📊 Performance Analysis Report

📝 Project Summary

The objective of this project is to classify online feedback based on toxicity. The dataset is multi-label, containing six categories that represent different types of toxic behavior: toxic, abusive, vulgar, menace, offense, and bigotry. However, the final evaluation focuses solely on the toxic label.

Model 1, a baseline Logistic Regression model using TF-IDF features, was successfully implemented and achieved an accuracy of 77% on the validation set. Due to time constraints, Model 2 (Transformer-based XLM-Roberta) remains incomplete.

🔍 Dataset Overview

Files Provided:

train.csv: Labeled data with feedback and six toxicity labels.

test.csv: Unlabeled data for prediction (focused on the "toxic" label).

Key Columns in train.csv:

id: Unique identifier

feedback_text: User-submitted feedback

toxic, abusive, vulgar, menace, offense, bigotry: Binary labels (1 = present, 0 = not present)

The dataset is primarily in English but may include multilingual text.

⚙️ Model 1: Logistic Regression (Baseline)

🔧 Preprocessing Steps

Lowercased all text

Removed stopwords, punctuation, and special characters

Applied lemmatization

Transformed text using TF-IDF Vectorization

🧠 Model Description

Algorithm: Logistic Regression (from scikit-learn)

Features: TF-IDF vectors

Training Strategy: Stratified split for training and validation

Focused on the toxic label for evaluation; auxiliary labels helped enhance feature representation

📈 Model 1 Evaluation Metrics (on 6,000 validation samples)

Class	Precision	Recall	F1-Score	Support
0 (Non-Toxic)	0.77	1.00	0.87	4,637
1 (Toxic)	0.12	0.00	0.00	1,363

Overall Accuracy: 0.77

Macro Average F1-Score: 0.44

Weighted Average Precision: 0.63

✅ Key Observations

The combination of TF-IDF and Logistic Regression provides a strong and interpretable baseline.

The model performs very well on non-toxic comments but struggles to detect toxic feedback, indicating class imbalance and limitation in feature expressiveness.

Fast training and low resource requirements make it suitable for initial deployment or exploratory analysis.

The model may fail to generalize to multilingual or subtle toxicity due to the limitations of TF-IDF and surface-level text features.

❌ Model 2: Transformer (XLM-Roberta)
A second model using XLM-Roberta (multilingual BERT) was planned for fine-tuning via Hugging Face Transformers to better capture linguistic context and support multilingual text. Unfortunately, this was not completed due to time constraints.

📌 Conclusion
The Logistic Regression model demonstrates decent performance with 77% accuracy, making it a solid baseline for toxic comment detection. However, its poor recall on toxic comments indicates that more sophisticated models like transformer-based architectures are necessary for real-world applications—especially where context, nuance, and multilingual understanding are essential.

Future work will focus on implementing and evaluating the XLM-Roberta model to improve overall detection accuracy and robustness.