Hate Speech Detection with Bias Analysis - Social Media Comments

Renusri Periketi - 121272409
Ushasree Vuchidi - 121195692
Group 20

The Challenge:

- Billions of daily social media interactions
- Toxic content threatens user safety
- Manual moderation impossible at scale
- Must balance accuracy with fairness

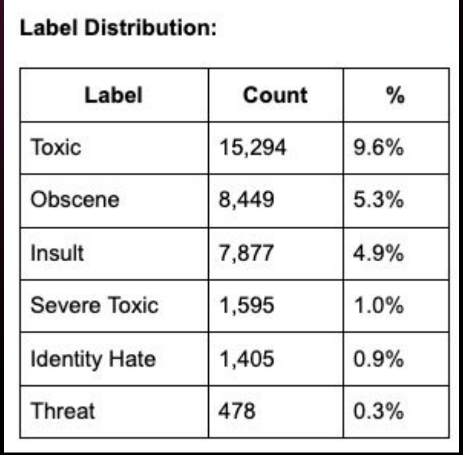
Research Questions:

1. How do traditional ML vs transformers compare?
2. What are the main error patterns?
3. Can models handle context (positive profanity)?
4. Do models show demographic bias?

Dataset

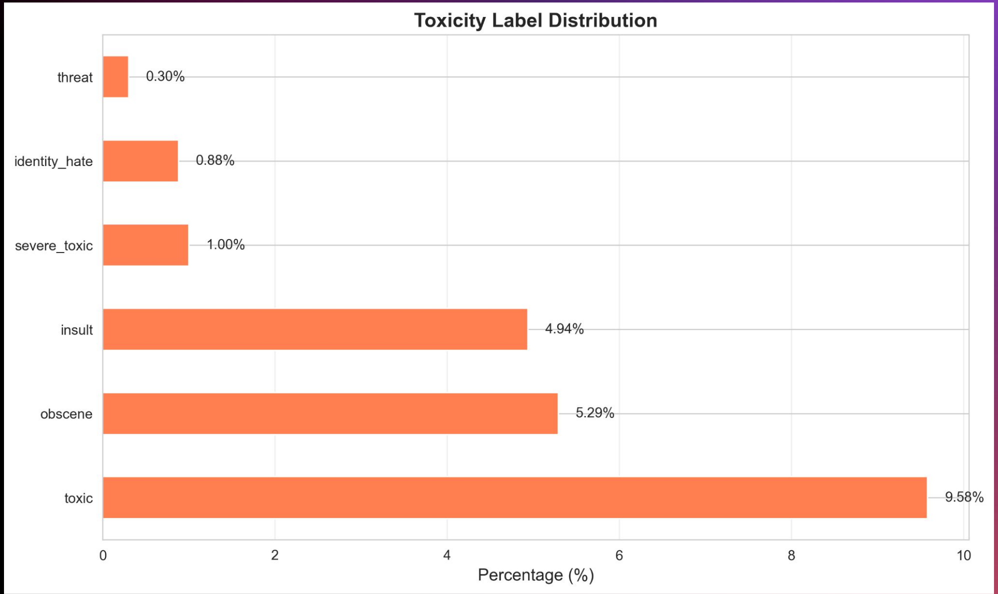
Jigsaw Toxic Comment Classification - Kaggle Challenge

- 159,571 Wikipedia comments
- 6 toxicity labels (multi-label)
- Severe class imbalance (90% clean, 10% toxic)



Label Distribution:

Label	Count	%
Toxic	15,294	9.6%
Obscene	8,449	5.3%
Insult	7,877	4.9%
Severe Toxic	1,595	1.0%
Identity Hate	1,405	0.9%
Threat	478	0.3%



Comparing Classification Models: Speed vs. Sophistication

Understanding the trade-offs between different machine learning models is crucial for effective deployment. Here, we delve into a comparison between a traditional baseline and an advanced neural network for text classification.

Baseline: Logistic Regression + TF-IDF

Pros: Fast training and inference, highly interpretable coefficients revealing feature importance. Utilizes 10,000 features (unigrams + bigrams) for robust representation.

Cons: Treats text as a "bag-of-words," ignoring word order and semantic context. Misses nuanced meanings.

Advanced: DistilBERT

Pros: Context-aware transformer model with 66 million parameters. Excellently captures word order and complex linguistic patterns for deeper understanding.

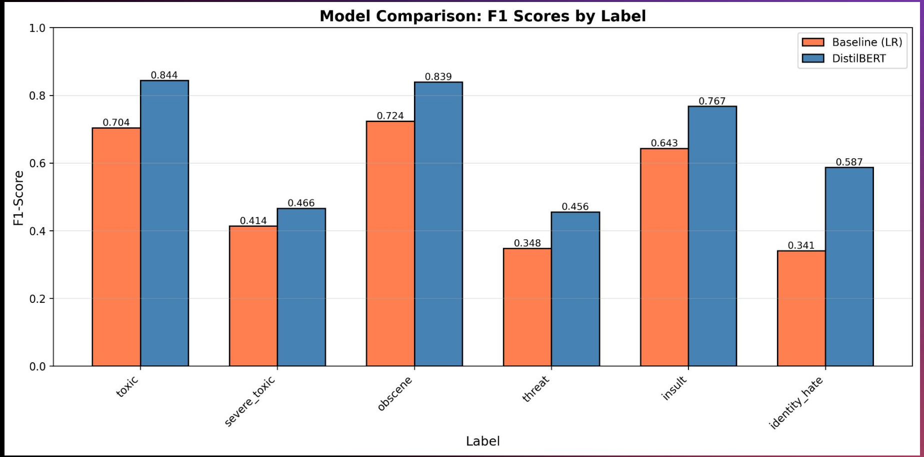
Cons: Slower inference times and significantly more computationally intensive for both training and deployment.

Training: 80/20 train/validation split - 3 epochs with early stopping - Macro F1 as primary metric

Results - Overall Performance

Per-Label Improvements:

- Best gains on rare labels:
  - Threat: +31% (0.348 → 0.456)
  - Identity Hate: +71.8% (0.341 → 0.587)
- Strong on frequent labels:
  - Obscene: 0.724
  - Insult: 0.643



Error Patterns & Bias Analysis

1. Error Analysis Highlights

- *DistilBERT Advantages*:
    - Catches 187 additional toxic comments (+3.7%)
    - Detects implicit toxicity
    - Understands context and word order
- *Common Failures*:
    - 60% false positives on positive profanity
    - Obfuscated text
    - Short comments (<5 words)
    - Implicit threats without profanity

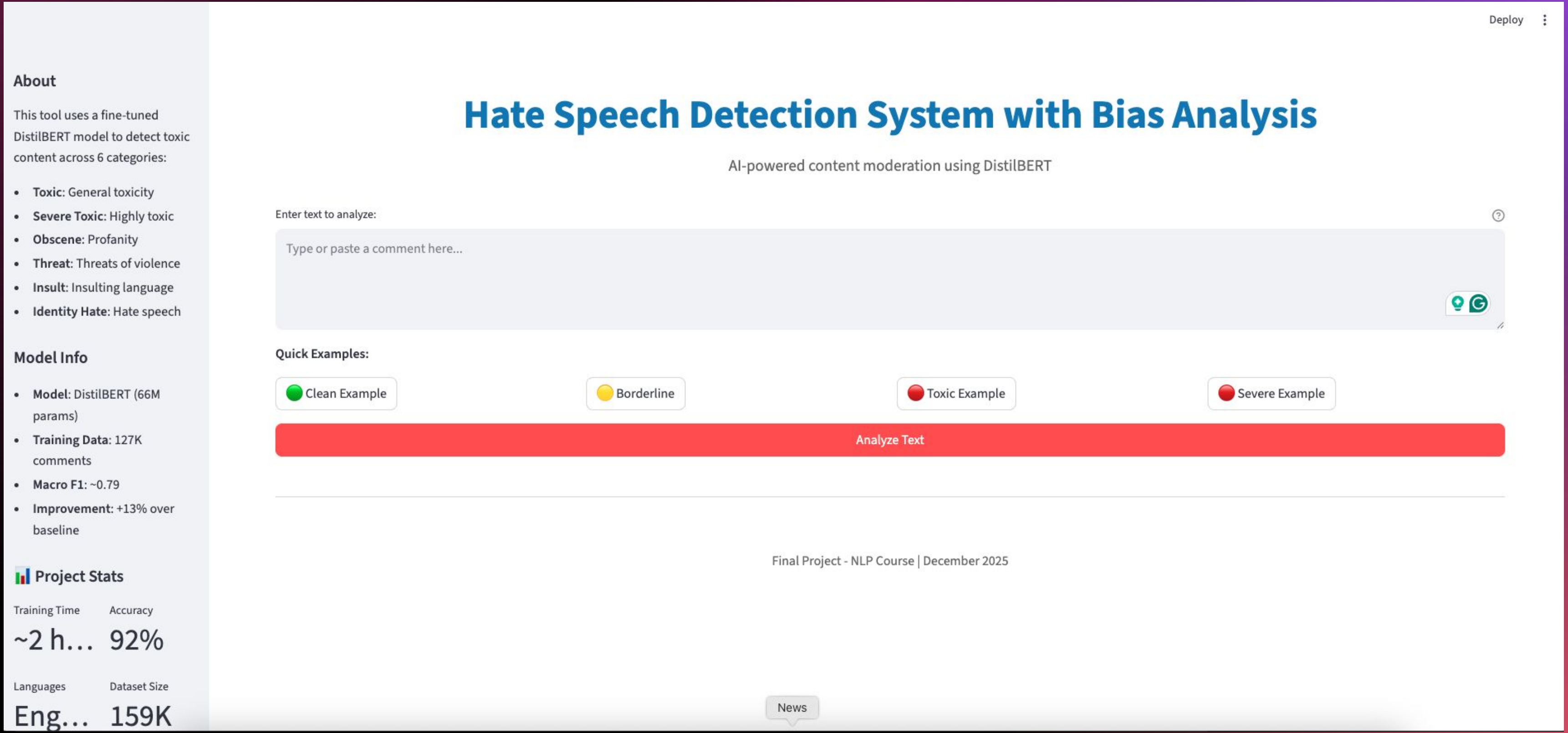
2. Bias Study Results

Identity Bias: Religion shows highest false positive rate (Muslim mentions 18.6%, p=0.024)

Dialect Bias: AAVE scored slightly higher (+3.8%, p=0.079, not statistically significant)

Impact: Even small biases can lead to disproportionate moderation of marginalized communities

Streamlit Interface



Deploy ⋮

About

This tool uses a fine-tuned DistilBERT model to detect toxic content across 6 categories:

- **Toxic**: General toxicity
- **Severe Toxic**: Highly toxic
- **Obscene**: Profanity
- **Threat**: Threats of violence
- **Insult**: Insulting language
- **Identity Hate**: Hate speech

Model Info

- **Model**: DistilBERT (66M params)
- **Training Data**: 127K comments
- **Macro F1**: ~0.79
- **Improvement**: +13% over baseline

📊 Project Stats

Training Time	Accuracy
~2 h…	92%
Languages	**Dataset Size**
Eng…	159K

Hate Speech Detection System with Bias Analysis

AI-powered content moderation using DistilBERT

Enter text to analyze:

Type or paste a comment here...

Quick Examples:

Final Project - NLP Course | December 2025

Live Demo

Network URL: Live demo
Github Link : Repository Link

Test instances:

1. Stop being so stupid and actually read the article.
2. That's the dumbest thing I've read all day.
3. You're a worthless piece of garbage and everyone knows it.

4. You should just die already, nobody would miss you anyway.
5. You're trash and your opinions are worthless garbage.
6.This is @#$% ridiculous and you know it.
7.Could you please clarify what you meant by that?

Key Takeaways

- DistilBERT improves macro F1 by +25%
- Best gains on rare, context-dependent labels
- Demonstrates value of transformers for toxicity detection

Limitations:

- Context sensitivity still imperfect (FP on positive profanity)
- Measurable bias against certain identity mentions
- Requires human oversight for fair deployment

Future Work:

- Adversarial debiasing for fairness
- Context classifier for profanity
- Multilingual support & explainable predictions

Thank you