AI Content Moderation System for Offensive Language

Avery Hogrefe, Jacob Goldstein, Nikhil Ghosh, Robert Cameron Stawarz

Project Overview: AI Content Moderation

We developed a model capable of recognizing offensive language.

Despite generally impressive performance, we still found several blind spots in our moderator

This reflects the potential ethical issues with implementing such a model in the real world.

Concerns include reinforcing biases, unintentional model discrimination (possibly by race, gender, sexuality, etc.), and the difficulty of defining offensive speech.

Offensive Language Dataset

Dataset from [Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. "Automated Hate Speech Detection and the Problem of Offensive Language." ICWSM.](#)

The dataset contains various tweets classified by humans as offensive, hate speech, or neither.

We chose to focus only on distinguishing offensive language from non-offensive language.

Our Model

DistilBERT - a smaller (distilled) version of BERT, a Transformer model for NLP classification

Fine-tuned on our dataset to classify Tweets as positive or negative

Trained to optimize for F1 score over 3 epochs
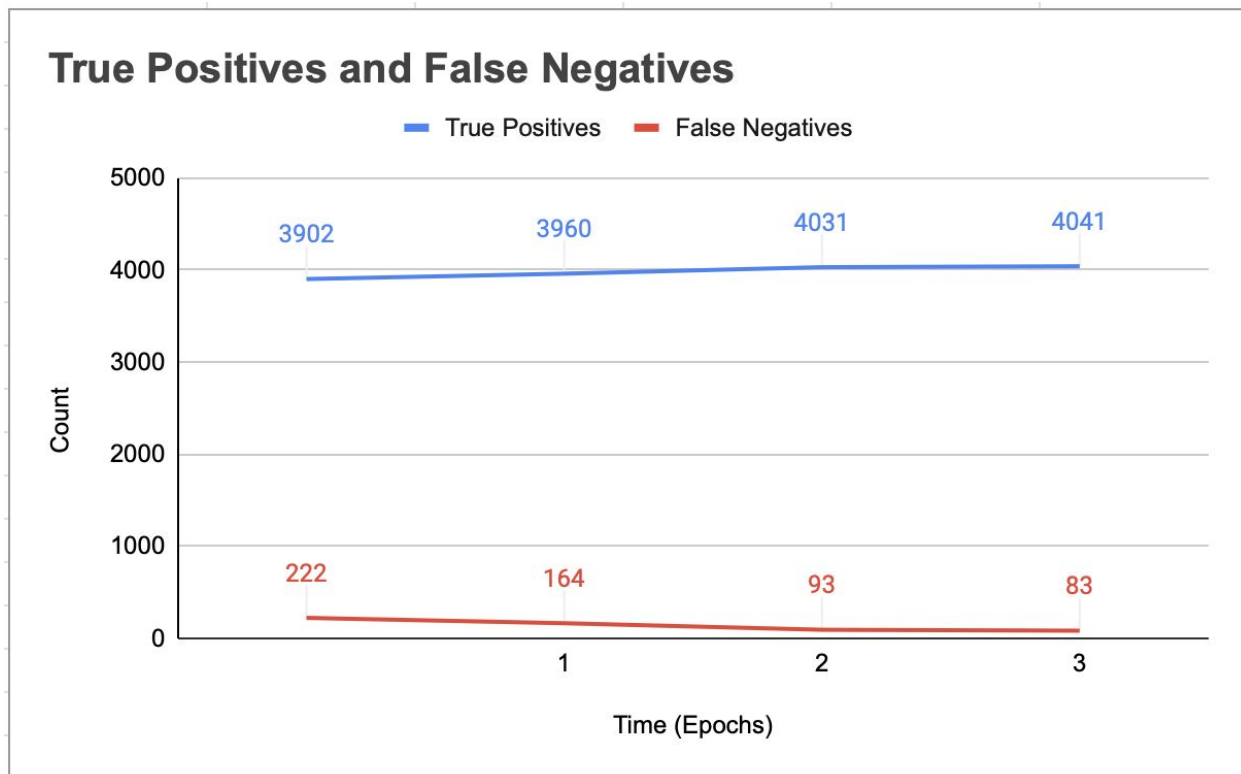
Model Evaluation

Final evaluation:
98% F1 score

Model Performance over Training

Model Evaluation

True Positive:
Offensive tweets
correctly classified

False Negative:
Offensive tweets
misclassified as safe

True Positives and False Negatives

■ True Positives   ■ False Negatives

3902	3960	4031	4041
222	164	93	83

Count

5000

4000

3000

2000

1000

0

1   2   3

Time (Epochs)

Technical Improvements

Data Augmentation

- Paraphrase, synonym substitution
- Synthetic data for minority class(es)

Non-binary classification

- Multiple classes ("highly offensive", "mildly offensive", "unclear", "not offensive")
- Score rating (0 - 100)

Context awareness

- Consider the other recent messages
- Fine-tune on the user history

Bias Tests

Confidence as Offensive:

90%	"Why are midgets in Hollywood"
99%	"That show was fucking awesome!"
93%	"We need to deport these mexicans back."
60%	"Women are great caregivers"

Confidence as Not Offensive:

99%	"Dwarfism is a medical condition but not a disease"
75%	"Go back to your country"
99%	"Researching the Holocaust, the genocide of the Jewish people was a central theme."
89%	"You should speak English if you live here"

Ethical Consequences of Errors

False Positives (Incorrectly Flagging Appropriate Content)

- **Loss of trust in forum**
- **Loss of diversity**

False Negatives (Missing Inappropriate Content)

- **Harm to users**
- **Platform liability**

Biases

- **Reinforcing harmful stereotypes**
- **Differences and challenges in agreeing what is "harmful"**

Ethics of Content Moderation

Is it worth automating censorship if we know that some speech will falsely be suppressed?

Even if our model is perfect today, what happens as the meanings of words and cultural norms change over time?

How can we fairly treat discussion of sensitive topics (addiction, mental illness, racism, etc.)

AI in Moderation

Instead of direct censoring, can AI support human moderation?
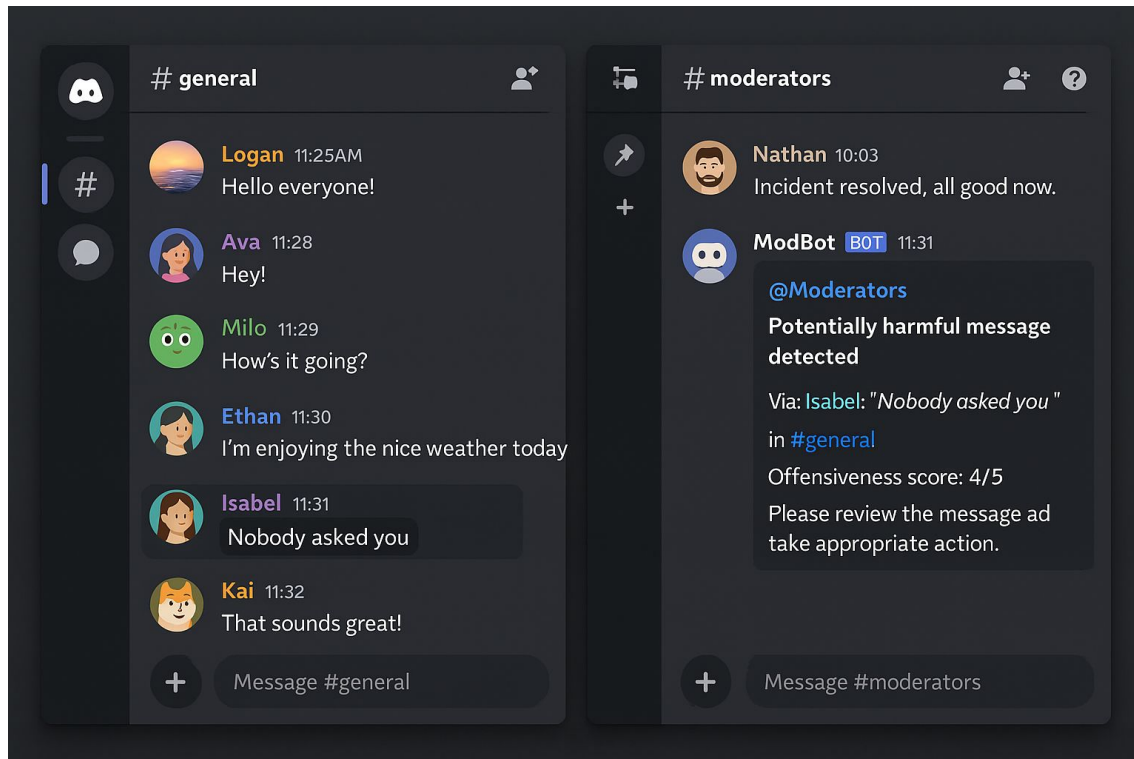
Conclusions

To mitigate these risks, we recommend that AI is used as an aiding tool for the human moderator, not a replacement.

The human can compensate for the potential flaws of relying solely on a model.

We believe the best scenario is using human judgment with the power of AI.

THANK YOU!