Real-Time NLP for Cybercrime Reporting

Problem Statement

Development of an NLP Model to guide citizens in filing cybercrime reports on the National Cyber Crime Reporting Portal (NCRP) correctly through a real time analysis of the description and incident supporting media files uploaded by the citizen.

Introduction

In a time when cyber threats are increasingly complex and widespread, it is vital for global organizations to effectively identify and manage cybersecurity incidents. This project contributes to addressing this necessity by creating an advanced classification system based on neural networks that automatically assigns categories to cyber incidents using text-based inputs. Utilizing a hierarchical method, the system discerns between general and detailed threat categories, enhancing its utility for security operations centers and response teams.

This initiative incorporates cutting-edge natural language processing (NLP) techniques and deep learning frameworks to analyze and sort data from cyber incident reports. With a substantial dataset containing more than 31,000 incidents, the model is designed to optimize the triage process and enhance the speed of responses by automating the initial sorting of incidents. This document outlines the results, measurement metrics, and further enhancement suggestions related to the project.

Significant Findings from NLP Analysis

The analysis of cybersecurity incidents using natural language processing revealed several critical insights about the nature and distribution of cyber threats:

Classification Performance

- The model achieved a primary category accuracy of 74.15% and subcategory accuracy of 53%, demonstrating stronger performance in broad categorization compared to specific incident types.

- The confidence scores show higher certainty in category predictions (mean: 0.76, std: 0.22) compared to subcategory predictions (mean: 0.56, std: 0.24).

Distribution and Patterns

1. **Dominant Incident Types**:
    - Online Financial Fraud emerged as the most prevalent category, with particularly strong model performance (F1-score: 0.86)
    - UPI-related frauds showed high recall (90.38%), indicating effective detection of these common incidents
2. **Challenging Categories**:
    - The model struggled with rare incident types such as:
        - Cyber Terrorism (F1-score: 0.0)
        - Ransomware (F1-score: 0.0)
        - Online Gambling & Betting (F1-score: 0.0)
    - This suggests a clear class imbalance issue in the training data
3. **Text Processing Insights**:
    - The implementation of domain-specific preprocessing, including handling of:
        - IP addresses
        - Email addresses
        - URLs
        - Technical terminology
    - Retention of security-relevant stop words improved context preservation
4. **Hierarchical Classification**:
    - The dual-output architecture (categories and subcategories) showed varying effectiveness
    - Better performance in broad categorization suggests successful capture of general incident patterns
    - Lower subcategory accuracy indicates challenges in fine-grained classification

Model Evaluation

Category-Level Metrics:

1. **Overall Performance**:

- Accuracy: 74.15%
- Weighted Average F1-score: 0.707
- Macro Average F1-score: 0.335

2. **Strong Performance Areas**:
    - Cyber Attack/Dependent Crimes (F1: 0.997)
    - Rape/Gang Rape/Sexually Abusive Content (F1: 0.950)
    - Online Financial Fraud (F1: 0.862)
3. **Areas Needing Improvement**:
    - Cyber Terrorism (F1: 0.0)
    - Online Gambling & Betting (F1: 0.0)
    - Ransomware (F1: 0.0)

Subcategory-Level Metrics:

1. **Overall Performance**:
    - Accuracy: 53%
    - Weighted Average F1-score: 0.486
    - Micro Average F1-score: 0.530
2. **Notable Results**:
    - UPI Related Frauds (F1: 0.682)
    - Debit/Credit Card Fraud/Sim Swap Fraud (F1: 0.692)
    - Cryptocurrency Fraud (F1: 0.563)

Implementation Plan

Short-term Improvements (1-3 months):

1. **Data Balancing**:
    - Implement class-weight adjustments
    - Apply SMOTE or similar techniques for minority classes
    - Collect additional data for underrepresented categories
2. **Model Architecture Enhancements**:
    - Experiment with transformer-based models (BERT, RoBERTa)
    - Implement attention mechanisms for better feature extraction
    - Add regularization techniques to prevent overfitting

3. **Preprocessing Refinements**:
    - Enhance domain-specific token handling
    - Implement advanced text cleaning for cybersecurity terminology
    - Create custom embeddings for technical terms

Long-term Deployment Strategy (3-6 months):

1. **System Integration**:
    - Develop REST API for model serving
    - Implement batch prediction capabilities
    - Create monitoring dashboard for model performance
2. **Maintenance Plan**:
    - Set up automated retraining pipeline
    - Implement drift detection
    - Establish feedback loop for continuous improvement

Conclusion

The cyber incident classification system has shown promising outcomes in automating the essential task of categorizing incidents, achieving an accuracy of 74.15% in primary classification. The hierarchical method is effective for broad categorizations, though it identifies areas that require enhancement in the classification of subcategories. The system is particularly adept at detecting prevalent cyber threats, notably within the domain of financial fraud, and it maintains high precision in crucial security scenarios.

However, the evaluation also uncovers significant challenges, especially in processing uncommon incident types and achieving uniform performance across various subcategories. The training data's class imbalance highlights an opportunity for enhancement through focused data gathering and sophisticated sampling methods. The deployment strategy detailed in this report suggests a mix of immediate technical enhancements and broader strategic initiatives.

Looking forward, the system's modular design and robust NLP foundation set the stage for future upgrades. With the suggested improvements and ongoing refinements, the classification system is poised to substantially improve the capabilities of cybersecurity incident responses, facilitating quicker, more precise threat identification and coordinated responses.

References and Dependencies

Primary Libraries:

- TensorFlow 2.x
- Keras
- NumPy
- Pandas
- Scikit-learn
- NLTK

Custom Components:

- CyberIncidentClassifier class
- Custom text preprocessing pipeline
- Hierarchical classification architecture

Academic References:

1. Vaswani et al. (2017 revised in 2023). "Attention Is All You Need"
2. Devlin et al. (2018 revised in 2019). "BERT: Pre-training of Deep Bidirectional Transformers"

Plagiarism Declaration

I hereby declare that this report is my original work, and all sources used have been properly cited. The implementation uses standard libraries and follows ethical coding practices.

Author: Dhruvkumar Patel, Vatsal Lavari

Date: November 7, 2024