College of Science

Hate Speech & Offensive Language Detection Using BERT Transformers

Jose Santiago Campa Morales and Claire Lynch

CSC 396

Motivation

- Detecting harmful speech can help minimize marginalized online groups and is essential for community safety & health

- Context, Intent, Demographics, Ambiguity… NOT EASY!

Warning: some of these contain text that may be sensitive to some readers

Hate Speech

"what's this [ch****] email? I'm moving to China and slicing his throat"

Offensive Language

"Wake your [a**] up [h**]"

Neither

"People who slam on the brakes at yellow lights should not be allowed to drive."

Model Explanation

Transformers.

- Better at understanding context (self-attention), not just keywords
- Pre-training (BERT)
- Reduce false alerts when offensive words appear in jokes or quotes
- Capture subtle tone differences (sarcasm, aggression, harassment)
- BERT has strong generalization since trained on large corpora and has good performance on short texts

Model Explanation

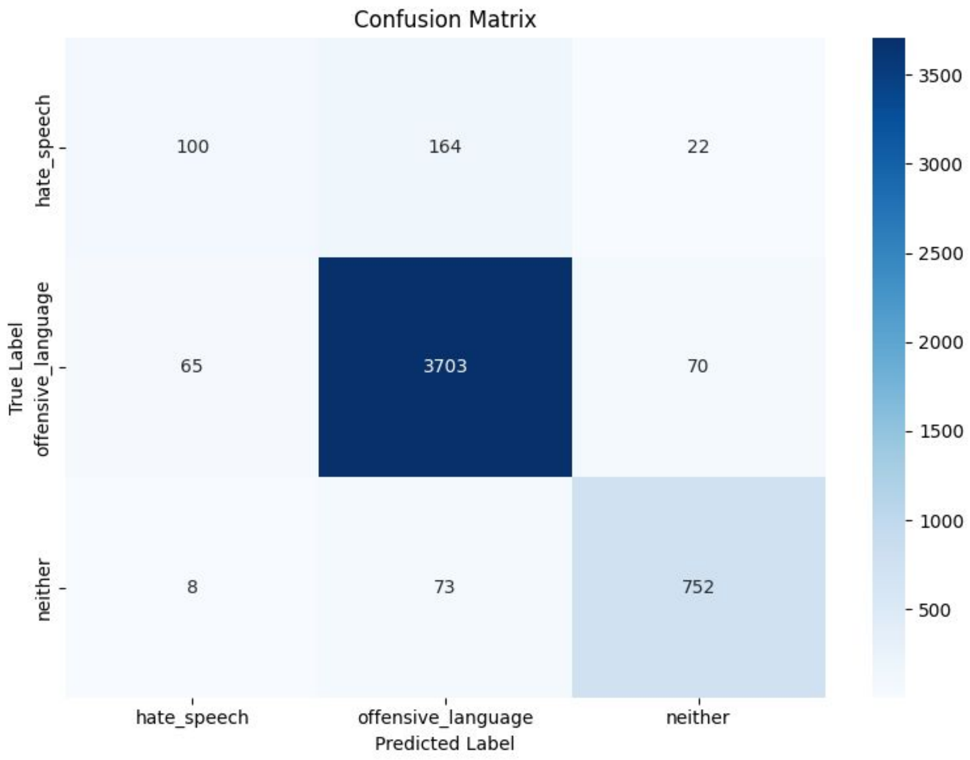
Dataset.

- Hate Speech and Offensive Language Dataset. (Davidson et al., 2017).
- Data split: 77% Offensive Language, 6% Hate Speech, 17% Neither
- Data imbalance.
- Very nuanced language implications and use of characters in posts

Why Multiclass Classification is Not Viable

- **Goal:** classify data into hate speech, offensive language, or neither
- **Outcome:** transformer slightly overfit + very imbalanced data

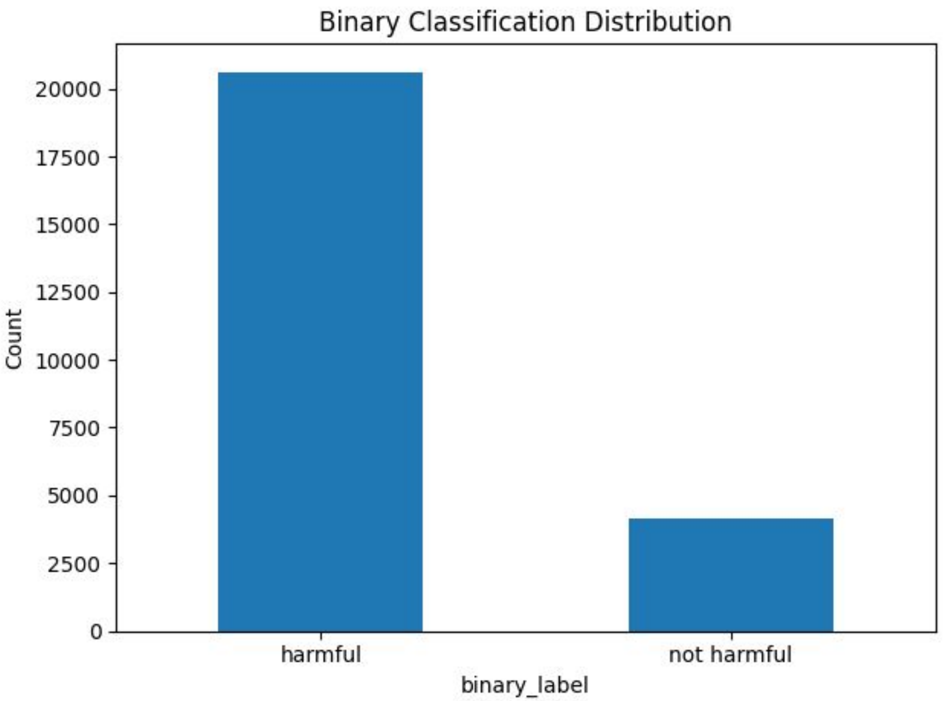
	precision	recall	f1-score
hate_speech	0.58	0.35	0.44
offensive_language	0.94	0.96	0.95
neither	0.89	0.90	0.90



Solution – Binary Classification

Harmful Speech vs.
Not Harmful Speech.

- Offensive Language & Hate Speech
- 83/17 split
- Predictable
- Applicable
- Removes ambiguity



Summary – Binary Classification

Epoch	Training Loss	Validation Loss	Accuracy
1	0.087	0.093	0.966
2	0.056	0.119	0.968
3	0.044	0.156	0.964

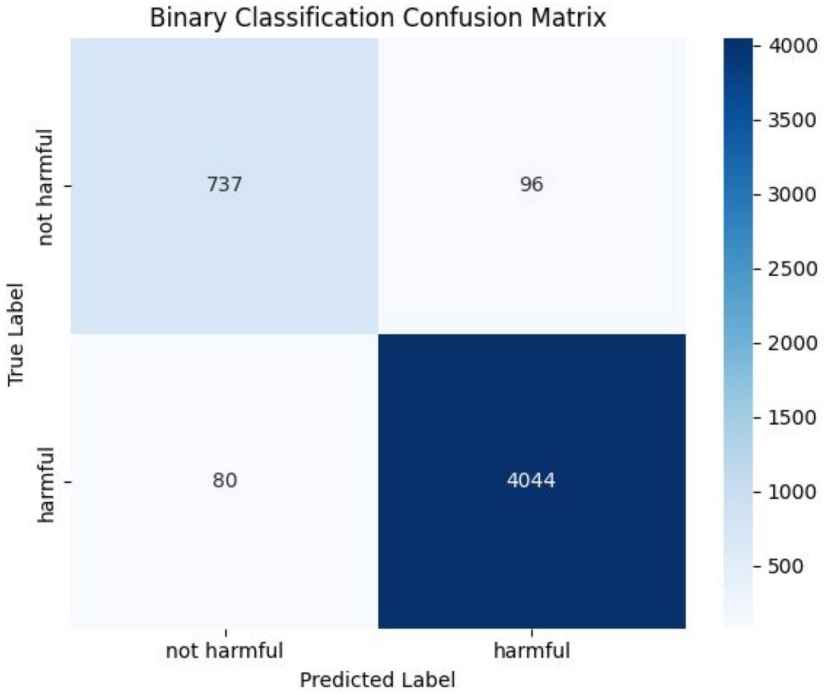
Results – Binary Classification

	Precision	**Recall**	**F1**	**Support**
Not Harmful	0.90	0.88	0.89	833
Harmful	0.98	0.98	0.98	4124
Accuracy			0.96	4957
Macro Avg	0.94	0.93	0.94	4957
Weighted Avg	0.96	0.96	0.96	4957

Binary Confusion Matrix

	precision	recall
non-harmful	0.90	0.88
harmful	0.98	0.98

- More false positives
  - (Predicts safe tweets as harmful)
- Less false negatives
  - (Predicts harmful tweets as safe)



Baselines

Random Classifier

- Predicts labels at random
- ~50% accuracy (binary)
- F1 ≈ 0.5

- Demonstrates how hard the dataset is without learning

Popularity Classifier

- Predicts all as harmful
- 83% accuracy
- F1 (harmful) = 1
- F1 (not harmful) = 0

- Misses every safe tweet → useless for real moderation

Error Analysis

False Negatives – *80 cases*
(Predicts harmful tweets as safe)

- “The Steelers new uniforms look like homosexual bumblebees. #wtf?”
- “I said something to you gorilla, stop listening to monkey music”

- Aggression used without slurs
- Harmful intent was implied or subtle: language is specifically negative

False Positives – *96 cases*
(Predicts safe tweets as harmful)

- “i'd never send my kids to private school because then they would never experience ghetto public school fights”
- “Taco Bell is super trash when you're sober”

- Patterns for posts containing profanity used in non-hostile context or jokingly
- Posts quoting lyrics or offensive memes

Warning: some of these contain text that may be sensitive to some readers

Observations, Limitations & Biases

- Over-reliant on keywords → False Positives
- Class Imbalance → False Negatives

- Length & subtlety bias
- Slang Misclassification & social bias

Insights, Impact & Conclusion

- Transformers are highly effective for harmful-language detection
- Multiclass labeling suffers from data imbalance

- **Real-world impact:** Small improvements scale to millions of users
- **Not a simple task:** requires careful, ethical deployment

Self-Evaluation

- 96% accuracy
- Relatively small project leads to meaningful real-world impact
- AI can support safer digital spaces
- Strong technical + ethical awareness

Future Work

- Larger, more balanced dataset
- Specialized model (HateBERT)
- Thread-level context aware detection
- Better fairness analysis

Contributions

- Claire
  - Project Idea
  - Dataset Search
  - Model Implementation
  - Baselines
  - Presentation
  - Report
- Jose
  - Dataset Search
  - Data Cleaning
  - Model Implementation
  - Binary Solution
  - Presentation
  - Report

Thank you! Any Questions?