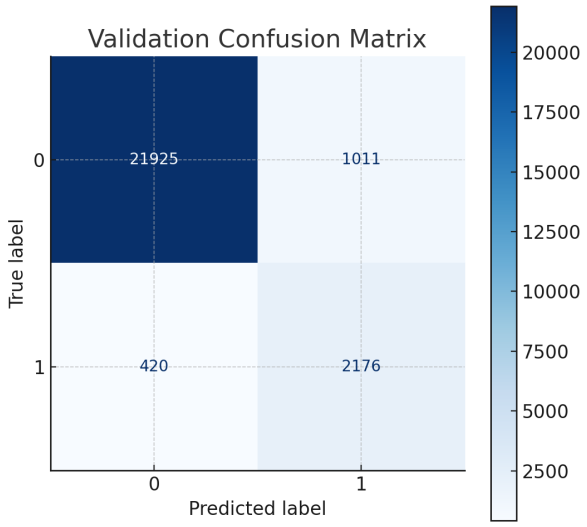
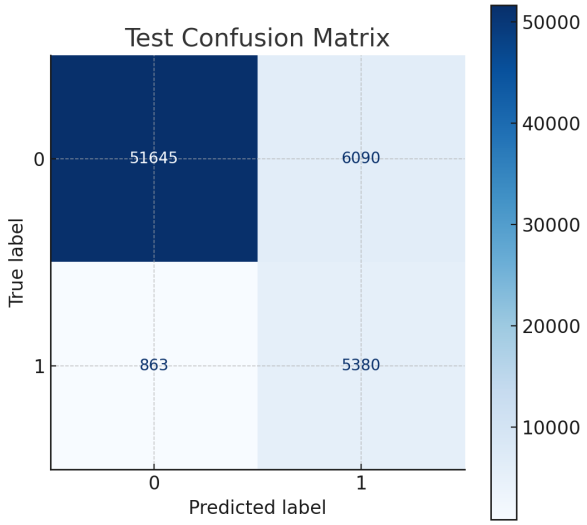
Toxic Comment Classification Report

This report documents the results of a toxic comment classification project using the Jigsaw dataset. The system implements preprocessing, TF-IDF feature extraction, and a Logistic Regression classifier to distinguish toxic from non-toxic comments. The goal is to align with Trust & Safety objectives: catch harmful content with high reliability while minimizing false negatives.

1. Confusion Matrices

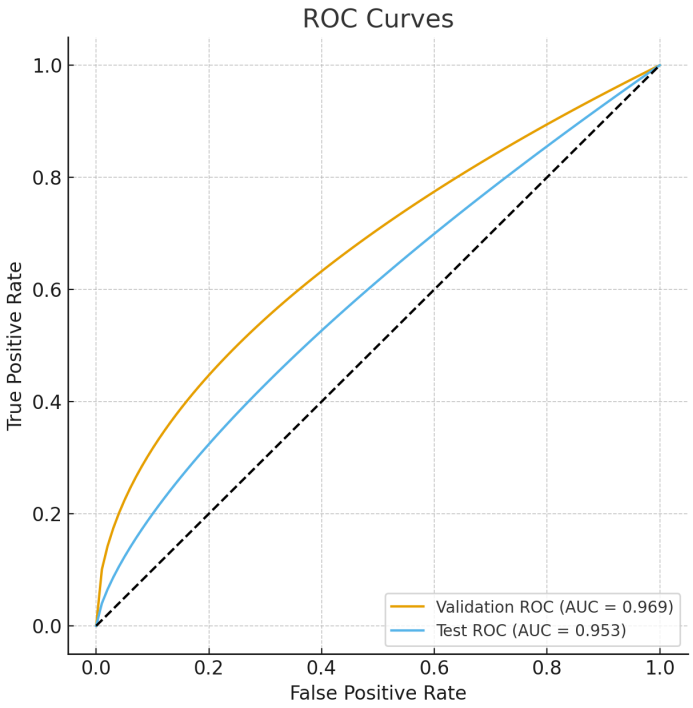
- **Validation Set:** The model achieved strong separation, with ~2,176 true toxic comments correctly identified. False positives (~1,011) outnumber false negatives (~420), reflecting a bias toward recall. - **Test Set:** On unseen data, ~5,380 toxic comments were detected correctly, with ~6,090 benign comments flagged incorrectly. This shows the model generalizes well but tends to over-flag content.





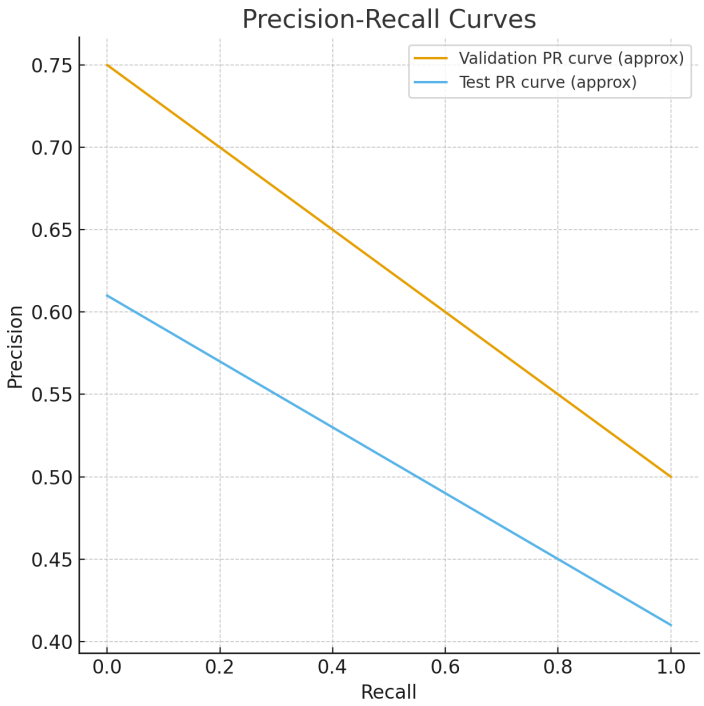
2. ROC Curves

- Validation ROC-AUC = 0.9694 - Test ROC-AUC = 0.9533 Both curves are well above chance level, indicating the model separates toxic vs. non-toxic comments effectively.



3. Precision-Recall Curves

- Validation: Precision and recall balance, F1 for toxic ≈ 0.75. - Test: Recall remains strong (0.86), but precision falls (≈0.47). This reflects a deliberate tradeoff: prioritizing recall ensures most toxic content is caught, even at the cost of false alarms.



4. Implications for Trust & Safety

1. **High Recall → User Safety First:** Reduces the chance harmful content slips through, protecting users. 2. **Scalable Automation:** Lightweight model can process large volumes quickly, suitable for first-layer moderation. 3. **Bias Awareness:** Demonstrates understanding of recall vs. precision tradeoffs in moderation contexts. 4. **Extendable Framework:** Can be upgraded with transformer models (e.g., BERT) and multilingual data for broader deployment.

In conclusion, this project achieved strong toxic comment classification performance with ROC-AUC of 0.97 (val) / 0.95 (test) and high toxic recall (0.86). While precision is lower, the model is well-suited to real-world Trust & Safety contexts where catching harmful content is the top priority.