



Social Beings™:

A New Social Networking Platform
Moderated by Generative AI

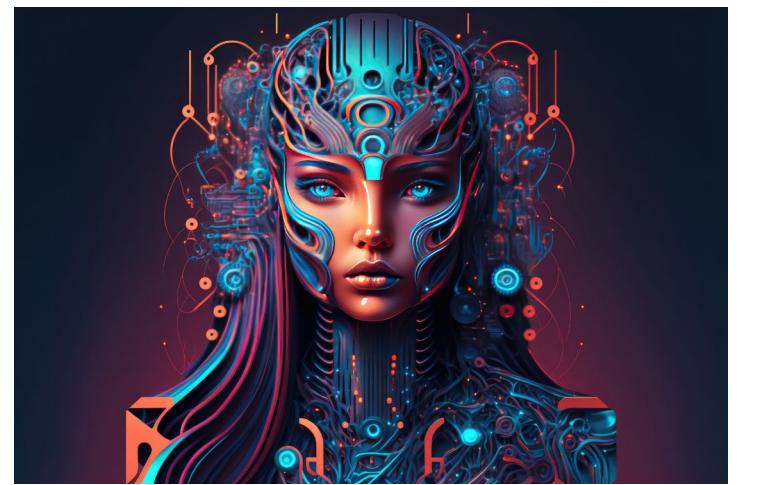
Detecting and Mitigating Implicit and Explicit Bias on Social Networks Using ML and Generative AI

Celia Banks

cbbanks@umich.edu

Submitted to the University of Michigan's MADS Capstone for Fulfillment of Masters of Applied Data Science Degree

Project Mentors: Capstone: Laura Stagnaro, PhD | Milestone II: Jinseok Kim, PhD

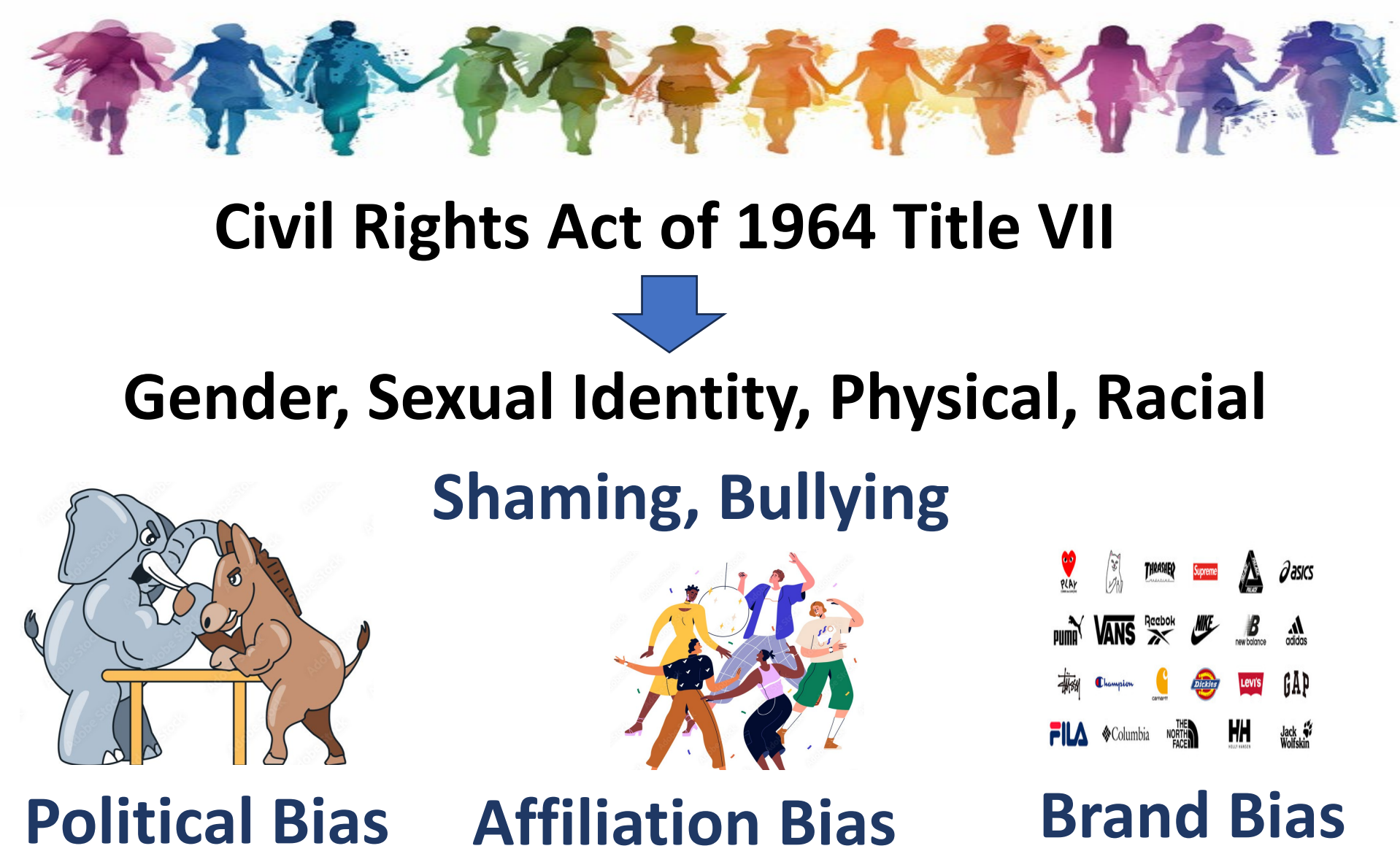


JenAI-Moderator™

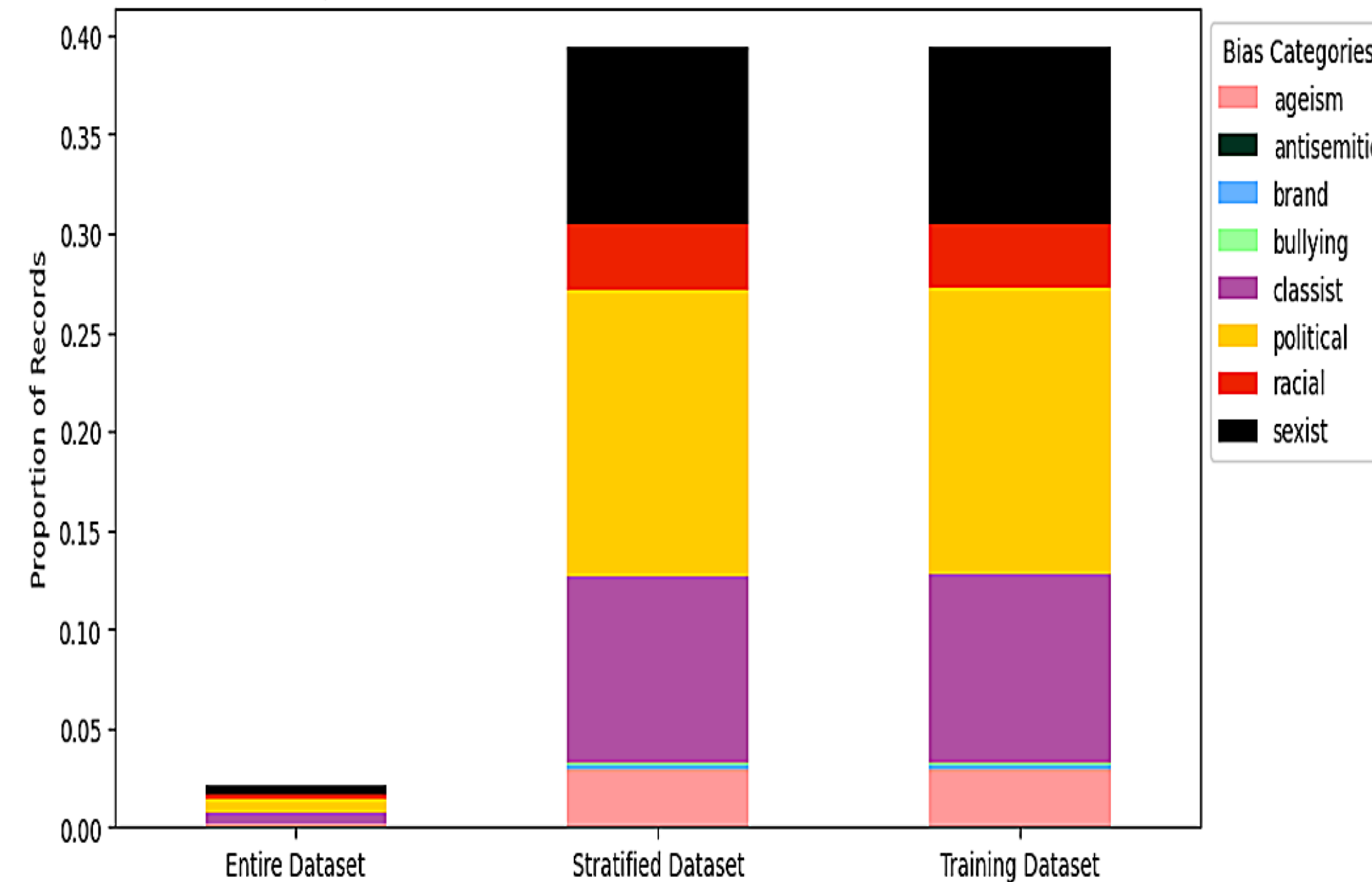
ABSTRACT

Social networks have introduced complex challenges in mitigating biases—implicit and explicit—embedded within digital interactions. Leveraging generative AI, this project employs a novel way to use machine learning (ML) to address these biases, aiming to tone down the negativity found within online platforms.

DIFFERENT TYPES OF BIAS



Bias in Entire v. Stratified v. Training Datasets

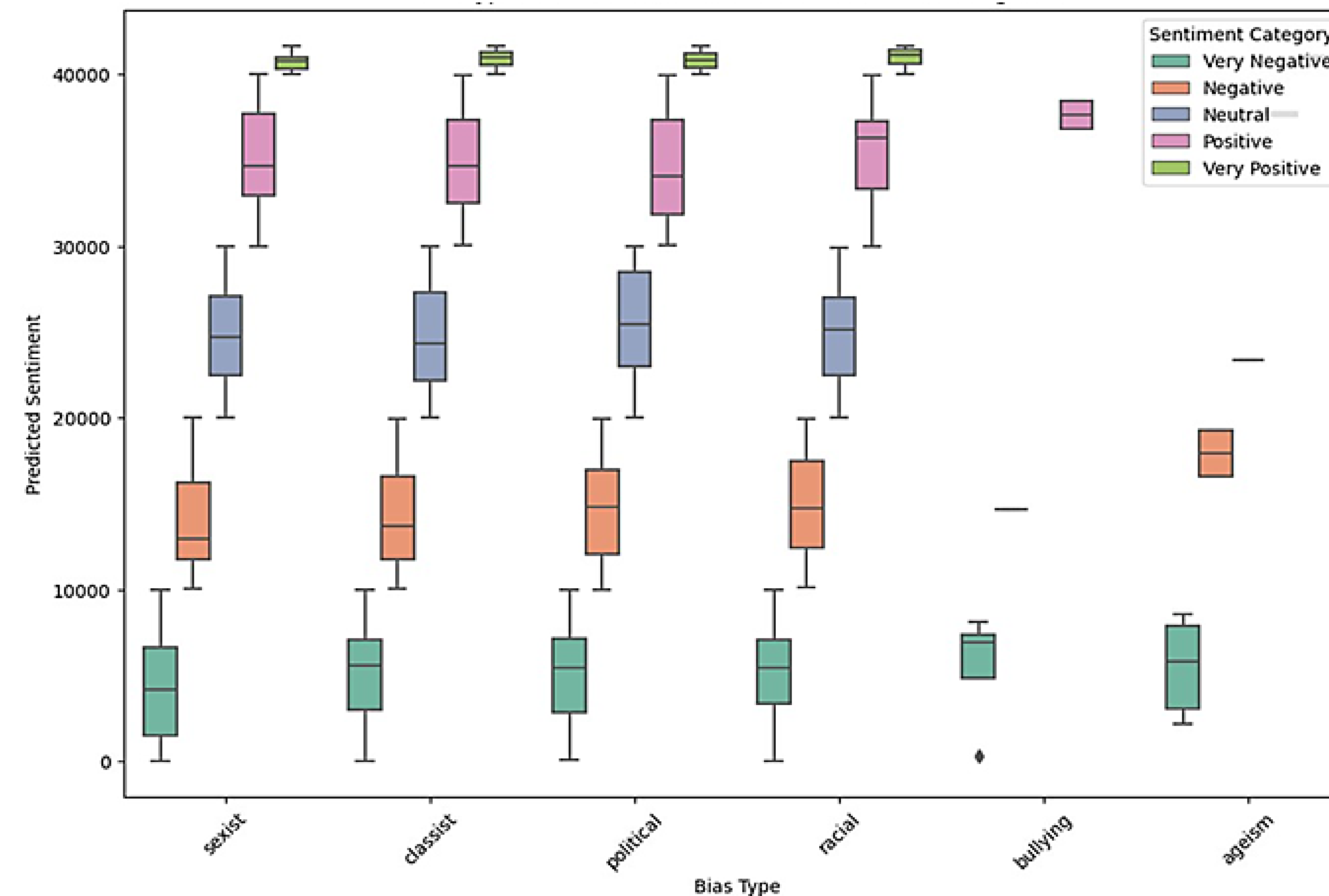


METHODS

- ❑ Leveraged deep learning and pre-trained models to classify, analyze, and address bias in U.S. English one million Twitter tweets and one million Reddit comments.
- ❑ Trained RoBERTa model for sentiment analysis to distinguish between implicit and explicit bias.
- ❑ Pre-trained DistilBERT as base model on dataset modified on RoBERTa sentiment classification.
- ❑ Used large language model to mitigate implicit and explicit bias in Twitter tweets and Reddit comments.

How can artificial intelligence (AI) assist in unlearning implicit bias in social network environments?

Bias Type Distribution with Predicted Sentiment Categories



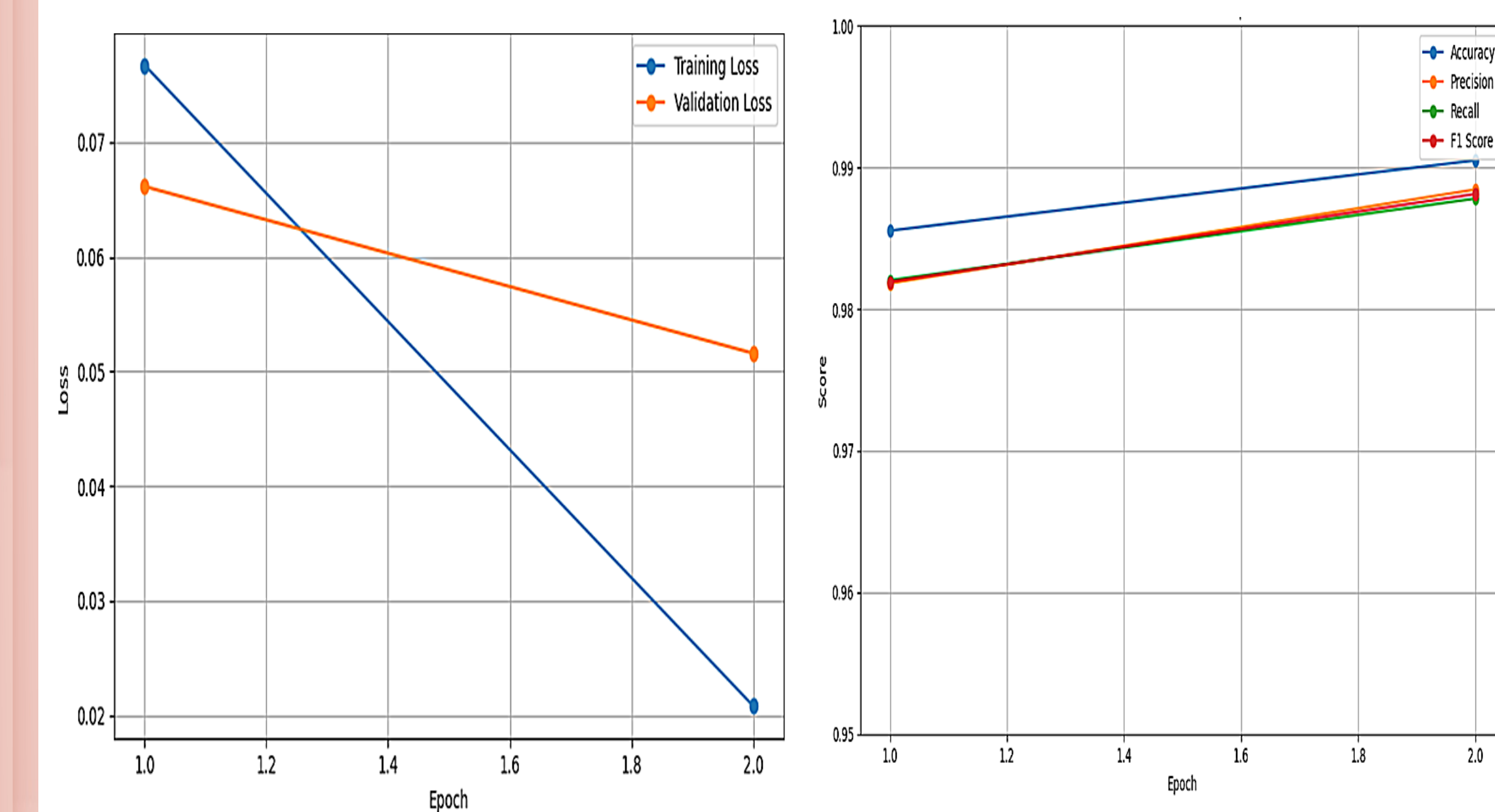
Sample of Mitigation Responses

Social Network	Bias Type	User Comment	JenAI-Moderator Responds
Twitter	Classist	White trash vs us... We were outnumbered. http://twitpic.com/2y3jp	It sounds like you're trying to unpack some complex language and intent in that comment. While the original statement does seem to convey a sense of division and perhaps derogatory labeling, an alternative phrasing aiming for neutrality could help reduce bias. For example, you might consider rephrasing it to something like, "We felt outnumbered in a challenging situation." This way, it avoids labeling any group negatively and focuses on the experience without implying judgment. \n\nIf you're considering the context more closely, it's always a good idea to think about how language can affect perceptions and relationships among people. Being mindful of the words we choose can really make a difference! Let me know if there's more you'd like to explore on this topic.
Twitter	sexist	@ [redacted] Oh, God, ow. That must have hurt like a bitch.	Thank you for sharing that comment! It seems like you're looking to identify potential bias and ensure that the tone remains neutral and respectful. \n\nThe original comment does express a strong emotional reaction, which could imply bias through the use of intense language. To adjust it for a more neutral tone, you could rephrase it to something like: "That sounds really painful. I hope you're doing okay." This way, you maintain empathy while avoiding any language that could be perceived as biased or overly emotional.\n\nIf you have any more examples or need further clarification, feel free to share! I'm here to help.
Reddit	racial	'migrant workers run on diesel?	Thank you for pointing that out! It's really important to be mindful of our language. To promote inclusion and clarity, we could rephrase the comment as: "There are concerns about the environmental impact of diesel use by certain sectors, including migrant workers." This way, we address the issue without implying bias or stereotypes. If you'd like more suggestions or further assistance, feel free to ask!

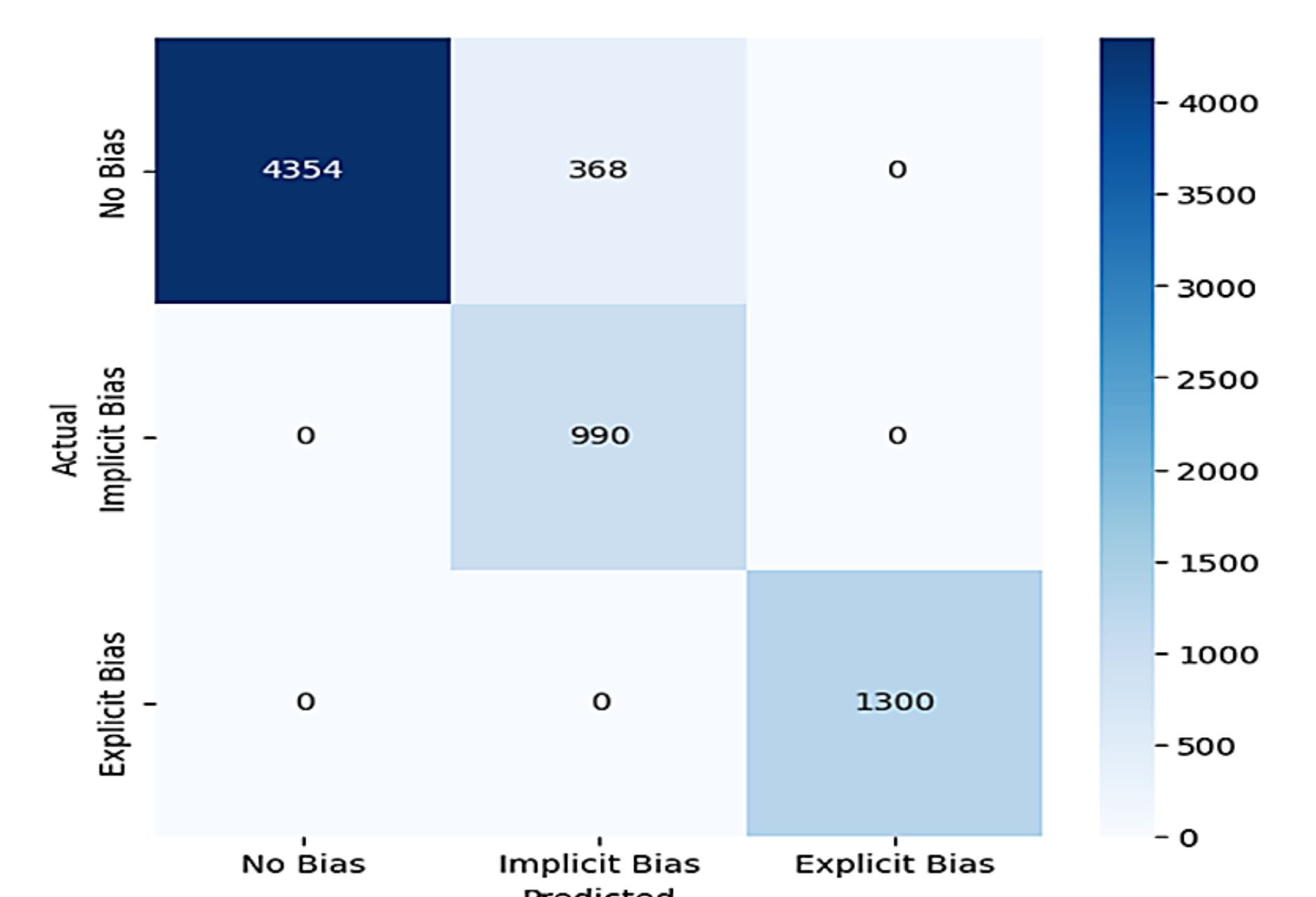
RESULTS

Produced a robust, generalizable model with .99 F1 on no bias and .99 F1 on has bias. Proved on testing with example comments.

DistilBERT Train v. Validate Loss DistilBERT Model Accuracy



RoBERTa Sentiment Confusion Matrix



Model Test with Example comments

- ❖ Text: Wetbacks are criminals and should not be allowed to immigrate to this country.
❑ Prediction: Bias Detected
- ❖ Text: I love learning about different perspectives in social media.
❑ Prediction: No Bias Detected
- ❖ Text: You are white privileged to think that way.
❑ Prediction: Bias Detected
- ❖ Text: I find that N----s are more critical of themselves than they are of other races especially white.
❑ Prediction: Bias Detected

HEAR THE PROJECT DISCUSSION

