Task 1: LSTM training

This task addresses multi-class toxic text classification. Given a user query and an associated image description, the objective is to predict the corresponding toxicity category. Due to class imbalance and semantic overlap between categories, macro F1 score was used as the primary evaluation metric.

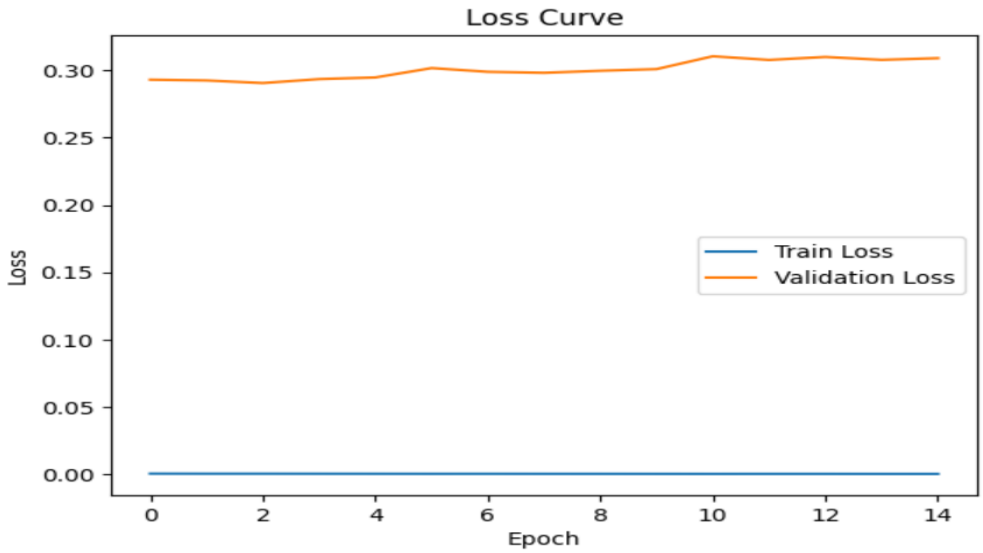
For data preprocessing, duplicate samples were removed to prevent data leakage and inflated performance. Toxicity categories were encoded numerically using label encoding. A stratified train–test split was applied to preserve class distribution across subsets.

To enrich contextual information, the query and image description were concatenated into a single text sequence. The combined text was tokenized and padded to a fixed length to enable batch processing by the neural network.

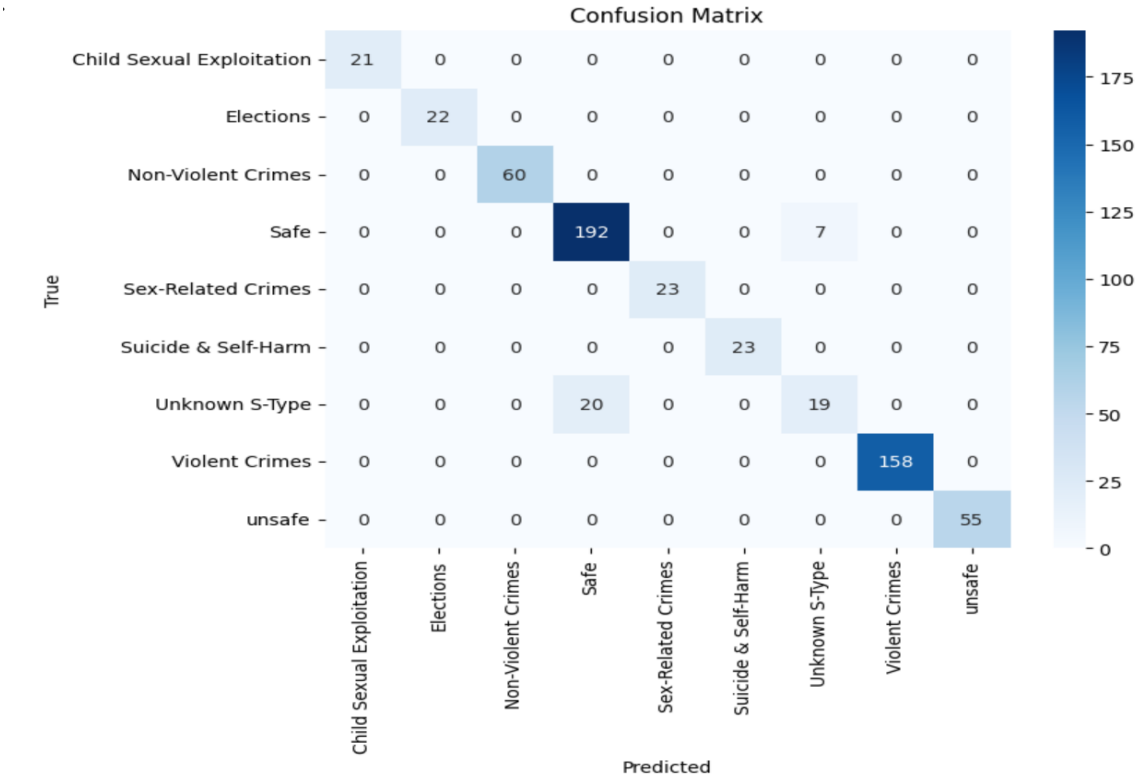
A Bidirectional LSTM was employed to capture contextual dependencies in both forward and backward directions. An embedding layer was used to learn dense semantic representations of tokens, followed by a Bidirectional LSTM layer and a softmax output layer for multi-class classification.

```
Macro F1 Score: 0.9465468837731612
```

The proposed model achieved a macro F1 score of approximately 0.95, indicating strong and balanced performance across all toxicity categories. The confusion matrix demonstrates a strong diagonal, confirming accurate classification for both majority and minority classes, with only minor confusion between semantically similar categories.



The training and validation loss curves indicate rapid convergence of the model. Training loss decreases sharply and approaches zero within a few epochs, demonstrating that the Bidirectional LSTM is able to fit the training data effectively. Validation loss remains relatively low throughout training but shows a slight upward trend in later epochs. This behavior suggests **mild overfitting**, which is expected given the increased model capacity and limited dataset size. However, the gap between training and validation loss remains small, and validation performance stays consistently high, indicating that the model retains good generalization ability rather than memorizing the training data.



The confusion matrix reveals strong classification performance across nearly all toxicity categories, with a dominant diagonal indicating correct predictions for the majority of samples. Classes such as *Child Sexual Exploitation*, *Elections*, *Non-Violent Crimes*, *Sex-Related Crimes*, *Suicide & Self-Harm*, *Violent Crimes*, and *unsafe* are classified almost perfectly. Minor confusion is observed between the *Unknown S-Type* and *Safe* categories, which is understandable due to their semantic similarity and inherent ambiguity