

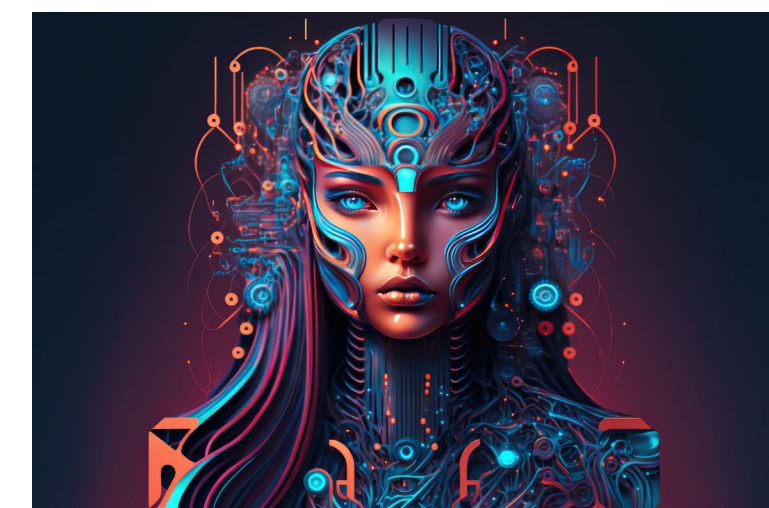


Social Beings™:

A New Social Networking Platform
Moderated by Generative AI

Using Generative AI to Mitigate Bias in Social Network Environments

Celia Banks



JenAI™

Submitted to the University of Michigan’s MADS Capstone for Fullillment of Masters of Applied Data Science Degree

ABSTRACT

The rapid growth of social networks has introduced complex challenges in mitigating biases—both implicit and explicit—embedded within digital interactions. Leveraging generative AI, this project explores strategies to address these biases, aiming to improve equity and inclusivity within online platforms. The study highlights how biases, rooted in human cognition and amplified by algorithmic decision-making, influence user experiences and perpetuate social inequality.

OBJECTIVE

Bias in social networks can emerge in various forms, from user-generated content to algorithmic recommendations. Rooted in structural inequalities, biases such as those related to race, gender, religion, bullying, and ageism often manifest as explicit or implicit actions. These issues and other biases like political representation, group affiliation, and brand bias are exacerbated by the design of generative AI systems, which can inadvertently perpetuate harmful stereotypes through flawed data selection, outdated patterns, and algorithmic reinforcement.

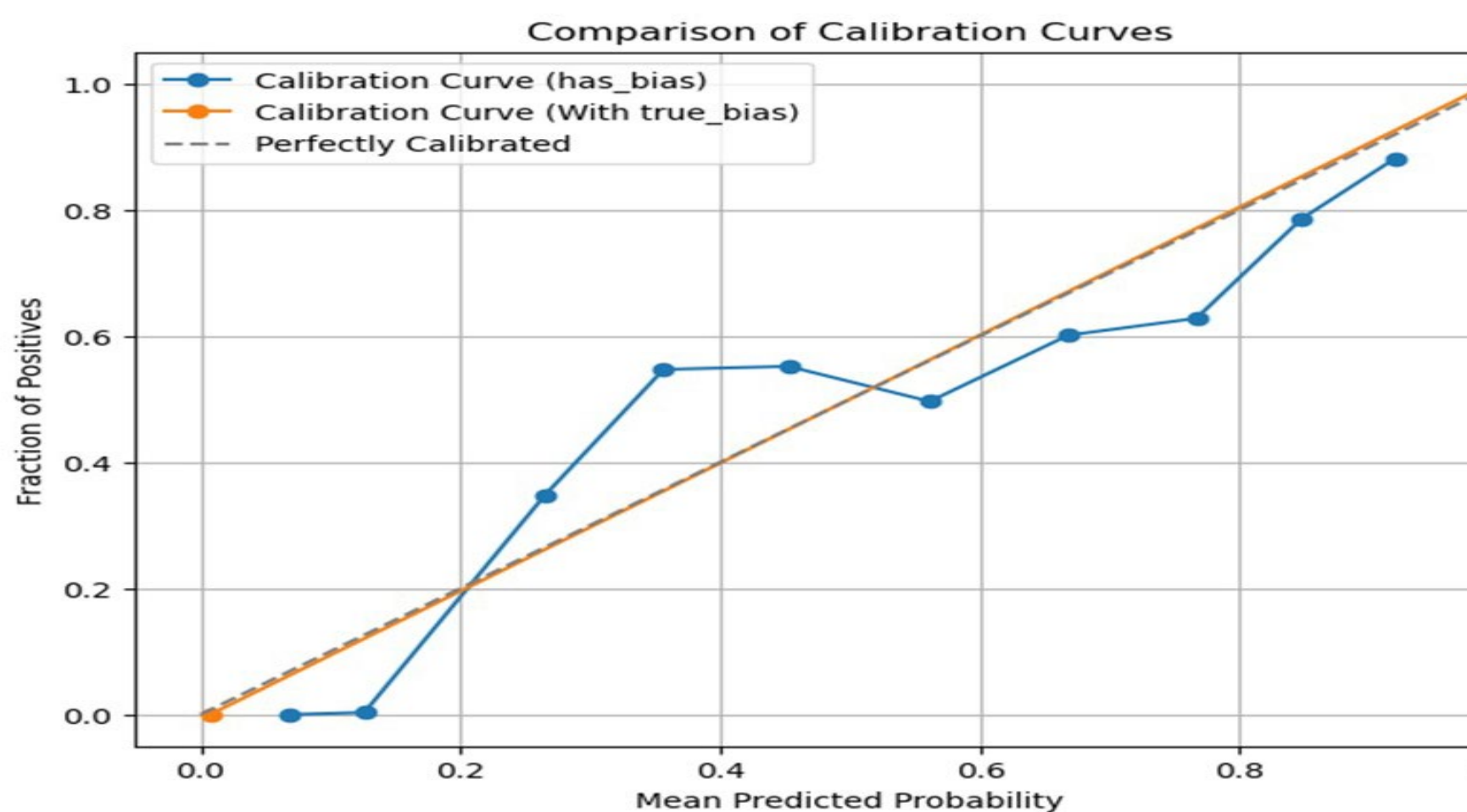
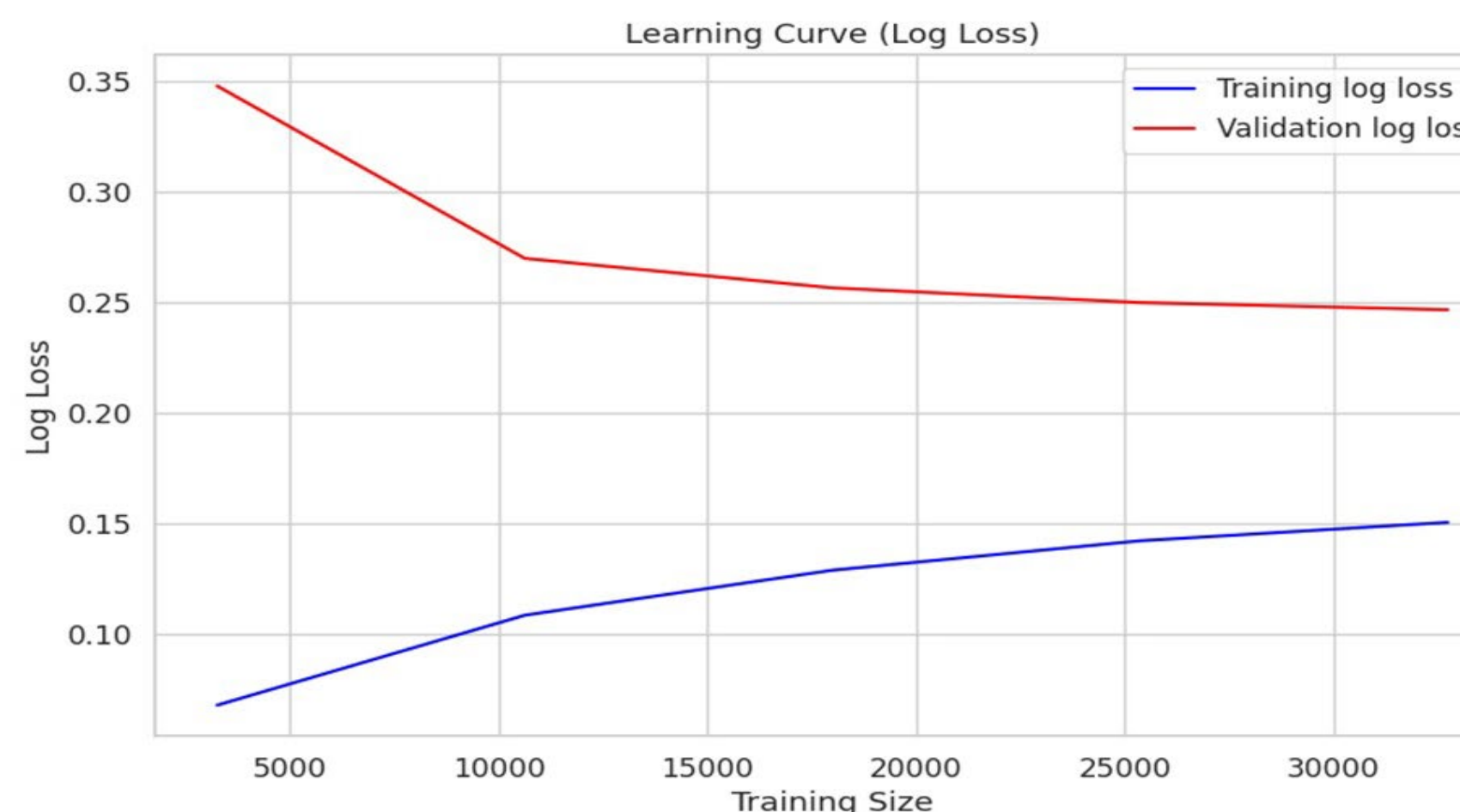
This project addresses the critical research question: *How can AI assist in unlearning bias in social network environments?* By developing mitigation strategies, the study aims to:

- 1.Detect and address algorithmic biases without introducing further inequities.
- 2.Provide users with tools to engage and learn from biased content constructively.
- 3.Minimize unwanted outputs, such as biased recommendations, that lead to user frustration or disinformation.

METHODS

We developed a bias detection and mitigation framework using a supervised learning model trained on 2 million combined U.S. English Twitter and Reddit comments from the Sentiment140 dataset and Watchful1’s Reddit repository (2009). Bias was labeled as implicit or explicit using templates of bias terms and safe phrases. False bias, such as benign use of terms like “grandma,” was minimized by refining the templates and tuning the XGBoost Classifier. The model achieved 99% accuracy.

Bias mitigation utilized a large language model (ChatGPT) to process comments, identify bias, and trigger action events: explicit bias invoked strong responses, while implicit bias generated suggestive replacements and user alerts to promote learning. This method avoids reliance on personal user data, addressing privacy concerns while outperforming traditional fairness algorithms that may inadvertently harm target groups.



METHODOLOGY DETAILS

The model was developed on the University of Michigan's Great Lakes cloud platform using a single-core processor, 64GB RAM, and 1 GPU.

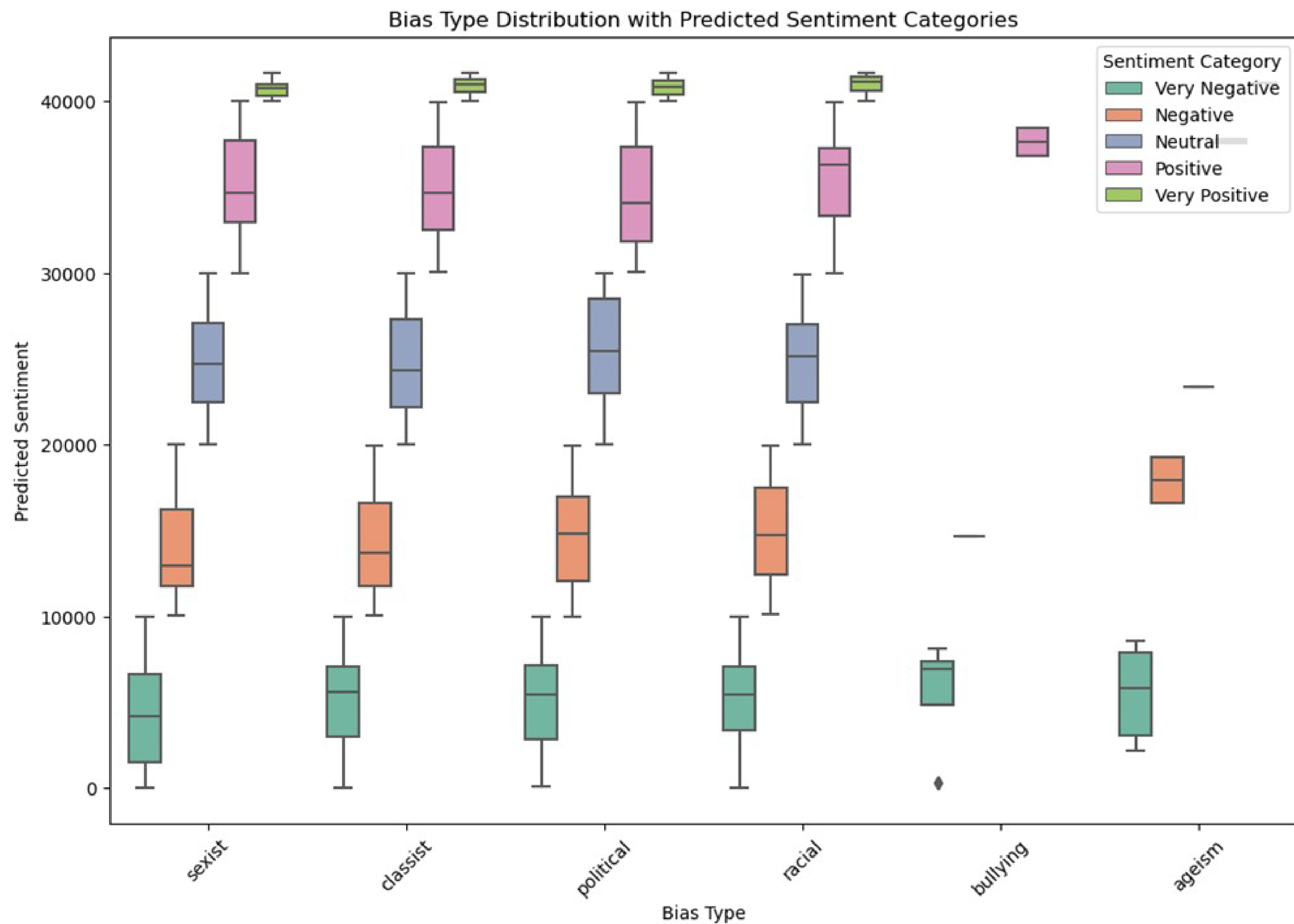
Data Processing: Tokenized and lemmatized text, removed extraneous characters, and created templates of bias terms and safe phrases. Data was stratified by bias presence and processed in batches for efficiency.

Bias Terms Association: Coded bias categories using label encoders, creating data visualizations like occurrence tables and heatmaps.

Bias Classification: Added columns for true_bias and implicit_explicit labels. Records with false bias terms were updated accordingly

Model Training: The XGBoost Classifier was trained and validated with train/test/validate splits. Key features included processed tokens and true_bias labels, yielding 99% accuracy.

Cross-Validation: A 5-fold cross-validation confirmed the absence of overfitting with an optimal 100 boosting rounds, ensuring robust real-world performance.



RESULTS

The ability to generalize a model for bias detection and mitigation is critical for consistent performance in a live social network environment. While this project did not test the model in a live environment, its design and outcomes demonstrate its potential to deliver effectively in such settings. Visualizing the model's capability to generalize to unseen data highlights its effectiveness.

Sample of Mitigation Responses

Social Network	Bias Type	User Comment	JenAI 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 labelling, 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!