Detecting Phishing Emails Leading To Ransomware Using A Lightweight Transformer Model

Tạ Duy Huy
Trường Đại học Công Nghệ Thông Tin

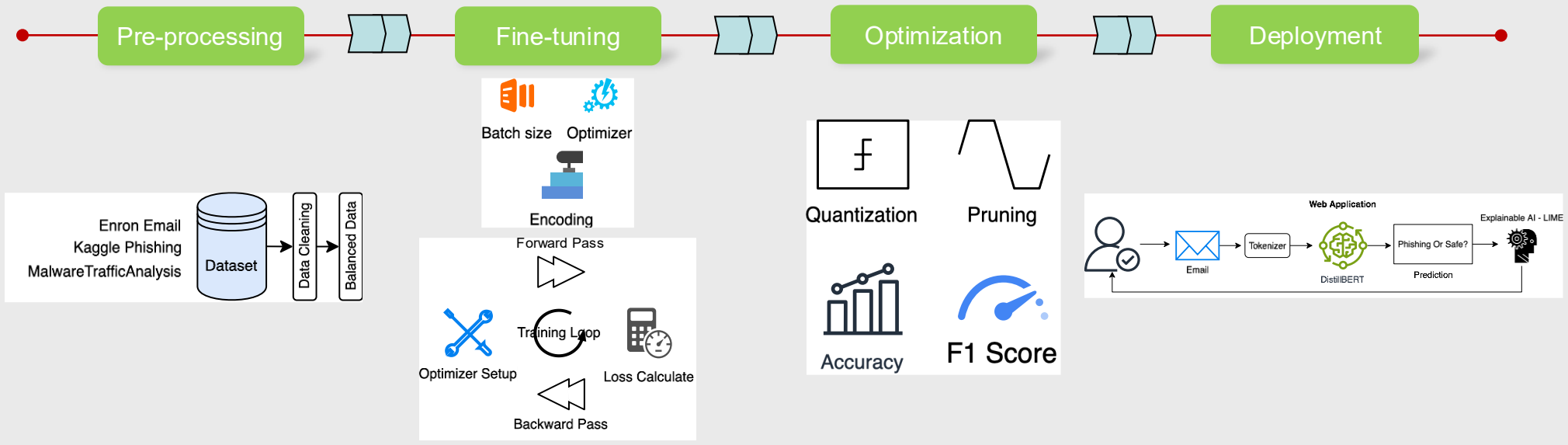
What ?

- A lightweight Transformer-based (DistillBERT) model integrated with Explainable AI (LIME) for detecting phishing emails leading to ransomware.
- Preprocesses email data, classifies emails as "phishing" or "legitimate" with high accuracy (F1-score > 95%), and uses LIME to highlight suspicious keywords (e.g., "urgent").
- Designed for deployment on low-resource devices such as laptops and Raspberry Pi, with a model size under 100MB and inference time less than 1 second.

Why ?

- Phishing emails drive ransomware, causing $18.7 billion losses in 2023, targeting vulnerable users.
- Traditional methods (SVM, Random Forest) lack context analysis; BERT requires heavy resources.
- Growing ransomware threats (e.g., WannaCry, Colonial Pipeline) demand scalable, transparent solutions.
- Lack of user-friendly tools leaves non-experts exposed to sophisticated cyber threats.

Overview



Description

1. Pre-processing (Data Collection & Pre-processing)

- Data Collection: Gathers diverse email datasets, including Enron Email Dataset (500,000 legitimate emails), Kaggle Phishing Dataset (10,000 phishing emails), and MalwareTrafficAnalysis (1,000 ransomware-related emails from campaigns like Ryuk and WannaCry).
- Pre-processing: Removes duplicates, HTML tags, and special characters (e.g., emojis) using BeautifulSoup; tokenizes emails with BERT tokenizer (max 512 tokens); addresses class imbalance using class weighting to enhance model performance on minority classes.

2. Training (Fine-tuning)

- Fine-tunes DistilBERT (66 million parameters) on Google Colab with a Tesla T4 GPU.
- Training parameters: batch size of 16, learning rate of 2e-5, maximum 5 epochs, using Cross-Entropy loss and AdamW optimizer.
- Dataset split: 80% training, 10% validation, 10% testing, targeting an F1-score above 95%.

3. Optimization

- Applies quantization (float32 to int8) via Hugging Face Optimum, reducing model size to under 100MB and RAM usage below 8GB.
- Uses pruning to remove 20% of less significant weights, achieving inference time below 1 second per email.
- Converts the model to ONNX format with ONNX Runtime for optimized processing on CPU and ARM architectures.

4. Deployment

- Deploys the model through a Flask-based web application, allowing users to input emails and receive real-time classification results.
- Tests on low-resource devices: laptops (Windows/Linux, 8GB RAM) and Raspberry Pi 4 (4GB RAM).
- Integrates LIME for explainability, displaying key phishing terms (e.g., "urgent": 0.75) on the interface, with a user survey targeting > 4/5 understandability score.

Tạ Duy Huy – Trường Đại học Công Nghệ Thông Tin
TEL : +84916207579 Email : huytd.19@grad.uit.edu.vn