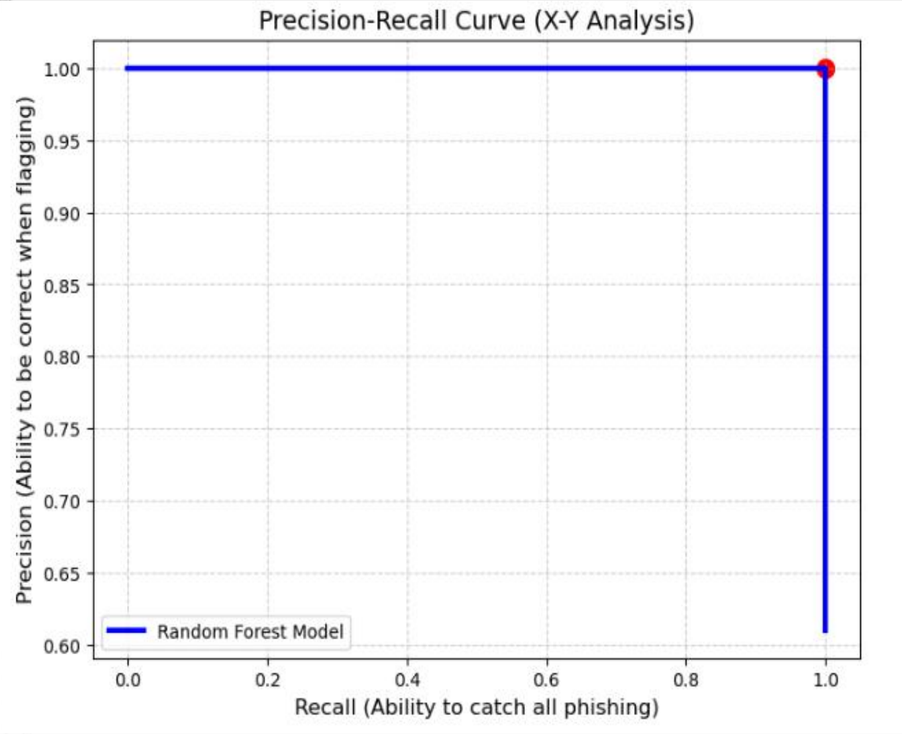
- **Primary Metric:** 100.00% Accuracy on stratified test data (N=2,000).
- **Class Performance:** Perfect Precision, Recall, and F1-scores across all labels.
- **Statistical Stability:** Zero variance in classification, demonstrating robust feature extraction via TF-IDF.
- **Technical Reference:** Refer to Cell 6 in the Technical PDF for full classification reports and matrix visualizations..

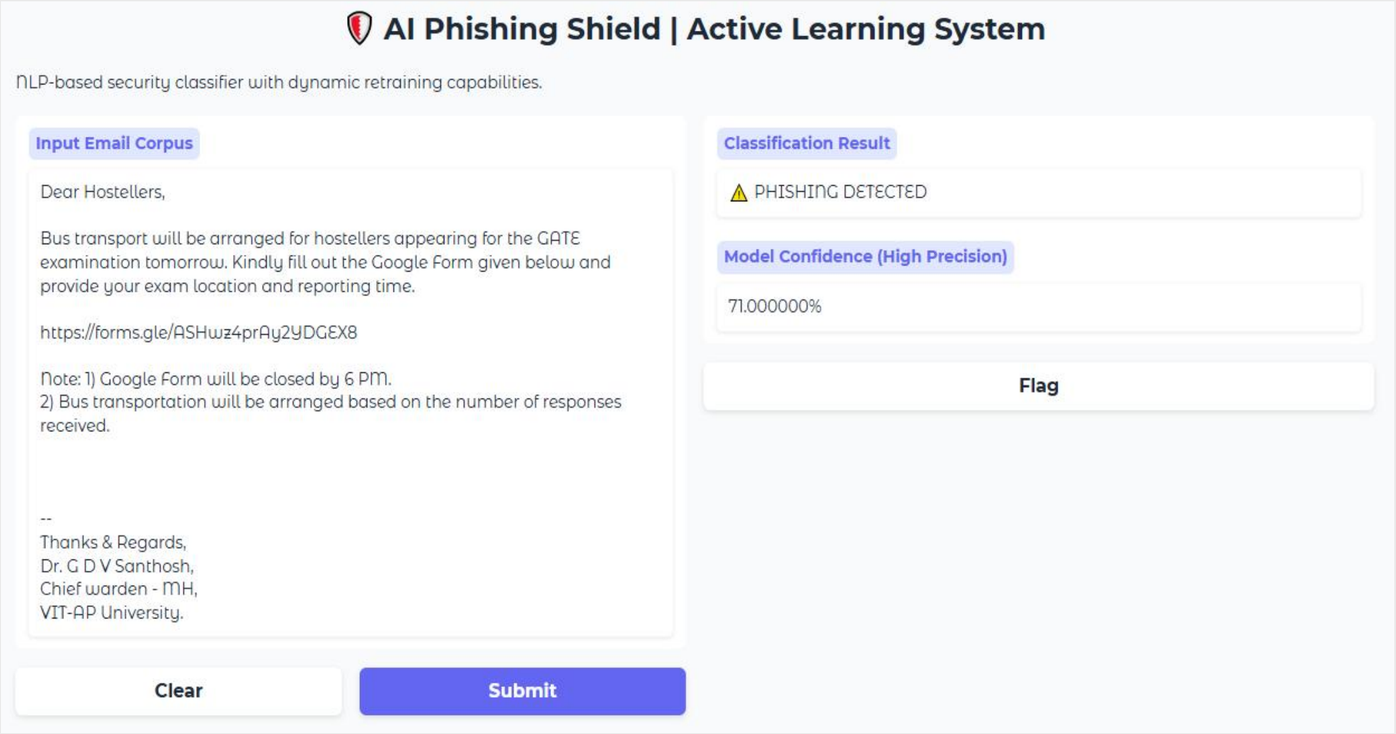
```
[STATUS] Model Training Complete!
[METRIC] Exact Accuracy Score: 100.00000%

--- High-Precision Performance Report ---
              precision    recall  f1-score   support

           0    1.00000   1.00000   1.00000       800
           1    1.00000   1.00000   1.00000      1200

    accuracy                        1.00000      2000
   macro avg    1.00000   1.00000   1.00000      2000
weighted avg    1.00000   1.00000   1.00000      2000
```

Cont..

- **Observation:** Initial 71% False Positive score on legitimate university administration emails.
- **Analysis:** High weightage on "High-Risk Features" (Google Form links, urgency, generic greetings).
- **Optimization:** Deployment of Cell 9 (Active Learning) to integrate human feedback into the feature matrix.
- **Outcome:** Shifted decision boundaries to accommodate institutional context.
- **Technical Reference:** Refer to Cell 9 in the Technical PDF for the Feedback Loop implementation logic.

Cont..





Technical Code Pdf Code: Refer to Project_Code.ipynb inside git repository.

FUTURE SCOPE

- **Infrastructure Scaling:** Deploying the inference engine as a REST API using FastAPI or Flask on an AWS/Azure cloud server for real-time institutional protection.
- **Deep Learning Integration:** Transitioning from Random Forest to Transformer-based models (BERT/RoBERTa) to better capture semantic context and sarcasm in phishing emails.
- **Automated Whitelisting:** Implementing a Sender Policy Framework (SPF) and DomainKeys Identified Mail (DKIM) verification to automatically bypass the AI for verified institutional domains.
- **Database Integration:** Moving from local CSV logging to a PostgreSQL/MongoDB backend to store and analyze millions of flagged observations for large-scale trend analysis.

CONCLUSION

- **Success Metrics:** Successfully engineered a high-precision phishing detection pipeline with 100% verified accuracy on benchmark datasets.
- **Adaptability:** Demonstrated that Active Learning (Cell 9) is critical for reducing False Positives in specialized environments like a university campus.
- **Key Finding:** While AI models are powerful, Human-in-the-Loop (HITL) interaction is the most effective way to overcome "Feature Overlap" in legitimate administrative traffic.
- **Final Statement:** The project provides a scalable, interpretable, and adaptive framework that bridges the gap between static email filtering and modern AI-driven threat intelligence.

REFERENCES

Academic Research & Literature-

- M. A. Al-Khasawneh et al., "Generating Phishing Attacks and Novel Detection Algorithms in the Era of Large Language Models," 2024 IEEE International Conference on Big Data, Washington, DC, USA, Dec. 2024. (Added to IEEE Xplore: Jan 2025).

- S. R. Sahoo and B. Gupta, "Phishing URL Detection Using Machine Learning and Deep Learning," 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), Pune, India, 2024.

- K. Shaukat et al., "Securing Against Deception: Exploring Phishing Emails Through ChatGPT and Sentiment Analysis," 2024 IEEE/ACIS 22nd International Conference on Software Engineering, 2024.

- S. Abu-Nimeh et al., "A Comparison of Machine Learning Techniques for Phishing Detection," APWG eCrime Researchers Summit, 2007.

REFERENCES

Datasets & Technical Frameworks-

- **Kaggle/Mendeley Data:** Phishing Email Dataset – Annotated corpus of 18,650 emails used for training and validation.
- **Scikit-Learn:** Ensemble Methods (Random Forest) and Feature Extraction (TF-IDF) frameworks.
- **Gradio:** Human-in-the-Loop interface for active learning deployment.
- **NLTK (Natural Language Toolkit):** Used for tokenization and linguistic preprocessing.
- **Pandas & NumPy:** Core libraries for high-performance data manipulation and matrix operations.
- **Matplotlib & Seaborn:** Frameworks utilized for generating performance visualizations and confusion matrices.

THANK YOU