Jigsaw Unintended Bias in Toxicity Classification

Hugo Mojica; Andrey Gonzales; Laura Paez.
'Universidad Distrital Francisco José de Caldas'

Abstract

In the digital age, toxic online communication is an ongoing challenge. This project addresses the "Jigsaw Unintended Bias in Toxicity Classification" competition, which highlights the difficulty of detecting toxic comments without unfairly penalizing identity-based language. We propose a systemic design that integrates classic NLP pipelines with modern Large Language Models (LLMs), chaos mitigation strategies, and fairness metrics. By evolving from a TF-IDF + Logistic Regression baseline to a RoBERTa-powered system, we achieved measurable improvements in both accuracy and bias reduction, despite deployment constraints.

INTRODUCTION

Automated moderation systems often rely on surface-level patterns that misclassify harmless identity mentions (e.g., "gay rights") as toxic. This results in unintended bias that disproportionately affects minority voices. The Jigsaw competition provides a realistic dataset containing multiple types of toxicity and annotated identity terms, enabling fairness-centric experimentation.Our approach is framed within systems engineering, where the classification problem is treated as a multi-component system. We progressively improved our architecture over three workshops, integrating preprocessing, contextual analysis, chaos management, and transformer-based embeddings.

GOAL

Research Question

How can systemic design and chaos-aware modules reduce unintended bias in toxicity detection systems

Objectives

Develop a fair and explainable toxicity classifier.

Minimize false positives in identity-related comments.

Handle linguistic ambiguity like sarcasm.

Expected Outcome:

A scalable model with ≥30% false positive reduction and improved performance on subgroup metrics.

PROPOSED SOLUTION

Modular Architecture

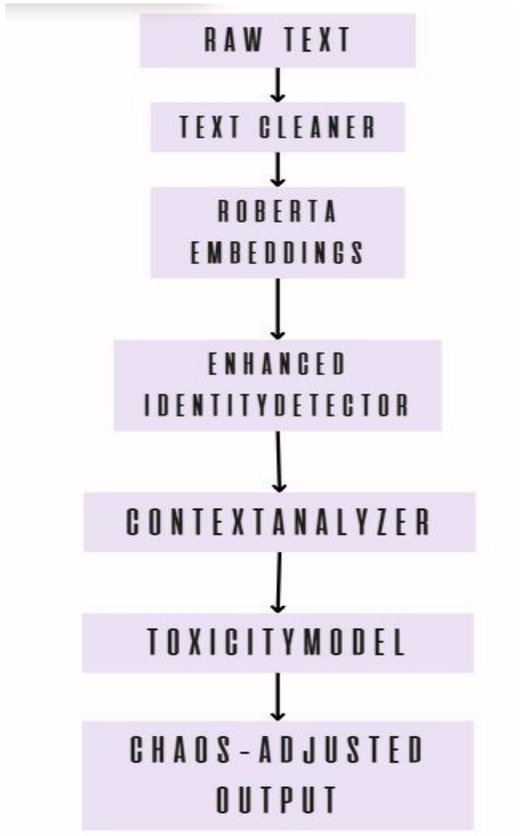

Figure 1. Arquitecture.

METHODS AND MATERIALS

Dataset: Jigsaw 2019 (Wikipedia comment corpus with toxicity annotations)

Tools: Python, Scikit-learn, HuggingFace Transformers, PyTorch

LLM Used: DistilRoBERTa for embeddings and context

Metrics:

Subgroup AUC (Fairness by identity group)

F1 Score (Sarcasm & Identity attacks)

FP Rate (False positives in non-toxic identity usage)

Conclusions

LLMs outperform traditional methods in handling ambiguous, identity-rich language.

Systemic modular design enabled better testing and bias control.

Deployment feasibility is as important as model performance.

Future Work: Integrate lightweight distilled models + Docker containers for portable, fair NLP pipelines.

DISCUSSION

Our architecture improved fairness and accuracy, but real-world deployment was blocked due to competition constraints (no internet access in Kaggle kernels). Despite preloading model weights and dependencies, the system failed silently due to memory limits and I/O restrictions.

Lessons learned:

Offline deployment must be designed from day 1.

LLMs are powerful but computationally expensive.

Hybrid models (lightweight + deep) offer the best trade-off.

Contact

Hugo Mojica, Laura Paez, Andrey Gonzales
Facultad de Ingeniería
Universidad Distrital Francisco José de Caldas

1. Jigsaw Unintended Bias: kaggle.com/jigsaw-unintended-bias
2. Liu et al., RoBERTa: arxiv.org/abs/1907.11692
3. Mehrabi et al., Bias Survey: arxiv.org/abs/1908.09635
4. Borkan et al., Subgroup Metrics: NAACL 2019