TransparentOx: A Compact Transformer-Based Toxicity Classifier

November 3, 2025

Abstract

TransparentOx fine-tunes transformer encoders to detect toxic online comments. It provides interpretable thresholding and an interactive Streamlit app. This report outlines dataset challenges, model design, results, and lessons learned on how data imbalance and comment length affect detection quality.

1 Introduction

Toxicity detection helps moderate online discussions but remains difficult when abusive language appears briefly or without context. *TransparentOx* fine-tunes `DistilBERT` on the Civil Comments dataset, using balanced loss and threshold tuning to improve interpretability and recall.

2 Dataset

We use the *Civil Comments* dataset (∼1.8M English comments, ∼9:1 non-toxic:toxic). Toxic comments are typically longer and more context-rich, leading to bias against short toxic phrases (e.g., *"idiot", "shut up"*). Sarcasm and coded language further challenge detection.

2.1 Preprocessing

Text is tokenized with the pretrained DistilBERT tokenizer (uncased), truncated or padded to 128 or 256 tokens. No extra cleaning or stopwording is used.

3 Methodology

3.1 Model and Training

We fine-tune `distilbert-base-uncased` for binary classification with:

- Class-weighted loss and optional focal loss,
- `WeightedRandomSampler` to rebalance training,
- Validation-based threshold selection to optimise F1.

Main hyperparameters for the best run (`distilbert-L128-e2-lr3e-5`): max length 128, learning rate 3×10^{-5}, 2 epochs, batch sizes 16/32, and warmup ratio 0.06. Evaluation uses accuracy, precision, recall, F1, ROC-AUC, and PR-AUC.

4 Results

4.1 Performance Summary

Using 1,000 test samples (910 non-toxic / 90 toxic):

Class	Precision	Recall	F1
Non-toxic	0.95	0.82	0.88
Toxic	0.24	0.59	0.35
Accuracy		0.80	
Macro-F1		0.61	
ROC-AUC		0.79	
PR-AUC		0.33	

4.2 Analysis

The model identifies longer toxic comments reliably but under-scores short, direct insults. Recall improves at lower thresholds (e.g., $t = 0.45$) but with more false positives. Default thresholds ($t \approx 0.5$) yield high accuracy but miss most toxic content.

5 Discussion

Length bias: learned patterns correlate with long, explicit text. **Class imbalance:** accuracy remains inflated by majority class dominance. **Thresholds:** adjusting t shifts precision–recall trade-offs; the Streamlit app allows users to explore this interactively.

6 Limitations and Future Work

Training used small subsets for speed, limiting generalisation. Short, high-toxicity examples are rare, and sarcasm remains hard to detect. Future work: data augmentation for short toxic phrases, full-dataset training, and stronger models (e.g., RoBERTa, DeBERTa). Explainability tools (e.g., SHAP) could highlight why specific words trigger toxicity.

7 Conclusion

TransparentOx demonstrates that lightweight transformers can classify toxicity effectively when properly balanced and thresholded. However, dataset imbalance and text-length bias reduce sensitivity to short toxic phrases, highlighting the need for better-curated data and threshold calibration.