# Progress Report:

# AI Content Moderation System for Offensive Language

Avery Hogrefe, Jacob Goldstein, Nikhil Ghosh, Robert Stawarz

# Project Overview and Goals

Our Project: An AI Content Moderation System for Offensive Language

Goals:

- Develop a model capable of recognizing offensive language.
- Study our model to discover any potential biases (such race, gender, sexuality, etc.)
- Use our findings to consider the ethical ramifications of implementing such a model in the real world.

# Offensive Language Dataset

Our model uses data 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 used the frequency of how often each tweet was flagged as offensive or hateful to assign it a score.

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

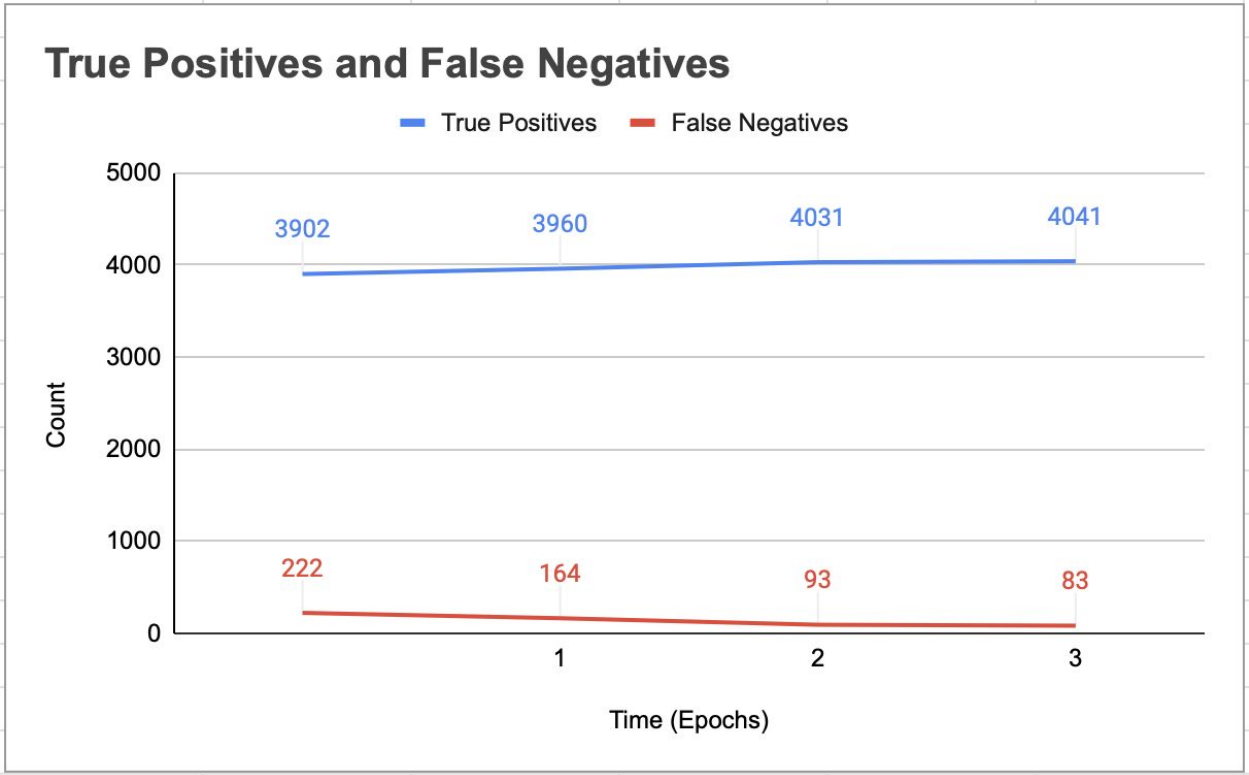
# Model Evaluation

Final evaluation:
98% F1 score

# Model Evaluation

True Positive:
Offensive tweets correctly classified

False Negative:
Offensive tweets misclassified as safe



# 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”

# Plan for Completion

1. Gather more data on the types of errors the model makes.
2. Discern any patterns to the model's biases.
3. Consider methods to improve the model and/or data.
4. Use our findings to develop a deeper ethical reflection on AI Content Moderation.
5. Use all of this to complete the final presentation and technical report.

# THANK YOU!