Status Report CSCI544 - Group 19
Real-World Bias in AI: Assessing and Mitigating Stereotypes in Multimodal Models

Archana Sondur[1], Hitanshu Panchal[2], Nehal Garg[3], Pranav Padhye[4], Vishwas Parekh[5]
University of Southern California, USA
[1]sondur@usc.edu, [2]hpanchal@usc.edu, [3]nehalgar@usc.edu,
[4]ppadhye@usc.edu, [5]vdparekh@usc.edu

1 Tasks Performed

1.1 Dataset Construction

We curated a dataset of 2,500 image-text pairs from platforms like Twitter, Reddit, news sites (e.g., CNN, BBC), and Unsplash to reflect diverse real-world content. Each sample was tagged with one of 20+ demographic labels spanning race, religion, nationality, gender identity, and disability. To improve subgroup granularity, we disaggregated "LGBTQ" into Gay, Lesbian, Transgender, and Others (e.g., Bisexual, Nonbinary), and added underrepresented nationalities (e.g., Indian, French, German) using targeted keyword filtering. All data was stored in structured JSON format with metadata for downstream analysis.

1.2 Manual Data Pre-processing and Curation

Given the noisy and ambiguous nature of real-world data, we employed a multi-stage manual filtering process to ensure quality and relevance. Each image was reviewed to confirm accurate demographic alignment, contextual clarity, and semantic neutrality. Visual artifacts such as excessive overlays, political watermarks, memes, or blurred content were excluded. To preserve fairness, we maintained approximate parity in sample size across subgroups. This ensures that no particular group dominates the model's learning during evaluation. Our human-in-the-loop review process, while time-intensive, was crucial for maintaining ethical standards and dataset reliability.

1.3 Bias Evaluation and Results

We evaluated bias using OpenAI's CLIP (ViT-B/32), which projects images and text into a shared embedding space via contrastive learning. Cosine similarity was computed between image embeddings and evaluative prompts (e.g., "This is a kind person" vs. "This is a dangerous person")

Our analysis included:

- **Top-k Attribution:** Extracted the 15 most associated attributes per subgroup, color-coded by sentiment for qualitative insights.
- **Pairwise Bias Scoring:** Computed Caliskan-style effect sizes to quantify bias between subgroup pairs.

Preliminary results showed:

Group	Subgroup_A	Subgroup_B	Bias_Score	Favored_Group
Disability	Mentally Disabled	Physically Disabled	-0.483158	Physically Disabled
Disability	Non-Disabled	Physically Disabled	1.256323	Non-Disabled
Disability	Non-Disabled	Mentally Disabled	1.374975	Non-Disabled
Nationality	Indian	Mexican	0.326241	Indian
Nationality	Indian	European	0.476713	Indian
Nationality	Indian	Arab	-0.115919	Arab
Nationality	European	Americans	-0.065231	Americans
Nationality	European	Chinese	0.097084	European
Nationality	European	Arab	-0.581694	Arab
Nationality	European	Mexican	-0.158761	Mexican
Nationality	Americans	Arab	-0.541179	Arab
Nationality	Indian	Americans	0.431759	Indian
Nationality	Americans	Mexican	-0.099504	Mexican
Nationality	Chinese	Arab	-0.662366	Arab
Nationality	Chinese	Mexican	-0.252563	Mexican
Nationality	Arab	Mexican	0.437058	Arab
Nationality	Americans	Chinese	0.164334	Americans
Nationality	Indian	Chinese	0.561675	Indian
Religion	Christians	Muslims	0.379974	Christians
Religion	Muslims	Hindu	-0.893823	Hindu
Religion	Jewish	Buddists	-0.690173	Buddists
Religion	Jewish	Sikhs	-0.583477	Sikhs
Religion	Hindu	Buddists	0.551409	Hindu
Religion	Hindu	Sikhs	0.518982	Hindu
Religion	Hindu	Jewish	1.114357	Hindu
Religion	Muslims	Buddists	-0.434891	Buddists
Religion	Muslims	Sikhs	-0.353555	Sikhs
Religion	Muslims	Jewish	0.246747	Muslims
Religion	Sikhs	Buddists	-0.037952	Buddists
Religion	Christians	Buddists	-0.079685	Buddists
Religion	Christians	Sikhs	-0.030233	Sikhs
Religion	Christians	Jewish	0.648318	Christians
Religion	Christians	Hindu	-0.648757	Hindu
Sexual Orientation	Gay	Heterosexual	-0.800475	Heterosexual
Sexual Orientation	Lesbian	Transgender	0.525688	Lesbian
Sexual Orientation	Gay	Transgender	0.313845	Gay
Sexual Orientation	Lesbian	Gay	0.228546	Lesbian
Sexual Orientation	Lesbian	Heterosexual	-0.628718	Heterosexual
Sexual Orientation	Heterosexual	Transgender	1.001290	Heterosexual

Figure 1: Pairwise bias scores computed using a Caliskan-style effect size. Positive values indicate stronger association with pleasant attributes for Subgroup A; negative values favor Subgroup B.

All experiments were implemented in a modular Python pipeline (`BiasDetectionCustom.ipynb`)

designed for reuse and future extension to models like ALBEF and ViLT.

2 Risks and Challenges

2.1 Methodological Fidelity

MMBias was curated with a focus on uniformity, whereas our real-world dataset encompasses a broader range of visual and linguistic variation. Even subtle differences in factors like lighting, tone, or text structure can significantly impact model behavior. Therefore, maintaining strict alignment with scoring formulas is essential to ensure consistent and meaningful comparisons. Furthermore, attention to these nuances is critical for accurately assessing and mitigating biases, ensuring the model's robustness across diverse real-world scenarios.

2.2 Noise and Ambiguity

Social media content often contains sarcasm, cultural symbols, and memes that Vision-Language Models (VLMs) may misinterpret. These elements can blur the distinction between model bias and dataset artifacts. As a result, it is crucial to minimize noise as much as possible.

2.3 Cross-Dataset Comparability

Unlike MMBias, sourcing representative real-world images for certain minority groups remains a significant challenge. To tackle this, we plan to use stratified sampling and normalization techniques to ensure a valid and fair comparison across different datasets. Additionally, we will continuously monitor and update our data collection methods to better reflect the diversity of real-world populations.

3 Mitigations and Next Steps

3.1 Scoring Alignment

We follow the Caliskan effect size formula similar to the one used for MMBias, reproducing target-attribute sets and normalizing cosine scores to ensure fair benchmarking.

3.2 Cleaning Pipeline

We used OpenCV, DLib, and manual filtering to remove noisy, ambiguous, or harmful samples. Each identity group was curated with attention to cultural and ethical representation.

3.3 Future Work

- Evaluate our pipeline on the original MMBias dataset for validation.
- Perform comparative analysis across datasets using identical model setups.
- Explore mitigation techniques such as counterfactual augmentation and adversarial fine-tuning.

4 Individual Tasks

Member Name	**Contribution**
Archana Sondur	• Developed coding script for Bias Detection. • Assisted with CLIP implementation. • Validated and analyzed results.
Hitanshu Panchal	• Researched and documented image collection APIs for CNN news, BBC news & Unsplash in Postman. • Implemented API for Unsplash & CNN news. • Generated dataset for Religion.
Nehal Garg	• Generated dataset for Nationality. • Cleaned images from API output.
Pranav Padhye	• Researched and documented API for Twitter & Reddit in Postman. • Generated dataset for Sexual Orientation and Disability. • Configured Reddit API.
Vishwas Parekh	• Scripted bias detection logic. • Ran CLIP model bias detection on MMBias and our dataset.

Table 1: Individual contributions of team members.

References

[1] Sepehr Janghorbani and Gerard De Melo. "*Multi-Modal Bias: Introducing a Framework for Stereotypical Bias Assessment beyond Gender and Race in Vision–Language Models.*" EACL 2023.

[2] Caliskan, Bryson, and Narayanan. "*Semantics derived automatically from language corpora contain human-like biases.*" Science, 2017.

[3] Guo and Caliskan. "*Detecting emergent intersectional biases...*" AAAI/ACM Conference on AI, Ethics, and Society, 2021.

[4] Moin Nadeem et al. "*StereoSet: Measuring stereotypical bias...*" ACL-IJCNLP 2021.

[5] Ben Hutchinson et al. "*Social biases in NLP models...*" arXiv:2005.00813.