LSTM Toxic Content Classification Report

1. Objective

The goal of this task is to train an LSTM-based deep learning model to classify textual queries into multiple toxicity categories.

2. Dataset Description

The dataset consists of user queries combined with image descriptions and labelled into several categories, such as Safe, Violent Crimes, Suicide & Self-Harm, Elections, Sex-Related Crimes, Child Sexual Exploitation, Non-Violent Crimes, Unknown S-Type, and Unsafe.

3. Preprocessing

To improve model learning, several preprocessing steps were applied:

- Combined query and image descriptions into one text field.
- Removed noise words such as: image, photo, shows, description.
- Removed punctuation, numbers, and URLs.
- Tokenized text using Keras Tokenizer.
- Applied padding to ensure equal sequence length.

4. Handling Class Imbalance

The dataset was highly imbalanced. To address this, class weights were computed and applied during training to ensure minority classes were properly learned.

5. Model Architecture

A Bidirectional LSTM architecture was used:

1. Embedding Layer
2. Bidirectional LSTM (128 units)
3. Dropout
4. Bidirectional LSTM (64 units)
5. Dense Layer
6. Softmax Output

6. Training Configuration

Parameter	Value
Epochs	10
Batch Size	64
Optimizer	Adam
Loss Function	Sparse Categorical Crossentropy
Evaluation Metric	F1 Score

7. Evaluation Results

Weighted F1 Score achieved by the model:

0.946

This demonstrates strong classification performance across all categories.

8. Visual Results

Insert the following figures if required by the submission:

- • Training Curve
- • Confusion Matrix

9. Conclusion

The LSTM model successfully captured semantic patterns in the text data. Applying preprocessing and class-weight balancing significantly improved model generalization and performance.