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

March 2025

Abstract

Bias in multimodal models, particularly Vision-Language Models (VLMs), extends beyond gender and racial biases to include religion, nationality, disability, and sexual orientation. While the MMBias dataset provides a benchmark for bias assessment, it has limitations, such as restricted subgroup representation and data imbalance. This study addresses these gaps through two key investigations. First, we analyze VLM bias on unfiltered social media and news content, comparing patterns with MMBias to assess whether models reflect or distort societal biases in dynamic media environments. Second, we propose a Fairness-Augmented Dataset (FAD) to mitigate bias by broadening demographic representation and incorporating counter-stereotypical associations. Additionally, we explore rephrasing strategies to neutralize biased language in textual data. By integrating real-world bias evaluation with fairness-aware fine-tuning, this research contributes to developing more equitable AI systems with implications for AI ethics, content moderation, and automated decision-making.

1 Introduction

Vision-Language Models (VLMs) such as CLIP, ALBEF, and ViLT have become fundamental in AI-driven applications, including image search, content moderation, and assistive technologies. While these models demonstrate impressive multimodal understanding, they also inherit and amplify biases present in their training data. Previous research has extensively documented gender and racial biases in VLMs, yet biases related to religion, nationality, sexual orientation, and disabilities remain significantly underexplored. Given the growing influence of AI-generated media on public discourse and decision-making, addressing these biases is crucial to ensuring fairness and equity in AI systems.

Despite progress in bias evaluation, existing studies primarily rely on controlled benchmark datasets that fail to capture how biases manifest in real-world digital environments such as social media and news platforms. These sources actively shape public narratives, influence opinions, and reinforce stereotypes, making them critical yet understudied domains for bias assessment. However, the extent to which VLMs align with or distort biases present in such dynamic media landscapes remains largely unknown. Moreover, most bias mitigation strategies focus on post-processing model outputs rather than addressing biases embedded in learned representations. This approach is limited in scalability, highly dataset-dependent, and insufficient to prevent bias propagation into downstream applications.

To address these gaps, our research extends the MMBias framework to evaluate biases in real-world images and captions sourced from social media (Twitter, Instagram, Facebook) and news platforms (CNN, BBC, Fox News).

2 Literature Review

Bias in artificial intelligence systems has been widely studied, particularly in language models, where most research focuses on gender and racial biases [2][3]. However, certain studies have expanded this focus to biases concerning aspects like profession, religion, and disability[4][5]. These biases are not confined to textual data alone, as image classification systems and multimodal models have also exhibited similar patterns of bias.

Thus far, research on multimodal bias assessment in self-supervised models has been largely limited to considerations of gender and racial biases. In [1], the authors address the limitations by introducing MMBias, a benchmark designed to assess biases across a broader spectrum of social categories, including religion, nationality, disability, and sexual orientation. Their study evaluates bias in several prominent multimodal vision–language models such as CLIP, ALBEF, and ViLT. Using a dataset of approximately 3,800 images and phrases spanning 14 population subgroups, they demonstrate that these models exhibit meaningful biases favoring certain groups.

Additionally, they propose a debiasing technique tailored for large pretrained models, which can be applied as a post-processing step to mitigate bias while maintaining model accuracy.

Building on these findings, our study aims to further investigate bias in multimodal models by comparing their performance on both the MMBias dataset and real-world data from social media.

3 Hypothesis and Plan

Real-World Bias Testing: Evaluating Bias in Social Media and News

Real-world bias testing builds on the MMBias framework by going beyond its curated dataset to see how Vision-Language Models (VLMs) behave when exposed to unfiltered content from social media and news. The key question we aim to answer is - Do AI models show the same biases when processing real-world images and text as they do in controlled benchmark datasets? AI models trained on structured datasets might not behave the same way when dealing with real, unfiltered data. By applying MMBias metrics to real-world content, we can explore whether AI biases align with biases found in media and society.

First, we collect real-world image-text data by scraping posts from platforms like Twitter, Reddit, as well as extracting headlines and images from news sources like CNN, BBC. Next, we pre-process the data by organizing text-image pairs, categorizing topics, and identifying demographic groups (e.g., race, religion, gender, nationality). Finally, we run bias evaluations using MMBias metrics, comparing how AI models, such as CLIP, ViLT, and ALBEF, associate certain terms with different groups. For example, we analyze whether a model is more likely to link 'crime' with black or Hispanic individuals when processing real news headlines.

The results of this study will help us understand how bias in AI models compares to bias in real-world media. If models reflect societal biases, it means that they are simply mirroring existing patterns in the data. If they amplify bias, it suggests that AI training methods may be reinforcing harmful stereotypes. However, if models show different biases than the media, it could indicate that data set curation plays a major role in shaping how bias is learned.

Fine-Tuning Vision-Language Models (VLMs) on Fairness-Augmented Data

The MMBias dataset represents a significant step forward in assessing bias in Vision-Language Models (VLMs). However, it has several limitations that must be addressed for a more comprehensive fairness evaluation. First, MMBias provides a limited representation of certain groups, as it focuses on only 14 subgroups while overlooking others that are crucial for fair evaluation. Second, it contains stereotypical image-text associations that reinforce cultural biases. Finally, the dataset is likely imbalanced, meaning some groups have significantly more images or text samples than others. This imbalance can lead to stronger biases in model predictions, such as disproportionately positive generalizations about Christianity compared to Judaism or Islam.

Hence, we propose a Fairness-Augmented Dataset (FAD), designed to mitigate MMBias's limitations and enhance fairness in AI evaluations. To improve group representation, FAD incorporates a wide range of ethnicities, gender identities, and social classes, ensuring a fair assessment across diverse demographics. To counteract cultural biases, FAD includes counter-stereotypical text-image pairs, e.g., alongside ”a Mexican laborer,” it presents ”a Mexican scientist” or ”a Mexican CEO” preventing AI models from associating specific identities with limited roles. Lastly, FAD ensures balanced representation by maintaining an equal number of images and text samples across demographic groups, reducing bias introduced by data imbalances.

4 Data Set

The original MMBias data set provides a valuable benchmark to evaluate stereotypical bias beyond gender and race, including religion, nationality, disability, and sexual orientation.

The existing MMBias dataset, while comprehensive, relies primarily on images sourced from Flickr. To evaluate the extent to which VLMs exhibit biases in real-world scenarios, we will create a new dataset of images extracted from social media platforms and news sources. This dataset will be able to reflect the biases present in the real-world. The data collection will focus on gathering a diverse set of images of the target groups. We will employ the following methods during data collection and filtering: (1) We will use a diverse set of keywords and search terms that are related to the target groups to filter out the images. (2) We will use location and language filters to obtain a more representative set of images. (3) We will remove remaining irrelevant images manually.

To mitigate the biases identified in VLMs, we will create a fairness-augmented dataset to fine-tune the models. This dataset will be generated by employing various data augmentation techniques. To begin with, the text will be augmented with synonym replacement and back-translation. Additionally, we will ensure that the augmented dataset is balanced across all target groups, preventing the model from learning to associate certain attributes with specific groups due to data imbalance. Furthermore, the images can be augmented using different techniques such as cropping, flipping, and color jittering.

By creating and utilizing these datasets, this project will provide a more comprehensive analysis of biases in VLMs.

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.*" In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics.

[2] Wei Guo and Aylin Caliskan. "*Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases.*" In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pages 122–133.

[3] Shikha Bordia and Samuel R Bowman. "*Identifying and reducing gender bias in word-level language models.*" In Proceedings of the 2019 Conference of the North.

[4] Moin Nadeem, Anna Bethke, and Siva Reddy. "*StereoSet: Measuring stereotypical bias in pretrained language models*". In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing.

[5] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. "*Social biases in nlp models as barriers for persons with disabilities.*" arXiv preprint arXiv:2005.00813.