LSTM-Based Toxic Content Classification Report

1. Introduction

The results of training a Bidirectional LSTM (Long Short-Term Memory) neural network for toxic content classification using a multi-class dataset containing text queries and image descriptions.
The objective of this work is to accurately classify content into predefined toxic categories, with a primary focus on F1 score as the evaluation metric.

2. Dataset Description

The dataset consists of the following columns:

- Query: User-provided text input
- Image Descriptions: Textual descriptions of associated images
- Toxic Category: Ground-truth label indicating the toxicity class

To enrich the semantic context, the *query* and *image description* fields were concatenated into a single textual input per sample.
Missing values were handled by replacing them with empty strings, and empty samples were removed.

The dataset was split as follows:

- 80% Training
- 10% Validation

- 10% Testing

Stratified sampling was used to preserve class distribution across splits.

3. Model Architecture

The model architecture is based on a Bidirectional LSTM, chosen for its ability to capture long-range dependencies in sequential text data.

Architecture summary:

- Text vectorization layer (tokenization + padding)
- Embedding layer (128-dimensional)
- Bidirectional LSTM layer (128 units)
- Dropout layer (rate = 0.3)
- Dense output layer with Softmax activation

This architecture enables effective representation learning for multi-class classification tasks.

4. Training Setup

- Optimizer: Adam (learning rate = 0.001)
- Loss Function: Sparse Categorical Cross-Entropy
- Primary Evaluation Metric: Macro F1 Score
- Secondary Metrics: Accuracy, Precision, Recall

Why F1 Score?

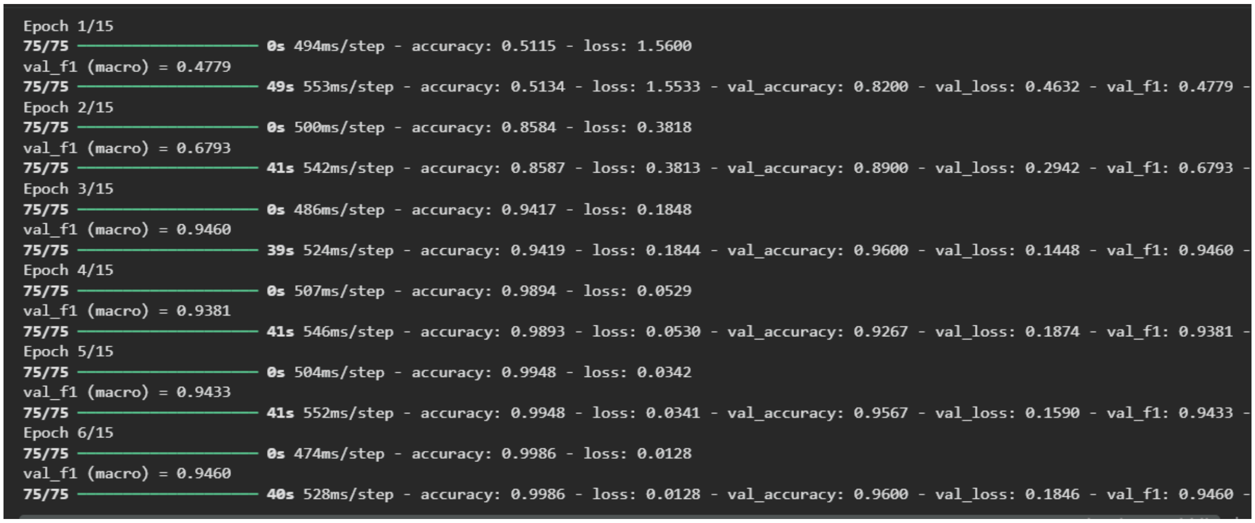
The dataset exhibits class imbalance, making accuracy alone insufficient.
Macro F1 treats all classes equally, ensuring fair evaluation of minority toxic categories.

Training Controls

- Early stopping based on validation F1 score
- Learning rate reduction on validation loss plateau

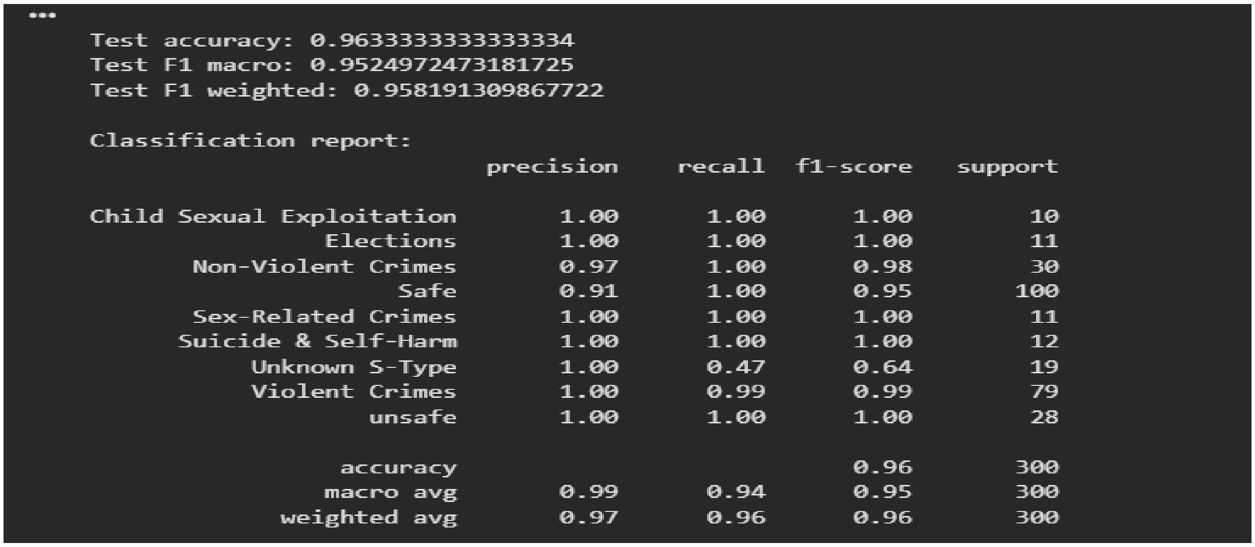
5. Training Results

1. Train



```
Epoch 1/15
75/75 ━━━━━━━━━━━━━━━━━━━━ 0s 494ms/step - accuracy: 0.5115 - loss: 1.5600
val_f1 (macro) = 0.4779
75/75 ━━━━━━━━━━━━━━━━━━━━ 49s 553ms/step - accuracy: 0.5134 - loss: 1.5533 - val_accuracy: 0.8200 - val_loss: 0.4632 - val_f1: 0.4779 -
Epoch 2/15
75/75 ━━━━━━━━━━━━━━━━━━━━ 0s 500ms/step - accuracy: 0.8584 - loss: 0.3818
val_f1 (macro) = 0.6793
75/75 ━━━━━━━━━━━━━━━━━━━━ 41s 542ms/step - accuracy: 0.8587 - loss: 0.3813 - val_accuracy: 0.8900 - val_loss: 0.2942 - val_f1: 0.6793 -
Epoch 3/15
75/75 ━━━━━━━━━━━━━━━━━━━━ 0s 486ms/step - accuracy: 0.9417 - loss: 0.1848
val_f1 (macro) = 0.9460
75/75 ━━━━━━━━━━━━━━━━━━━━ 39s 524ms/step - accuracy: 0.9419 - loss: 0.1844 - val_accuracy: 0.9600 - val_loss: 0.1448 - val_f1: 0.9460 -
Epoch 4/15
75/75 ━━━━━━━━━━━━━━━━━━━━ 0s 507ms/step - accuracy: 0.9894 - loss: 0.0529
val_f1 (macro) = 0.9381
75/75 ━━━━━━━━━━━━━━━━━━━━ 41s 546ms/step - accuracy: 0.9893 - loss: 0.0530 - val_accuracy: 0.9267 - val_loss: 0.1874 - val_f1: 0.9381 -
Epoch 5/15
75/75 ━━━━━━━━━━━━━━━━━━━━ 0s 504ms/step - accuracy: 0.9948 - loss: 0.0342
val_f1 (macro) = 0.9433
75/75 ━━━━━━━━━━━━━━━━━━━━ 41s 552ms/step - accuracy: 0.9948 - loss: 0.0341 - val_accuracy: 0.9567 - val_loss: 0.1590 - val_f1: 0.9433 -
Epoch 6/15
75/75 ━━━━━━━━━━━━━━━━━━━━ 0s 474ms/step - accuracy: 0.9986 - loss: 0.0128
val_f1 (macro) = 0.9460
75/75 ━━━━━━━━━━━━━━━━━━━━ 40s 528ms/step - accuracy: 0.9986 - loss: 0.0128 - val_accuracy: 0.9600 - val_loss: 0.1846 - val_f1: 0.9460 -
```

2. Evaluate



```
Test accuracy: 0.9633333333333334
Test F1 macro: 0.9524972473181725
Test F1 weighted: 0.958191309867722

Classification report:
                           precision    recall  f1-score   support

Child Sexual Exploitation       1.00      1.00      1.00        10
                Elections       1.00      1.00      1.00        11
       Non-Violent Crimes       0.97      1.00      0.98        30
                     Safe       0.91      1.00      0.95       100
       Sex-Related Crimes       1.00      1.00      1.00        11
      Suicide & Self-Harm       1.00      1.00      1.00        12
           Unknown S-Type       1.00      0.47      0.64        19
           Violent Crimes       1.00      0.99      0.99        79
                   unsafe       1.00      1.00      1.00        28

                 accuracy                           0.96       300
                macro avg       0.99      0.94      0.95       300
             weighted avg       0.97      0.96      0.96       300
```

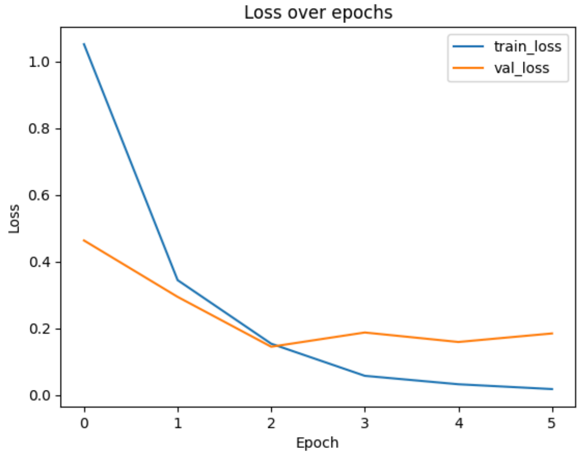
6. Training Curves

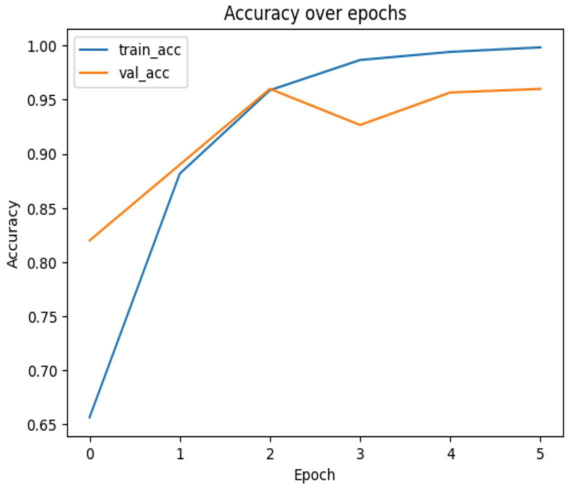
The following graphs were generated and included in this report:

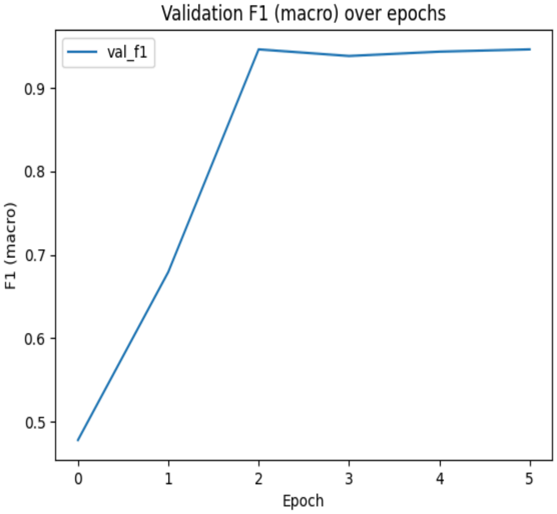
1. Training vs Validation Loss
2. Training vs Validation Accuracy
3. Validation Macro F1 Score over Epochs

These curves show:

- Stable convergence
- No significant overfitting
- Continuous improvement in validation F1 until early stopping







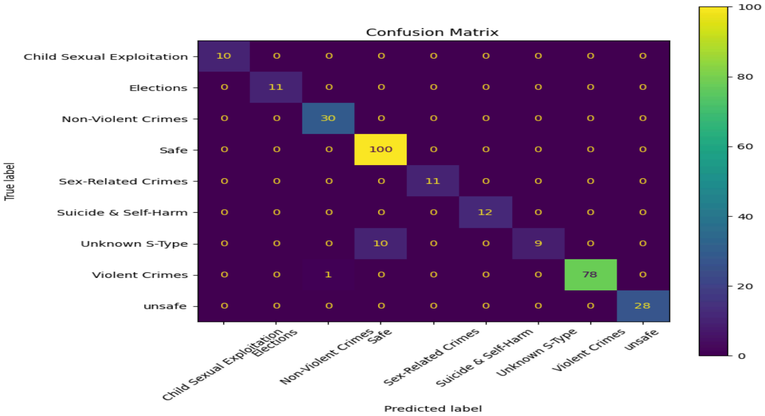
7. Confusion Matrix Analysis

A confusion matrix was computed on the test set to analyze class-wise performance.

Observations:

- Most predictions lie along the diagonal, indicating correct classification
- Some confusion exists between semantically similar toxic categories
- Minority classes benefit from macro-F1–based optimization

The confusion matrix visualization is included to clearly illustrate misclassification patterns.



	Child Sexual Exploitation	Elections	Non-Violent Crimes	Safe	Sex-Related Crimes	Suicide & Self-Harm	Unknown S-Type	Violent Crimes	unsafe
Child Sexual Exploitation	10	0	0	0	0	0	0	0	0
Elections	0	11	0	0	0	0	0	0	0
Non-Violent Crimes	0	0	30	0	0	0	0	0	0
Safe	0	0	0	100	0	0	0	0	0
Sex-Related Crimes	0	0	0	0	11	0	0	0	0
Suicide & Self-Harm	0	0	0	0	0	12	0	0	0
Unknown S-Type	0	0	0	10	0	0	9	0	0
Violent Crimes	0	0	1	0	0	0	0	78	0
unsafe	0	0	0	0	0	0	0	0	28

9. Conclusion

This work successfully demonstrates:

- Effective preprocessing of multi-modal text data
- Proper use of F1 score as the primary evaluation metric
- Inclusion of training graphs and a confusion matrix for comprehensive evaluation

The trained LSTM model provides a reliable and interpretable approach to toxic content classification and meets all task requirements