Enhanced Multilingual Toxic Comment Detection Pipeline using BERT: Threshold Optimization and Weighted Sampling for Improved Class Imbalance Handling

Introduction:

Online Toxicity comment harm individuals and community mental health. However, detecting these can be challenging because of the imbalanced datasets and multilingual language. This project helps detecting those toxic comments using four different approaches.

- Transformer based model: XLM-RoBERTa with weighted sampling and threshold optimization.
- GRU-based Deep Learning Model: Bidirectional GRU architecture with weighted sampling.
- LR: A baseline linear approach using TF-IDF features
- RF: An ensemble-based method leveraging decision trees on TF-IDF features.

Result Analysis: We evaluate the four models and the models are compared on **F1-score**, **precision**, and **recall** which is crucial in detecting minority class like toxicity.

1. BERT (XLM-Roberta-Base) with Weighted Sampling & Threshold Optimization

Epoch	Train Loss	Val Loss	Precision	Recall	F1-score
1	0.1747	0.0821	0.622	0.455	0.526
2	0.0898	0.0907	0.675	0.403	0.504
3	0.0473	0.1247	0.724	0.313	0.437

Best F1: 0.526 at threshold ≈ 0.00

Comments: This model achieves highest F1 score which shows the ability to generalize. This is achieved by exploiting weighted sampling and threshold tuning for optimum decision boundaries.

2. GRU-Based RNN

Epoch	Train Loss	Val F1
1	1.1687	0.3033
2	0.5155	0.3571
3	0.2621	0.3402
4	0.2191	0.3580
5	0.1922	0.2638
6	0.1243	0.2581
7	0.0978	0.2633

Best F1: 0.3580 at threshold ≈ 0.51

Comments: This model is underperformed compared to BERT because of it's limitation of understanding of complex multilingual context.

3. Logistic Regression (TF-IDF)

Class	Precision	Recall	F1-score	Support
Toxic (1)	0.18	0.80	0.29	134
Non-toxic (0)	0.89	0.30	0.45	705
Macro Avg	**0.53**	**0.55**	**0.37**	839

Comments: Recall is high but the precision is low which leads a greater number of false positives.

4. Random Forest

Class	Precision	Recall	F1-score	Support
Toxic (1)	0.29	0.27	0.28	134
Non-toxic (0)	0.86	0.87	0.87	705
Macro Avg	**0.58**	**0.57**	**0.57**	839

Comments: RF determines low recall for toxic class interprets that struggling of generalizing of toxic class. It performs better than LR in case of macro F1

Conclusion:

Model	Best F1	Precision (Toxic)	Recall (Toxic)	Strength
BERT	**0.526**	0.62 – 0.72	0.31 – 0.45	Contextual understanding, best toxic detection
GRU	0.358	Moderate	Moderate	Lightweight, average multilingual modeling
Random Forest	0.28	0.29	0.27	Better on non-toxic, limited toxic recall
Logistic Reg.	0.29	0.18	**0.80**	High recall, many false positives