AI SAFTEY POC TECHNICAL REPORT

CONTENTS:

1.INTRODUCTION:

The aim of this project is to develop a Proof of Concept (POC) for AI-driven safety models in conversational platforms. The goal is to identify harmful content, monitor conversation escalation, and detect signs of emotional distress in real-time. This POC demonstrates end-to-end functionality, from data preprocessing to model inference, and can be extended for production use.

The system focuses on four main components:

1. **Abuse Detection** – Identifying harmful or inappropriate language.
2. **Content Filtering** – Blocking age-inappropriate content.
3. **Escalation Detection** – Detecting increasing negativity in conversations.
4. **Crisis Detection** – Identifying potential emotional crises or self-harm indicators.

2.HIGH LEVEL DESIGN DECISIONS:

- **Modular Architecture:** Each model (abuse, content, escalation, crisis) is implemented separately and integrated via a Streamlit web interface.
- **Command-Line / Web Interface:** While the POC includes CLI demos for development, a Streamlit app provides a user-friendly interface for real-time testing.
- **Real-Time Processing:** The system can process individual messages and rolling conversation windows in near-real-time.
- **Python and ML Libraries:** Core libraries used include scikit-learn, Hugging Face Transformers, joblib, pandas, and Streamlit.

3.DATA SOURCES AND PRE-PROCESSING:

- **Abuse Detection:** Uses the publicly available hate_speech_offensive dataset. Text is cleaned and tokenized via TfidfVectorizer.
- **Crisis Detection:** Uses the sentinet/suicidality Hugging Face model for suicidal message detection.
- **Escalation Detection:** Uses a rolling window of offensive scores computed from the abuse detection model.
- **Content Filtering:** Uses predefined categories for age groups (child, teen, adult) to classify messages as safe or unsafe.

Pre-Processing Steps:

- Converting text to lowercase.
- Removing stop words (for abuse detection).
- Converting categorical labels to numerical values for model training.

4.MODEL ARCHITECTURES AND TRAINING:

4.1 Abuse Detection

- **Model:** Logistic Regression with TF-IDF features
- **Training:** Train-test split 80:20, class-weight balanced to handle dataset imbalance
- **Output:** Class label (0=Hate, 1=Offensive, 2=Neither) and probability scores

4.2 Escalation Detection

- **Model:** Uses rolling average of offensive scores from abuse detection
- **Window Size:** 3 messages
- **Alert Threshold:** 0.4 average score triggers escalation warning

4.3 Crisis Detection

- **Model:** Pretrained Hugging Face transformer (sentinet/suicidality)
- **Output:** Label (Suicidal / Non-Suicidal) and confidence score

4.4 Content Filtering

- **Method:** Rule-based filtering based on age categories and content labels
- **Output:** Blocked content types and confidence score

5. EVALUATION RESULTS:

- **Abuse Detection:** Accuracy ~86–89%, F1-score for offensive content ~0.84
- **Escalation Detection:** Qualitative evaluation with rolling conversation examples

- **Crisis Detection:** Uses pretrained model; confidence score indicates severity
- **Content Filtering:** Tested with sample inputs across age groups; successfully blocks inappropriate content.

6. ETHICAL CONSIDERATIONS:

- **Bias Mitigation:** Balanced class weights for logistic regression, ensures fair detection across types of messages
- **Privacy:** Only anonymized datasets used; no sensitive user data is stored
- **Transparency:** Streamlit interface shows model predictions and confidence scores for interpretability.

7.FUTURE IMPROVEMENTS:

The POC demonstrates modular design, allowing multiple team members to work on separate modules simultaneously. Clear interfaces between components make integration easier. Future improvements could include multi-language support, cloud deployment, and automated alerting for crisis detection.

8.CONCLUSION:

This POC provides a working demonstration of AI Safety Models for conversational platforms. It successfully integrates abuse detection, content filtering, escalation monitoring, and crisis detection into a unified, interactive system. The modular approach ensures it can be extended for production with minimal effort.