AI Project Report: Toxic Comment Detection from Image

1. Introduction

This project leverages OCR and Natural Language Processing (NLP) to extract and classify text from images. The core objective is to detect toxicity levels (toxic, severe_toxic, obscene, threat, insult, identity_hate) in user-generated content captured from images, particularly scanned or uploaded documents. It incorporates Tesseract OCR, grammar correction using a transformer model, and a multi-label classifier trained on labeled comment data.

2. Workflow Overview

Step 1: Data Collection

- **Dataset Used**: `train.csv` (from Jigsaw Toxic Comment Classification Challenge).
- **Columns**: `comment_text` (text), labels: `toxic`, `severe_toxic`, `obscene`, `threat`, `insult`, `identity_hate`.

Step 2: Text Preprocessing

- **From Image**:
  - Convert to grayscale.
  - Apply contrast enhancement and sharpening filter (if blurry).
  - OCR using `pytesseract` to extract text.
- **Cleaning**:
  - Remove non-ASCII characters, special symbols.
  - Normalize whitespace.
- **Grammar Correction**:
  - Using `vennify/t5-base-grammar-correction` transformer model.

Step 3: Feature Extraction

- TF-IDF Vectorizer with max 10,000 features.
- Stopwords removed from English vocabulary.
- Output: Sparse feature matrix for model input.

Step 4: Model Training

- **Algorithm**: One-vs-Rest Linear SVC (Support Vector Classifier).
- **Input**: TF-IDF features.
- **Output**: Multi-label predictions for 6 toxicity categories.
- **Split**: 80% training / 20% validation.

Step 5: Model Evaluation

- Classification report shows precision, recall, F1-score for each class.
- **Accuracy Achieved**: Printed as part of output logs.

3. Application Workflow

- User uploads an image (`.jpg`, `.jpeg`, `.png`).
- Image is preprocessed and passed through OCR.
- Text is grammatically corrected and vectorized.
- Model predicts toxicity levels.
- Labels are displayed with binary outcomes (Yes/No).

4. Technologies Used

- **Python**, **Google Colab**, **Tesseract OCR**, **Transformers**, **Scikit-learn**, **Pandas**, **NumPy**, **OpenCV**.
- **Model Persistence**: Joblib.
- **Model**: SVM (multi-label classification).

5. Results

- High accuracy for toxic category detection.
- Effective correction of low-confidence or blurry text via preprocessing.
- Multi-label classifier capable of recognizing overlapping toxic traits.

Submitted By:

M-Ali irtza (0860)

Saad Naseer (0790)