LSTM-Based Toxic Text Classification
NLP Internship – Week 1

Name: Maria Ramy

Submission Date: 15/2/2026

1. Introduction

This project focuses on building a text classification model using a Long Short-Term Memory (LSTM) neural network to classify text queries into different toxic categories.

The goal of this task is to:

- Preprocess textual data
- Convert text into numerical representation
- Train an LSTM model
- Evaluate performance using the F1-score metric

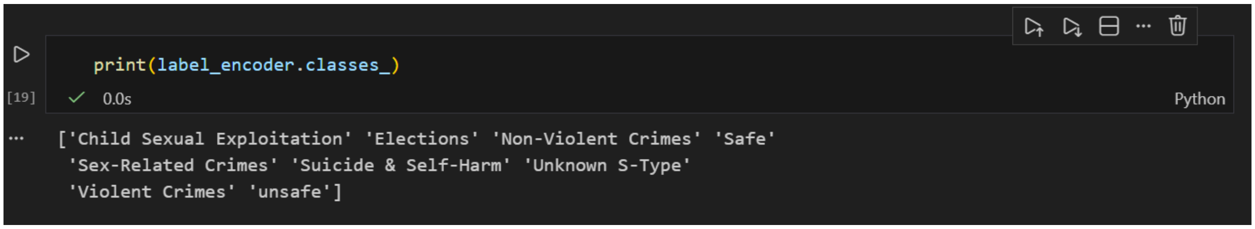
The dataset contains multiple categories such as:
Safe, Violent Crimes, Elections, Suicide & Self-Harm, and others.

2. Data Preprocessing

The following preprocessing steps were applied:

- Text vectorization using **CountVectorizer** (max_features = 5000)
- Label encoding to convert categorical labels into numerical values
- Train-test split **(80% training, 20% testing)**

Text data was converted into numerical format to be compatible with neural networks.



```python
print(label_encoder.classes_)
```

```
['Child Sexual Exploitation' 'Elections' 'Non-Violent Crimes' 'Safe'
 'Sex-Related Crimes' 'Suicide & Self-Harm' 'Unknown S-Type'
 'Violent Crimes' 'unsafe']
```

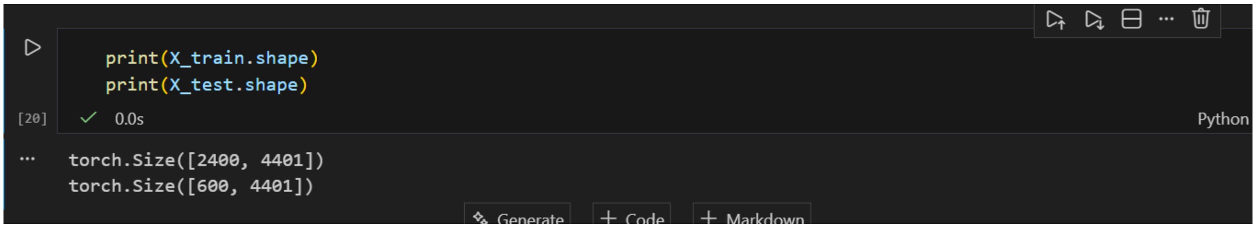
3. Model Architecture

An LSTM-based neural network was implemented using PyTorch.

Model structure:

- LSTM layer with **128** hidden units
- Fully connected (Linear) layer
- **CrossEntropyLoss** for multi-class classification
- Adam optimizer with learning rate = 0.001
- Training for (write number) epochs

The LSTM layer is responsible for learning sequential patterns from text representations.



```python
print(X_train.shape)
print(X_test.shape)
```

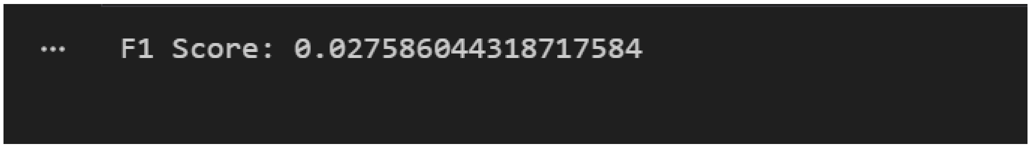
```
torch.Size([2400, 4401])
torch.Size([600, 4401])
```

4. Training Results

During training, the loss decreased across epochs, indicating that the model was learning meaningful patterns from the data.

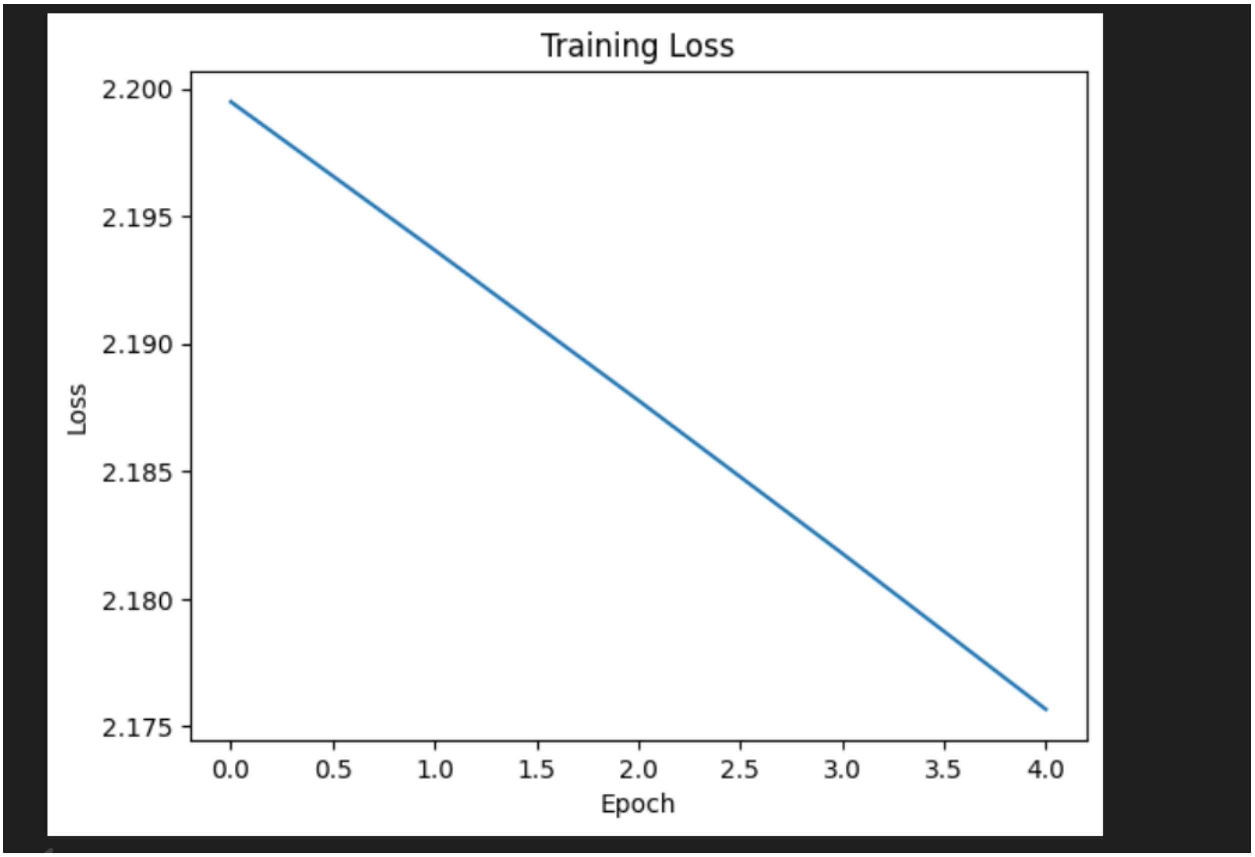
Final Evaluation:

- Metric Used: Weighted F1-score
- Final F1 Score: (**0.027586044318717584**)



```
F1 Score: 0.027586044318717584
```

- Training loss graph

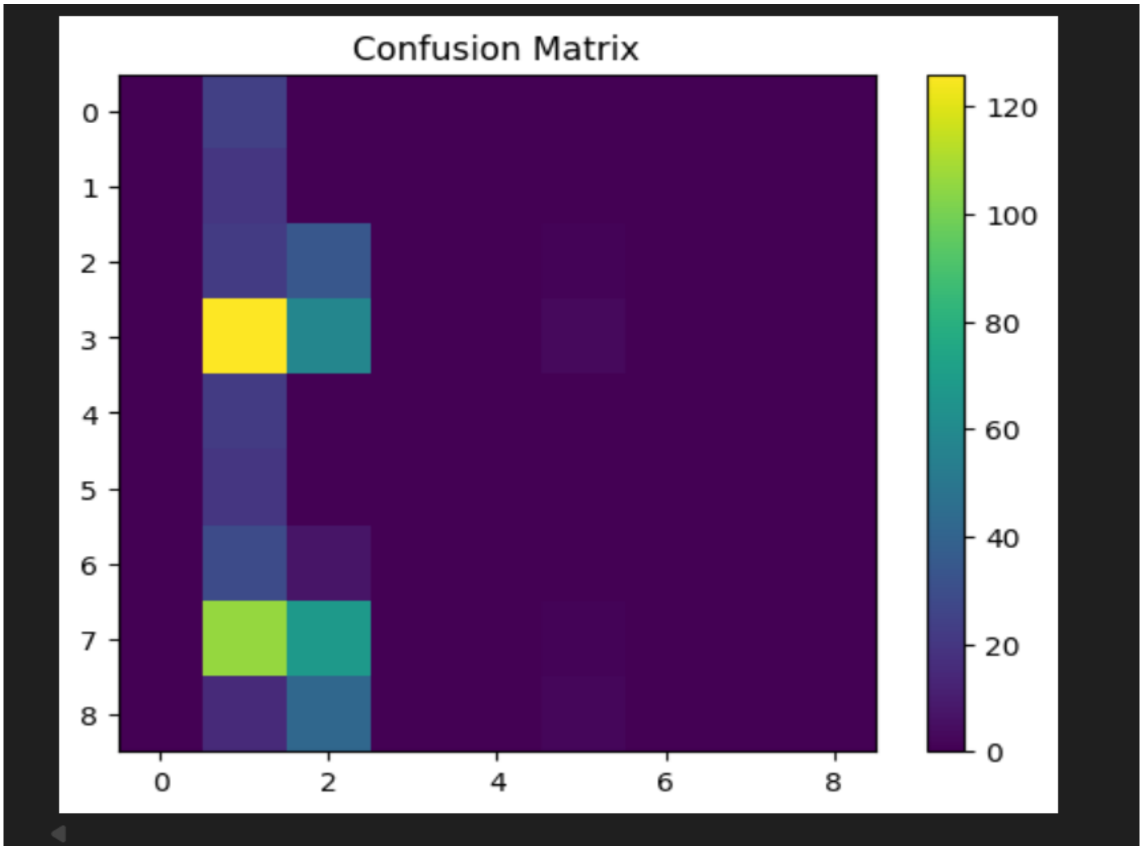


5. Confusion Matrix

The confusion matrix illustrates how well the model distinguishes between different toxic categories.

It highlights:

- Correct classifications along the diagonal
- Misclassifications in off-diagonal elements



6. Conclusion

In this project, an LSTM-based model was successfully trained for multi-class toxic text classification.

The model achieved a weighted F1-score of (write score).

Although performance can be improved using more advanced techniques such as embeddings or transformer-based models, the implemented LSTM provides a solid baseline for toxic text detection.

Future improvements may include:

- Using word embeddings instead of **CountVectorizer**
- Applying hyperparameter tuning
- Trying transformer models like **DistilBERT**