Performance Analysis Report: Offensive Language Classification

Md. Nurnobi Islam

April 13, 2025

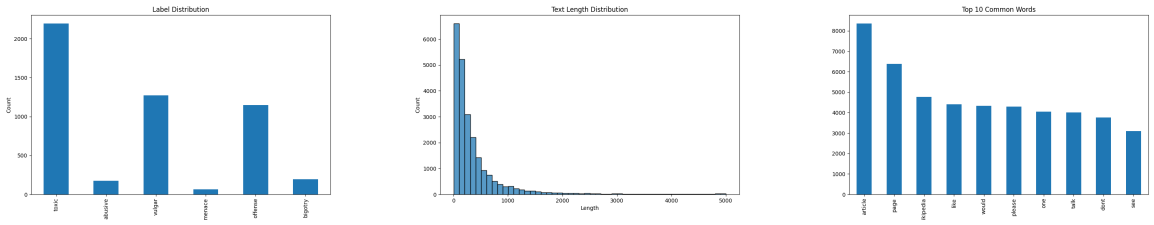
1 Introduction

This report summarizes the development and evaluation of machine learning models for detecting toxic content in online feedback, as part of the Offensive Language Classification task. Three models were implemented: Logistic Regression (baseline), GRU (sequential), and BERT (transformer-based). The goal was to predict the binary `toxic` label, leveraging fine-grained labels during training, with evaluation on multilingual validation and test data.

2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) revealed key dataset characteristics:

- **Label Distribution**: The training data includes six binary labels (`toxic`, `abusive`, `vulgar`, `menace`, `offense`, `bigotry`). The `toxic` label is the most frequent, but class imbalance exists (Figure 1a).
- **Text Length**: Feedback lengths vary widely, with most under 500 characters, indicating diverse comment styles (Figure 1b).
- **Word Frequency**: Common words (post-stopword removal) include domain-specific terms, suggesting context matters (Figure 1c).
- **Missing Values**: No significant missing data was found, ensuring robust preprocessing.



(a) Label Distribution (b) Text Length (c) Word Frequency

Figure 1: EDA Visualizations

3 Model Performance

Three models were trained and evaluated on the validation set, focusing on the `toxic` label. Metrics include accuracy, precision, recall, F1-score, and AUC-ROC.

3.1 Logistic Regression

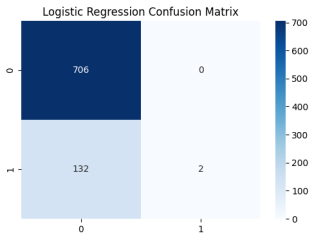
- **Setup**: TF-IDF features (5000 max), tuned with Grid Search (`C=10`).
- **Metrics**:
  - Accuracy: 0.8429
  - Precision: 1.0000
  - Recall: 0.0149
  - F1-Score: 0.0294
  - AUC-ROC: 0.5861
  - Tuned F1-Score: 0.0426
- **Visualizations**: See Figures 2a and 3a.

3.2 GRU

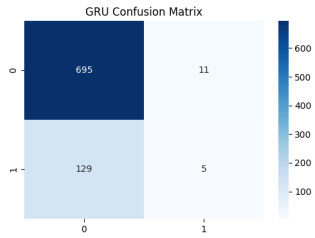
- **Setup**: Bidirectional GRU (64+32 units), 5 epochs, batch size 32.
- **Metrics** (Epoch 5):
  - Accuracy: 0.8333
  - Precision: 0.3125
  - Recall: 0.0373
  - F1-Score: 0.0667
  - AUC-ROC: 0.4873
- **Training**: Train accuracy 0.9861, validation loss 1.3478 (overfitting).
- **Visualizations**: See Figures 2b and 3b.

3.3 BERT

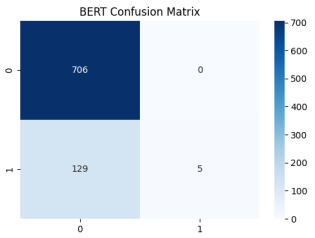
- **Setup**: Fine-tuned `bert-base-multilingual-cased`, 3 epochs (initial), 4 epochs (tuned), batch sizes 16 and 8.
- **Metrics** (Epoch 3):
  - Accuracy: 0.8464
  - Precision: 1.0000
  - Recall: 0.0373
  - F1-Score: 0.0719
  - AUC-ROC: 0.8296
- **Tuned Metrics** (Epoch 4):
  - F1-Score: 0.0000 (overfitting).
- **Training**: Train loss 0.1585 (Epoch 3), validation loss 0.6634.
- **Visualizations**: See Figures 2c and 3c.



(a) LR Confusion Matrix

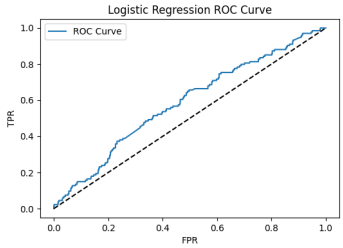


(b) GRU Confusion Matrix

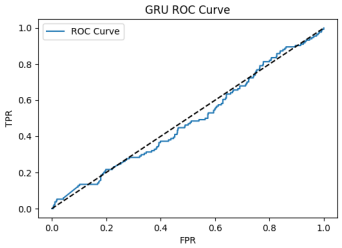


(c) BERT Confusion Matrix

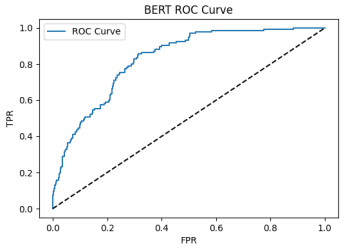
Figure 2: Confusion Matrices



(a) LR ROC Curve



(b) GRU ROC Curve



(c) BERT ROC Curve

Figure 3: ROC Curves

4 Model Comparison

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.8429	1.0000	0.0149	0.0294	0.5861
GRU	0.8333	0.3125	0.0373	0.0667	0.4873
BERT	0.8464	1.0000	0.0373	0.0719	0.8296

Table 1: Model Performance Comparison

- **BERT**: Highest AUC-ROC (0.8296) and F1-score (0.0719), excels in multilingual contexts.
- **Logistic Regression**: High precision, low recall (0.0149), limited by TF-IDF.
- **GRU**: Poor AUC-ROC (0.4873), overfits, not multilingual.

5 Challenges and Solutions

- **Class Imbalance**: Low recall (e.g., 0.0373 for BERT) due to few `toxic` samples.
  - *Solution*: Considered weighted loss; future: SMOTE.
- **Multilingual Data**: Non-English validation/test data challenged LR/GRU.
  - *Solution*: BERT's pre-training handled this well.
- **Overfitting**: GRU (val loss 1.3478) and tuned BERT (F1=0.0000) overfit.

  - *Solution*: Early stopping in BERT; GRU needs regularization.

- **Computation**: BERT training slow (1915s).
  - *Solution*: Used Colab GPU.

6 Conclusion

BERT is the most effective model (AUC-ROC 0.8296, F1 0.0719), ideal for multilingual moderation. Low recall across models suggests imbalance handling is needed. Future work: threshold tuning, weighted loss, ensembles.