Toxic Comment Detection from Images Using OCR and Machine Learning

# Objective:

Develop a system that allows users to upload images containing text (such as screenshots of social media comments or chat messages). The system will extract the text using Optical Character Recognition (OCR) and classify it as toxic, offensive, or safe using a machine learning model.

# Problem Statement:

Online platforms are inundated with user-generated content, some of which can be toxic or offensive. Manually moderating this content is challenging and time-consuming. Automating the detection of such content, especially from images, can significantly aid in maintaining a healthy online environment.

# Methodology:

1. Image Upload: Users upload an image containing textual content.  
2. Text Extraction: Utilize OCR (e.g., Tesseract) to extract text from the image.  
3. Preprocessing: Clean and preprocess the extracted text (e.g., lowercasing, removing punctuation).  
4. Classification: Feed the preprocessed text into a trained machine learning model (e.g., Naive Bayes or Neural Network) to classify the content.  
5. Output: Display the classification result to the user, indicating whether the content is toxic, offensive, or safe.

# Dataset:

Name: Jigsaw Toxic Comment Classification Dataset  
Source: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge  
Description: This dataset contains a large number of Wikipedia comments labeled by human raters for toxic behavior. The labels include: Toxic, Severe Toxic, Obscene, Threat, Insult, Identity Hate.

# Tools & Technologies:

- OCR: Tesseract OCR  
- Programming Language: Python  
- Libraries: pytesseract, scikit-learn or TensorFlow/Keras, Flask or Streamlit, Pandas, NumPy, Matplotlib, Seaborn

# Applications:

- Content Moderation: Assist moderators in identifying and filtering toxic content from images.  
- Cyberbullying Prevention: Detect and prevent the spread of offensive content in online communities.  
- Parental Control: Monitor and control the exposure of children to harmful content.

# Evaluation Metrics:

- Accuracy: Measure the overall correctness of the model.  
- Precision: Measure the proportion of true positives among all positive predictions.  
- Recall: Measure the proportion of true positives among all actual positives.  
- F1-Score: Harmonic mean of precision and recall.

# Deliverables:

- Source Code: Complete codebase of the project.  
- Trained Model: Serialized model ready for deployment.  
- Documentation: Detailed report covering: Introduction, Literature Review, Methodology, Experiments and Results, Conclusion  
- Presentation: Slide deck summarizing the project.

### **Idea Presented by:**

1. **Ali irtza 0860**

**Saad Naseer 0790**