INDIAAI CYBERGUARD AI HACKATHON

NLP-based Cybercrime Classification and Citizen Assistance

Team Name: *AI-Asaatid*

Team Members
Owais Ahmad Lone (*IIT Kharagpur*)
Kaisar Imtiyaz (*IIT Kharagpur*)

Submission Date: November 22, 2024

Abstract

The exponential rise in cybercrime cases has created a pressing need for scalable and efficient solutions to process, classify, and act upon textual descriptions of such incidents. This report outlines two distinct approaches leveraging cutting-edge NLP models: **BERT-base-uncased** and **DistilBERT**. Each approach is thoroughly evaluated for its technical merit, with results demonstrating the efficacy and limitations of these methods. Furthermore, this document details the conceptualization of an *end-to-end citizen assistance pipeline*, integrating these models into an interactive system for accurate cybercrime reporting. The report also explores scalability, future enhancements, and potential use cases, making this a comprehensive blueprint for a state-of-the-art AI solution in cybercrime prevention.

1 Introduction

Cybercrime: A threat that grows more complex with every passing day. Statistics from the NCRP database reveal that India faces approximately 6,000 reported cybercrime cases daily. These cases span a spectrum of crimes such as financial fraud, phishing, and identity theft, underscoring the urgent need for robust AI solutions.

Objective and Significance

The primary goal is to develop an NLP model capable of:

- Categorizing cybercrime reports into **specific categories and subcategories**.
- Assisting citizens in filing accurate cybercrime reports through **real-time feedback**.
- Enhancing law enforcement capabilities by providing **structured and actionable data**.

Key Contributions of This Report:

- A two-pronged approach using BERT and DistilBERT for text classification.
- Conceptualization of a scalable pipeline for citizen assistance.
- Extensive discussion on scalability, future advancements, and use cases.

2 Dataset Description

2.1 Overview

The datasets provided by the National Cybercrime Reporting Portal (NCRP) contain textual descriptions of various cybercrime incidents. These datasets are structured to include both categorical labels and textual data fields for classification tasks.

2.2 Fields and Features

The dataset consists of the following key fields:

- **crimeaditionalinfo:** The primary textual description of the incident, often including detailed victim narratives and other contextual information.
- **category:** The broader classification of the crime, e.g., *Online Financial Fraud*, *Identity Theft*.
- **subcategory:** A more specific classification within the broader category, e.g., *Fraud CallVishing*, *UPI Related Frauds* are the sub categories of the category *Online Financial Fraud*.

2.3 Class Distribution

The dataset includes a variety of classes with varying levels of representation. The most frequent class is **Online Financial Fraud**, while rare classes include **Report Unlawful Content** and **Crime Against Women and Children**. These were excluded during preprocessing to ensure balanced training.

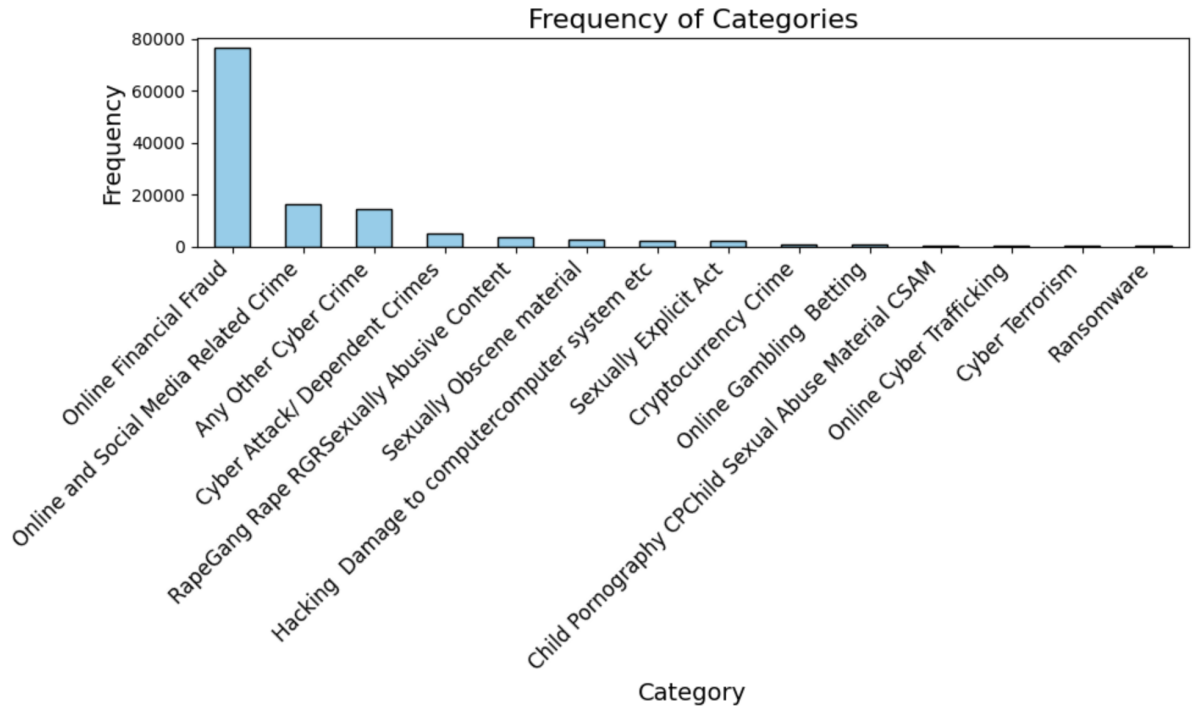

Figure 1: Category distribution after removing extremely rare categories

2.4 Challenges Identified in the Dataset

- **Imbalanced Classes:** Certain categories dominate, making it challenging to train unbiased models.
- **Text Quality:** Variability in text length, grammar, and structure due to citizen-generated content.
- **Rare Subcategories:** Subcategories with limited samples lead to challenges in generalization.

2.5 Insights for Model Design

The dataset's structure and challenges necessitated:

- Using robust preprocessing techniques to normalize text quality.
- Addressing class imbalance through undersampling methods. Over sampling in text based data is not useful.
- Selecting models capable of handling sparse subcategories effectively.

3 Data Preprocessing

3.1 Dataset Overview

The datasets provided by NCRP consist of textual descriptions of cybercrimes, categorized by the main *category* and *sub-category*. However, the data posed challenges such as:

- **Class Imbalance:** Certain categories were underrepresented, leading to potential bias.
- **Irregular Text Quality:** Variations in language and grammar required normalization.

3.2 Preprocessing Steps

Step 1: Dataset Consolidation and Cleaning

- Combined datasets into a unified pandas dataframe.
- Dropped rare classes like **'Report Unlawful Content'** and **'Crime Against Women and Children'** to avoid overfitting.
- Encoded 'category' and 'subcategory' columns as integers for compatibility with neural models.

Step 2: Text Normalization using NLTK and SPACY

- Applied tokenization using BERT and DistilBERT tokenizers.
- Implemented *lemmatization* with `spaCy` to ensure words were reduced to their base forms, e.g., *'running'* $\rightarrow$ *'run'*.
- Removed stop words (e.g., *'the', 'and'*) using `NLTK` to retain only impactful terms.
- Balanced the dataset by limiting the dominant class (**Online Financial Fraud**) to a threshold of 20,000 samples.

3.3 Impact of Preprocessing

The preprocessing steps improved the data's uniformity and ensured a fair distribution of categories during training, addressing **key issues such as data imbalance and noise**.

4 Approach 1: BERT-based Classification

4.1 Model Overview

BERT, or Bidirectional Encoder Representations from Transformers, is renowned for its contextual understanding of text. By analyzing both the left and right context of words, BERT captures nuanced relationships within text data.

4.2 Technical Implementation

- **Model Architecture:**
  - Used `bert-base-uncased`.
  - Added two dense layers for category and subcategory classification.
- **Loss Function:** Multi-task loss, combining cross-entropy losses for both outputs.
- **Training:**
  - Frozen pretrained layers to preserve general language understanding.
  - Fine-tuned classification heads over 3 epochs with a learning rate of $2e^{-4}$.

4.3 Results

Performance Metrics:

- Category classification accuracy: **74%**.
- Subcategory classification accuracy: **51%**.

Key Challenges:

- Insufficient representation of subcategories limited the model's effectiveness.
- Computational cost of BERT, especially during fine-tuning.

5 Approach 2: DistilBERT-based Sequence Classification Task

5.1 Model Overview

DistilBERT is a lightweight transformer model distilled from BERT. It retains 97% of BERT's accuracy while being 40% smaller and 60% faster.

5.2 Technical Implementation

- **Model Architecture:**
  - Used `distilbert-base-uncased`.
  - Added a single dense layer for multi-label classification (category + subcategory).
- **Loss Function:** Multi-label classification loss.
- **Training:**
  - Fine-tuned the entire model for 3 epochs.
  - Optimized using AdamW optimizer with a learning rate of $2e^{-4}$.

5.3 Results

Performance Metrics:

- Category classification accuracy: **80%**.
- Subcategory classification accuracy: **60%**.

Advantages Over BERT:

- Faster training and inference.
- Better generalization due to enhanced training conditions.

6 Future: End-to-End Pipeline for Citizen Assistance

6.1 Pipeline Overview

The proposed system integrates advanced NLP models, enhanced by Retrieval-Augmented Generation (RAG), into a comprehensive end-to-end pipeline designed to assist users at every stage of cybercrime reporting and prevention. By leveraging the capabilities of Large Language Models (LLMs), the pipeline empowers users to *identify*, *act on*, and *prevent* cyber threats in real-time, offering a seamless and intuitive experience in the modern digital landscape.

This pipeline combines state-of-the-art classification models with a dynamic, intelligent chatbot interface to create an innovative solution capable of guiding users through the complexities of cybersecurity.

Stages of the Pipeline:

1. **IDENTIFY: User-Level Security Through RAG-Based LLM Integration** At the first stage, the system employs a RAG-based chatbot to help users detect and evaluate potential threats. The chatbot:

   - Screens suspicious messages, emails, or notifications in real-time.
   - Classifies content as genuine or fraudulent using pre-trained cybercrime detection models.
   - Proactively flags phishing attempts, fraudulent transaction requests, or malicious links, enabling users to avoid falling victim to scams.

   For example, the chatbot can analyze an uploaded email screenshot and respond with a detailed assessment of whether it is a phishing attempt, including references to known fraud patterns.

2. **ACT ON: Guided Reporting and Assistance** Once a user identifies a cybercrime incident, the pipeline transitions to provide tailored reporting assistance. Key functionalities include:

   - **Dynamic Report Generation:** The chatbot extracts key details from the user's input and pre-fills forms on the NCRP portal.
   - **Real-Time Feedback:** Provides suggestions to refine the report for clarity and accuracy (e.g., prompting for missing details like transaction IDs or specific dates).
   - **Interactive Walkthrough:** Offers step-by-step guidance for completing the reporting process, minimizing errors and confusion.

   Additionally, the chatbot ensures that users understand the next steps, such as tracking the status of their complaint or contacting the appropriate authorities.

3. **PREVENT: Proactive Threat Mitigation and Education** Beyond reactive measures, the pipeline incorporates proactive features to reduce the likelihood of future cybercrime incidents:

   - **Dynamic Alerts:** Sends personalized warnings about emerging cyber threats based on trends observed in the user's geographical area or behavior.
   - **Cybersecurity Education:** Provides interactive tutorials, quizzes, and tips tailored to the user's needs, helping them stay vigilant against common scams.

- **Behavioral Insights:** Analyzes user activity to identify potential vulnerabilities and offers customized recommendations, such as strengthening passwords or avoiding risky websites.

6.2 Integration of RAG-Based Chatbot

The core of this pipeline is an advanced RAG-based chatbot that combines retrieval and generation for precise, context-aware assistance:

- **Retriever Module:** Uses a vectorized knowledge base, dynamically updated with the latest cybercrime advisories, FAQs, and government regulations.
- **Generator Module:** Employs fine-tuned LLMs to provide human-readable, actionable responses grounded in retrieved information.
- **Feedback Loop:** Continuously improves retrieval and generation through user feedback, ensuring accuracy and relevance.

This chatbot operates seamlessly across all pipeline stages, making it the central intelligence that connects users, law enforcement, and the broader cybersecurity ecosystem.

6.3 Integration with Law Enforcement and Scalability

To ensure end-to-end efficiency, the pipeline facilitates robust integration with law enforcement and scalability for nationwide deployment:

- **Law Enforcement Dashboard:** Provides structured, categorized reports with metadata for streamlined case prioritization and investigation.
- **Scalable Design:** Cloud-based deployment ensures responsiveness during peak reporting periods, while hybrid retrieval mechanisms handle diverse user queries effectively.
- **API Compatibility:** Enables seamless integration with existing cybercrime monitoring systems, fostering collaboration across agencies.

7 Conclusion

The integration of advanced NLP and RAG-based models into a comprehensive end-to-end pipeline offers a groundbreaking approach to cybercrime reporting, prevention, and user empowerment. By combining robust classification techniques with a dynamic RAG-based chatbot, the proposed solution provides real-time, actionable assistance to users while addressing key challenges in cybe rsecurity.

The system not only helps users *identify* and *act on* potential threats but also equips them to *prevent* future incidents through proactive alerts, personalized education, and behavior-driven recommendations. The chatbot's ability to dynamically retrieve and generate accurate, context-aware responses ensures precision and reliability at every stage of user interaction.

Furthermore, seamless integration with law enforcement systems accelerates case processing by delivering structured and actionable reports. The scalable and modular design of the

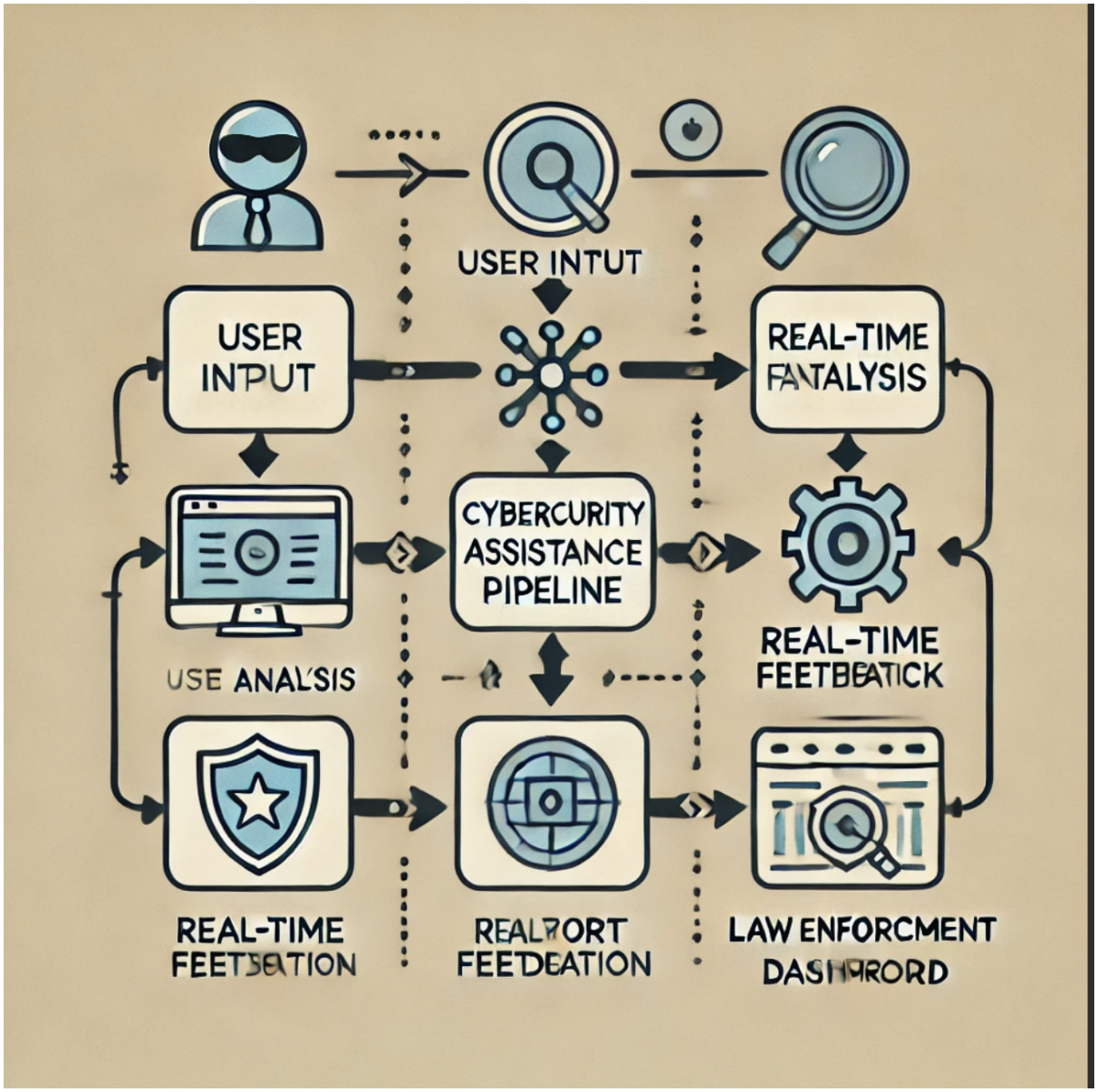

Figure 2: Simplified Workflow for the Cybercrime Assistance Pipeline.

pipeline ensures adaptability to evolving cyber threats and increasing user demands, making it a future-ready solution for modern digital ecosystems.

This project represents a significant leap in leveraging AI to address cybercrime, fostering a safer online environment and empowering individuals with the tools and knowledge needed to navigate the complexities of cybersecurity effectively.

References

1. Devlin, J., et al. ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.”

2. Sanh, V., et al. ”DistilBERT, a distilled version of BERT: smaller, faster, cheaper, and lighter.”

3. spaCy Documentation: https://spacy.io/