AI Prompt Injection Detection - Home Assignment

Overview

Your task is to develop a machine learning model that can classify AI prompts as either potentially malicious (PROMPT_INJECTION) or safe (BENIGN).

This assignment will test your ability to handle data, train models, and deploy them in a production-ready manner.

Assignment Details

Task 1: Model Development

1. Use the "prompt_injection_detection" dataset from HuggingFace (https://huggingface.co/datasets/jackhhao/jailbreak-classification)
2. Implement a binary classification model using any of the following approaches:
    - Fine-tuning a pre-trained transformer model
    - Creating a custom neural network
    - Using traditional machine learning approaches

NOTE: You must support all input lengths

Task 2: Training and Evaluation

1. Split the dataset into training, validation, and test sets
2. Implement and document your training pipeline
3. Create evaluation metrics including:
    - Accuracy
    - Precision
    - Recall
    - F1 Score
4. Provide confusion matrix visualization
5. Include examples of correctly and incorrectly classified prompts

Task 3: Deployment

1. Upload your trained model to HuggingFace Hub

2. Create a simple inference pipeline that can:
    - Accept a text prompt as input
    - Return the classification (PROMPT_INJECTION/BENIGN) and confidence score
    - Handle basic error cases

Deliverables

1. A Google Colab notebook containing:
    - Data loading and preprocessing
    - Model implementation and training
    - Evaluation code and results
    - Clear documentation and comments
2. A link to your deployed model on HuggingFace Hub
3. A README.md file explaining:
    - Your approach and design decisions
    - Model architecture and training strategy
    - Key results and observations
    - Instructions for running the inference pipeline

Technical Requirements

- Use Python 3.8+
- Recommended libraries: transformers, torch, pandas, numpy
- Include requirements.txt
- Proper code organization and documentation
- Optional: Git commit history showing incremental progress

Evaluation Criteria

Your submission will be evaluated based on:

1. Code Quality (25%)
    - Clean, well-organized code
    - Proper documentation
    - Error handling
2. Model Performance (25%)
    - Accuracy and other metrics
    - Training efficiency
    - Inference speed
3. Technical Implementation (25%)
    - Proper use of libraries and best practices
    - Data preprocessing and validation
    - Model deployment

4. Documentation (25%)
    - Clear explanation of approach
    - Well-documented notebooks
    - Comprehensive README

Time Expectation

- Expected completion time: 4 hours
- Maximum submission time: 7 days from receipt

Submission Instructions

1. Share the Colab notebook link with edit access
2. Provide the HuggingFace model repository link
3. Optional: Submit all code and documentation via a GitHub repository

Notes

- Focus on creating a working solution first, then improve if time permits
- Comment your code and document your decisions
- Feel free to ask clarifying questions if needed
- Include any assumptions you made during development

Good luck! We're excited to see your solution