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 bias
- Unwarranted censorship
- Unintentional model discrimination (possibly by race, gender, sexuality, etc.)
- The difficulty of defining offensive speech at all.

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 24,000 tweets classified by humans as offensive, hate speech, or neither.

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

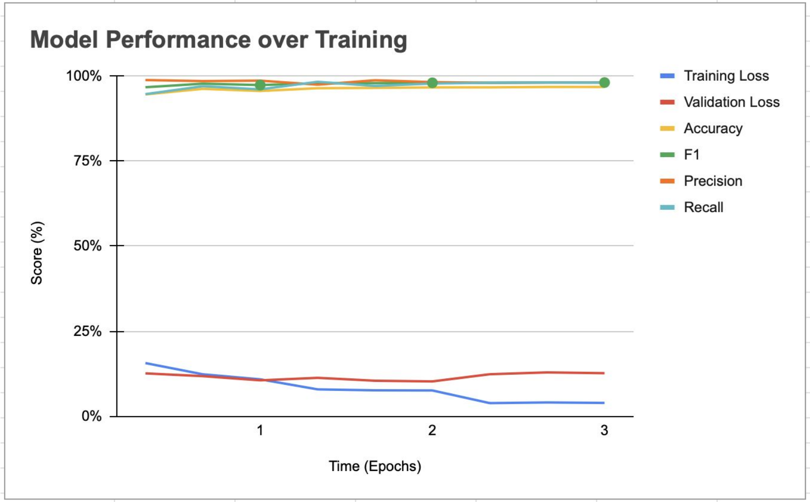
Our First Model

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

Chose F1 score as training metric

<u>After 3 epochs, 98% F1 score</u>

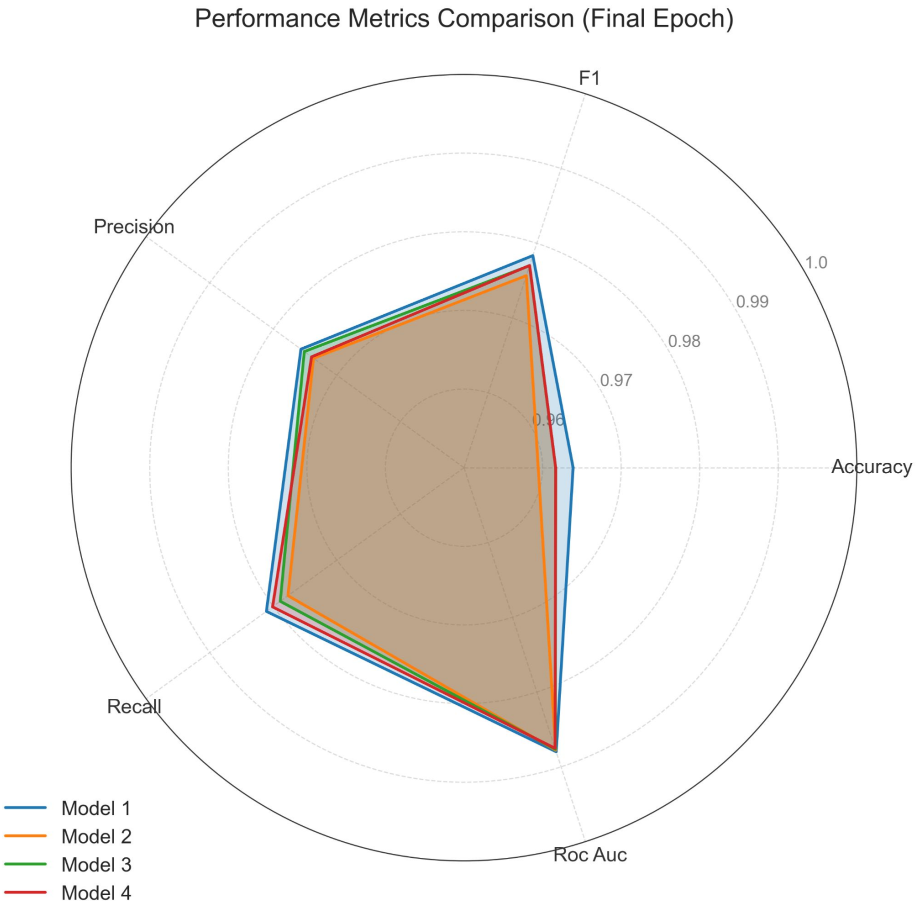


Second Run

4 Models

- Cross-validation
- Hyperparameter optimization
- Address class imbalance
- Error analysis
- Additional metrics

Insignificant differences:
scores 96.2 - 98.9%

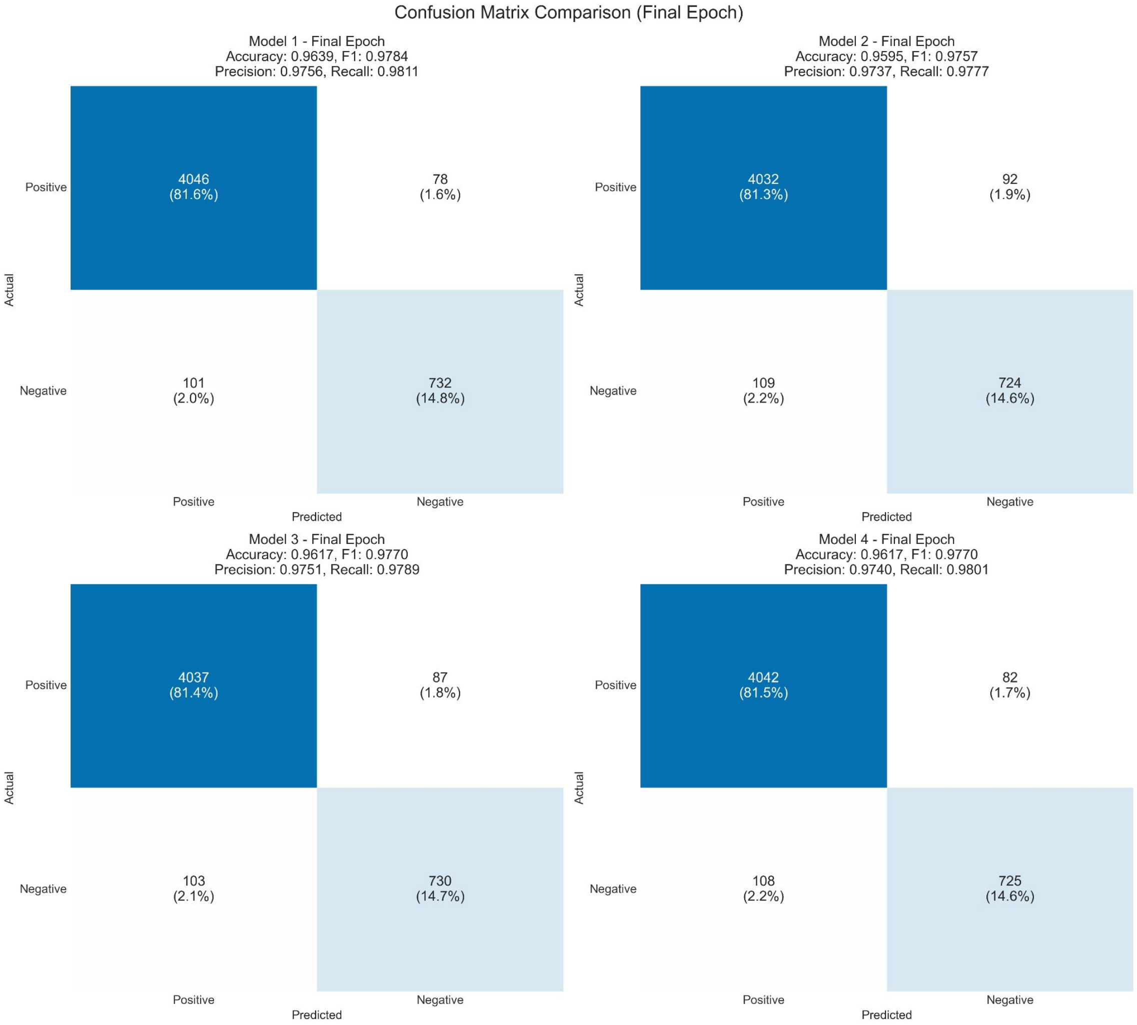


Performance Metrics Comparison (Final Epoch)

Second Run

Confusion Matrices

- Very similar



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 other recent messages, user history

Reinforcement

- Fine-tune from feedback after human corrections

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

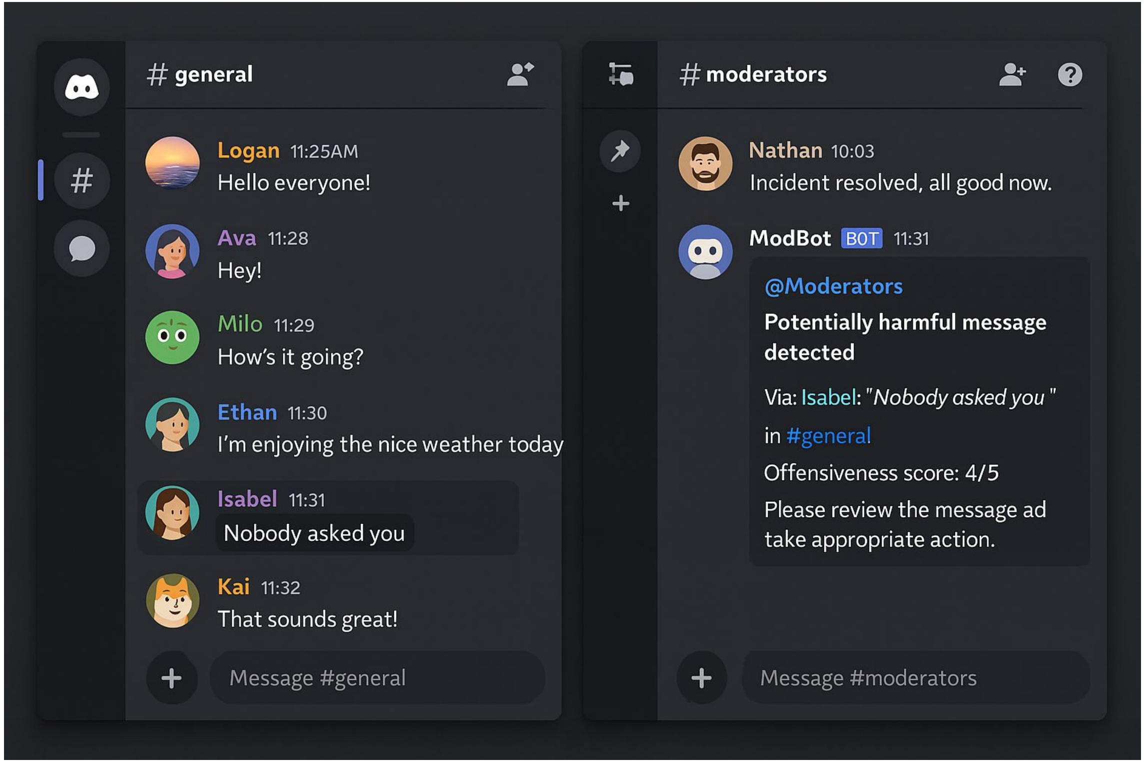
Is it worth automating censorship if we know that some speech will falsely be suppressed?

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?



general

Logan 11:25AM
Hello everyone!

Ava 11:28
Hey!

Milo 11:29
How's it going?

Ethan 11:30
I'm enjoying the nice weather today

Isabel 11:31
Nobody asked you

Kai 11:32
That sounds great!

moderators

Nathan 10:03
Incident resolved, all good now.

ModBot BOT 11:31

@Moderators

Potentially harmful message detected

Via: Isabel: "*Nobody asked you* "

in #general

Offensiveness score: 4/5

Please review the message ad take appropriate action.

Conclusions: What’s the Solution?

AI will not be perfect.

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!