Performance Analysis Report

Toxic Content Detection Using ML & NLP Models

Name: Mohammod Abdullah Bin Hossain

Date: 07/04/2025

Assignment: SM Technology AI Developer Task

1. Executive Summary

This task aims to detect toxic content in online feedback using machine learning and NLP models. Three models were implemented: Random Forest as a baseline, LSTM for sequential learning, and XLM-RoBERTa for multilingual understanding. The XLM model showed the most balanced performance, achieving 78% accuracy and an ROC-AUC of 0.54 on the test set. Random Forest and LSTM performed well on non-toxic content but struggled with detecting toxicity.

2. Problem Statement

The goal is to build a model that detects various types of toxic content in user feedback. The key challenges include class imbalance (with fewer toxic examples) and multilingual content in validation and test datasets, requiring models that can generalize across languages.

3. Dataset Overview

The dataset includes two primary files:
- train.csv (labeled data)
- test.csv (unlabeled data for prediction)
- validation.csv (labeled data)
- test_labels.csv

Each entry contains a 'feedback_text' and binary labels like 'toxic', 'abusive', 'vulgar', 'menace', 'offense', and 'bigotry'. The dataset is imbalanced, with significantly more non-toxic examples.

4. Modeling Approach

- Random Forest: Used as a strong, interpretable baseline model.
- LSTM: Captures long-term dependencies in textual data.

- XLM-RoBERTa: Leverages multilingual Transformer embeddings for broader generalization.

5. Evaluation Metrics

The following metrics were used:
- Accuracy: Overall correctness.
- Precision, Recall, F1-Score: To measure performance on individual classes.
- ROC-AUC: Indicates the model's ability to distinguish between classes. Recall and ROC-AUC for the toxic class are especially crucial due to its low frequency and importance.

6. Performance Results

Model	Accuracy (Val)	ROC-AUC (Val)	Accuracy (Test)	ROC-AUC (Test)	Notes
Random Forest	59.17%	0.5470	61.30%	0.5021	Baseline: weak on toxic recall
LSTM	84.05%	0.5000	77.28%	0.5000	Overfitted to non-toxic class
XLM-RoBERTa	83.93%	0.5295	78.18%	0.5405	Best balance; multilingual support

Random Forest underperformed on the minority class due to imbalance. LSTM had strong accuracy but failed to generalize for toxic class. XLM-RoBERTa provided the most reliable results across languages and class distributions.

▩ Verdict

Overall, the evaluation results reflect a commendable implementation effort across baseline and advanced models. While the Random Forest model served as a reliable starting point, its limited ability to capture contextual nuances led to low performance on the toxic class. The LSTM model exhibited strong generalization on non-toxic content but failed to detect toxic feedback, indicating overfitting to the majority class. XLM-RoBERTa emerged as the most balanced and promising model, especially in handling multilingual data and improving toxic content detection. Although current results are not yet optimal for production-level deployment, they lay a solid foundation for future improvements through data balancing, fine-tuning, and advanced training strategies.

7. Challenges Faced

- Severe class imbalance leading to low toxic recall.
- Models overfitting on majority class.
- Difficulty generalizing to multilingual data.

8. Future Improvements

- Implement focal loss or class-weighted loss functions.
- Apply data augmentation for minority class.
- Use pretraining with more multilingual corpora.
- Explore ensemble methods.
- Expand hyperparameter search ranges.

9. Conclusion

The XLM-RoBERTa model showed the most promise in balancing accuracy and recall across classes and languages. With further tuning and data strategies, the model can be adapted for real-world toxic content detection tasks.

10. Appendix (Optional)

- Training logs
- RUC AUC curve, Confusion Matrix
- Snippets of training code used in PyTorch and HuggingFace
- saved in the repository(all images and result matrix)