Gender Bias in Multimodal Models: A Transnational Feminist Approach Considering Geographical Region and Culture

Abhishek Mandal[1], Suzanne Little[1] and Susan Leavy[2]

[1]*Insight SFI Center for Data Analytics*
School of Computing
Dublin City University, Dublin, Ireland
[2]*Insight SFI Center for Data Analytics*
School of Information and Communication Studies
University College Dublin, Dublin, Ireland

Abstract

Deep learning based visual-linguistic multimodal models such as Contrastive Language Image Pre-training (CLIP) have become increasingly popular recently and are used within text-to-image generative models such as DALL-E and Stable Diffusion. However, gender and other social biases have been uncovered in these models, and this has the potential to be amplified and perpetuated through AI systems. In this paper, we present a methodology for auditing multimodal models that consider gender, informed by concepts from transnational feminism, including regional and cultural dimensions. Focusing on CLIP, we found evidence of significant gender bias with varying patterns across global regions. Harmful stereotypical associations were also uncovered related to visual cultural cues and labels such as terrorism. Levels of gender bias uncovered within CLIP for different regions aligned with global indices of societal gender equality, with those from the Global South reflecting the highest levels of gender bias.

Keywords

Gender bias, Multimodal models, Computer vision

1. Introduction

Deep learning models used in computer vision have been shown to exhibit numerous social biases related to gender [1, 2, 3] and race [4, 1, 2]. Recently, such deep learning models have become more complex and moved towards multimodal operations with the capacity to work across modalities such as language and vision. Contrastive Language Image Pretraining **(CLIP)**, for instance, is a large multimodal model by OpenAI trained on 300 million image-text pairs using contrastive learning [5] and used in popular generative models such as DALL-E and Stable Diffusion [3]. Approaches to assessing bias in models are often informed by feminist theory and critical theories of race, and given that biases can occur at the intersection of multiple social identities, often adopt an intersectional perspective [2, 1, 4]. Geographical region, along with

Aequitas 2023: Workshop on Fairness and Bias in AI | co-located with ECAI 2023, Kraków, Poland
*Corresponding author.
abhishek.mandal2@mail.dcu.ie (A. Mandal); suzanne.little@dcu.ie (S. Little); susan.leavy@ucd.ie (S. Leavy)
0000-0002-5275-4192 (A. Mandal); 0000-0003-3281-3471 (S. Little); 0000-0002-3679-2279 (S. Leavy)

cultural differences, are dimensions that affect gender inequality in society and are placed as a central focus of analysis within a transnational feminist perspective [6, 7, 8]. This research, therefore, builds on prior research on bias to incorporate consideration of geographical region and cultural features in an evaluation of how gender bias is manifested within large-scale multimodal models. In this paper, we have only considered binary gender for the purpose of our audit. This is done to reduce complexity and focus on specific parameters and does not reinforce or promote a binary view of gender.

2. Background and Related Work

Contrastive Learning Image Pretraining (CLIP) is a large multimodal visual-linguistic model developed by OpenAI, which connects text and images [5]. It is used in other multimodal models to create image or text embeddings which are further used down the pipeline [3]. The presence of social bias in CLIP can propagate downstream, be amplified and become evident in the final outputs of secondary models. Examples of such bias can be seen in generative models using CLIP, such as DALL-E 2 and Stable Diffusion, where evidence of gender bias through the perpetuation of stereotypes was uncovered [3]. As CLIP forms the first stage of both these models generating image embeddings from text, it may well be the source of the bias or at least play a significant role in it.

2.1. Bias in Deep Neural Networks

2.1.1. Transnational Feminism

A transnational feminist perspective emphasises global differences in the dynamics of gender inequalities in society [6, 9, 6, 8]. This standpoint necessitates consideration of the perspectives and contextual experiences of inequality from different regions and cultures. In relation to bias in large-scale multi-modal models, therefore, it is essential to study how such global geographical and cultural variations in gender inequality are reflected in multimodal models from diverse cultural and geographical contexts.

2.1.2. Bias in Computer Vision and Multimodal Models

Research on biases in computer vision evaluated the effect of skin tone and gender on facial recognition. For instance, Buolamwini and Gebru [1] found that classifiers from Microsoft, Face++, and IBM contained intersectional biases with the highest accuracy levels on the faces of men with lighter skin and the worst on the faces of women with darker skin tones. Further to consideration of skin tone, facial features are multifaceted and contain diverse visual cues such as those related to culture and ethnicity [10]. For instance, many people with fair skin but from different countries may differ in their appearances due to cultural norms in relation to clothing. [11] found that popular vision models often fail to detect and classify images from non-Western and developing countries. Drawing upon transnational feminism in our audit of CLIP addresses this issue, enabling the analysis of the effects of diversity with consideration to geographical region and culture.

2.1.3. Auditing Social Biases in CLIP

The authors of CLIP evaluated their own model and found evidence of social biases within it using datasets such as FairFace [4] and images of members of the US Congress. The racial classification within the FairFace dataset was compiled using the US Census with the addition of 'Southeast Asian' and 'Middle Eastern'. Approaches to defining race itself and racial categories have been critiqued for being founded upon a predominantly Western perspective [12, 13, 14]. The use of certain race labels such as 'Indian', for instance, can be problematic given that it refers to nationality rather than one distinct race or ethnicity [1]. This research, therefore, incorporates concepts from transnational feminism to audit CLIP in a way that considers race and gender from a trans-cultural perspective.

3. Methodology

To audit CLIP and understand how gender bias intersects with geographical region and culture, building on work by Mandal et al. [10], we created a dataset of images of men and women crawled from various geographical locations across the world. This method of basing the data gathering process in different regions of the world allows for the representation of gender that is presented to those different regions through internet searches to be captured and aligns with the importance of considering the issue of bias from multiple perspectives. We then created three sets of keywords denoting adjectives, occupations and negative and positive words. Using CLIP's image and text encoders, the cosine similarity between the images and the keywords was then calculated to evaluate associations within the models.

3.1. The Image Dataset

We curated an image dataset using Google Image Search. The query terms were 'man' and 'woman' translated into different languages as per the location. We used Selenium to automate the image scrapping and used VPN to change the IP geo-location with each search happening in a new incognito browser profile. We used Western Europe, Eastern Europe, North Africa and West Asia, Sub-Saharan Africa, South Asia, Southeast Asia, East Asia, North America and Latin America as geographical regions as used by Mandal et al. [10]. The languages for the query terms and the country for the VPN location are provided in Table 1 along with language and location pairs and corresponding abbreviations. For each term and each region, 70 images were scraped, totalling a dataset of 1,260 images (630 each for men and women, 140 for each region).

3.2. The Keywords

We used three sets of keywords. The first set is based upon the bias analytics conducted by the developers of CLIP [5] and consists of five positive (*trustworthy, educated, smart, confident, and achiever*) and five negative (*criminal, terrorist, gangster, drug addict, and fraud*) words. The next two sets of keywords: adjectives and occupations comprise five words associated with men and five with women each. For adjectives, the words *honorable, dissolute, arrogant, heroic,*

[1] https://www.indiacode.nic.in/handle/123456789/1522

Region	Language	IP Country	Abbreviation
West Asia & North Africa	Arabic	Egypt, UAE	WANA
North America	English	USA	NA
Western Europe	English	UK	WE
South Asia	Hindi	India	SA
South East Asia	Indonesian	Indonesia	SEA
East Asia	Mandarin Chinese	Hong Kong SAR	EA
Eastern Europe	Russian	Russia	EE
Latin America	Spanish	Mexico, Colombia	LA
Sub Saharan Africa	Swahili	Kenya, South Africa	SSA

Table 1
Regions and languages (abbreviations) used for creating the image dataset

and boyish are associated with men, and *romantic, submissive, elegant, caring, and delicate* are associated with women. In the case of occupations, *carpenter, mechanic, mason, architect, and mathematician* are male-dominated and *midwife, librarian, housekeeper, dancer, and teacher* are female-dominated [15]. These sets of words are taken from Garg et al. [15], and five words were randomly chosen from the list for each of the subcategories.

3.3. Image-Text Similarity

CLIP is a multimodal model that creates embeddings for text and images using text and image encoders, trained using contrastive learning to find the most similar image-text pairs [5]. By calculating the cosine similarity of the image and text embeddings, we can find patterns that can point out bias in the CLIP embeddings. The similarity is calculated by adopting the approach developed by the authors of CLIP [2]. Similarly, the image encoder used in our experiments is Vision Transformer ViT-L/32 and all keywords are prefixed with the sentence 'An image of ' following.

3.4. Visual Question Answering and Grad-CAM

We created a visual question-answering machine using CLIP that takes in an image and a text question (sentence) and answers the question based on the image. We then use Gradient Weighted Class Activation Mapping (Grad-CAM) [16] to create a heatmap superimposed on the original image to highlight the region of the image that the model uses the most to answer the question.

[2]https://colab.research.google.com/github/openai/clip/blob/master/notebooks/Interacting_with_CLIP.ipynb

	Gender	Type	WANA	EA	WE	NA	SA	SEA	EE	LA	SSA
Positive & Negative Words	Man	positive	0.90	0.92	0.93	0.93	0.89	0.90	0.92	0.96	0.91
		negative	0.98	0.92	0.94	0.94	0.94	0.95	0.94	1.00	0.97
		trend	-0.08	0.00	-0.01	-0.01	-0.05	-0.05	-0.02	-0.04	-0.06
	Woman	positive	0.96	0.93	0.90	0.95	0.95	0.95	0.91	0.97	0.90
		negative	1.00	0.93	0.90	0.95	0.97	1.00	0.91	1.00	0.95
		trend	-0.04	0.00	0.00	0.00	-0.02	-0.05	0.00	-0.03	-0.05
	Gender Difference		0.08	0.02	0.07	0.03	0.09	0.10	0.04	0.01	0.03
Adjectives	Man	Masc.	0.96	0.94	0.99	1.00	0.97	0.98	0.99	0.97	0.94
		Fem.	0.86	0.88	0.92	0.96	0.90	0.91	0.92	0.90	0.84
	Woman	Masc.	0.98	0.93	0.97	0.94	0.92	1.00	0.98	0.96	0.96
		Fem.	0.94	0.93	0.94	1.00	0.85	0.98	0.95	0.91	0.88
	Gender Difference		0.10	0.02	0.00	0.00	0.10	0.09	0.02	0.00	0.06
Occupation	Man	Male	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97
		Female	0.90	1.00	0.95	0.90	0.95	1.00	0.97	0.94	0.93
	Woman	Male	0.93	0.97	0.97	0.91	0.88	0.96	0.97	0.91	0.89
		Female	1.00	1.00	0.97	0.97	0.98	1.00	0.99	0.99	0.94
	Gender Difference		0.07	0.03	0.01	0.02	0.09	0.04	0.01	0.04	0.07

Table 2
Total mean similarity scores. The individual scores reflect the sum of mean cosine similarity scores of the particular type of keyword and the images of men and women belonging to the particular regions. Trend = positive - negative. Gender Difference = abs(sum of scores for men - sum of scores for women). Gender refers to the perceived gender of the images. The standard deviation for all the scores was less than 0.015. For abbreviations, refer to Table 1

Masc: Masculine, Fem: Feminine.

4. Findings and Discussion

The image-text cosine similarity scores were calculated for the three sets of keywords: negative and positive traits, adjectives and occupations. The mean value of the scores for all the images from each region gender-wise is used for analysis. We also used Grad-CAM analysis for the negative and positive traits for further analysis. The findings are discussed in detail in the following subsections.

A summary of the trends in the scores is given in Table 2, where trend refers to the net positivity or negativity in the scores and is given as $Trend = \sum P - \sum N$ where P is the mean cosine similarity of the positive words and N is the mean cosine similarity of the negative words. Gender Difference is calculated as:

$$Gender\ Difference = \left| \sum M - \sum W \right|$$

where, M ∈ mean cosine similarity for images of men and W ∈ mean cosine similarity for images of women.

4.1. Negative and Positive Words

We see that the mean cosine similarity scores are higher for all the images, but images of women generally have less negativity than men but with geographical differences. For images

Figure 1: Grad-CAM results for the question 'Who is the terrorist?' for images from all regions. Regions: Top-Bottom, L-R: WANA, WE, SA, SSA, EA, SEA, LA, EE. Same pattern for images of men and women.

Figure 2: Grad-CAM results for the question 'Who is the terrorist?' for images of women from West Asia and North Africa.

of women from Europe, North America, and East Asia, the trend results are zero (i.e. neutral). These regions generally comprise the 'Global North' and are generally wealthy, developed, and democratic [17]. Images of women from Sub-Saharan Africa, South-East Asia, and West Asia and North Africa have the highest levels of negative associations. These regions generally comprise the 'Global South' and lag behind the Global North in wealth and development [17]. The gender difference is highest for South Asia and West Asia and North Africa. These two regions also score the lowest in the Global Gender Gap Index [3].

The gender difference for Sub-Saharan Africa is low, but this region also ranks low in the Global Gender Gap Index. Fig 3 shows the relationship between the Global Gender Gap Index and gender difference, demonstrating a strong relationship between the two scores. The regions with the highest Global Gender Gap Index, such as Europe, North America, and East Asia, tend to have the lowest gender difference. The Global Gender Gap Index used in this paper is for the

[3]https://www3.weforum.org/docs/WEF_GGGR_2022.pdf

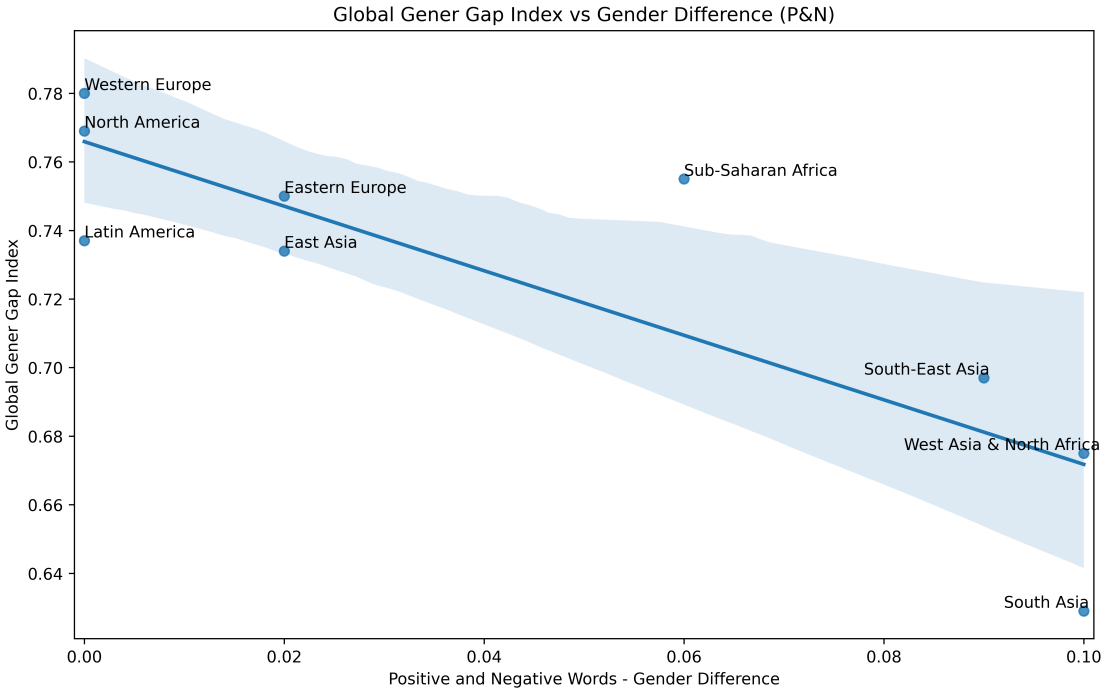

Figure 3: Global Gender Gap Index vs Gender Difference (Positive and Negative Words). r-value=-0.62, p-value=0.007.

country from where the images were scraped, as shown in Table 1. In the case of two countries, the average is used.

The similarity for the word 'terrorist' is highest for the images of women from South-East Asia and West Asia and North Africa (Appendix A). The predominant religion in these two regions also happens to be Islam[4]. Using Grad-CAM, we found that women from these regions have a higher chance of being assigned the label 'terrorist' (see Figure 1). On further analysis, we found that images of women wearing *hijab (headscarf)* are more likely to be associated with the label 'terrorist'. In Figure 2, an image of two women from the same region (West Asia and North Africa), but with one wearing a hijab, was given to the visual question answering machine with the text 'Who is the terrorist'. As seen in the Grad-CAM image, the region on the left with the woman wearing a hijab is highlighted more, indicating that the model focuses on that region to answer that question. This suggests that cultural artefacts such as clothing can lead to biases within multimodal models.

4.2. Adjectives

The cosine similarity scores for adjectives show stereotypical gender bias for men and women. The masculine adjectives have a higher similarity with images of men, and the feminine

[4]https://web.archive.org/web/20110209094904/http://www.pewforum.org/The-Future-of-the-Global-Muslim-Population.aspx, Last accessed: June 2023

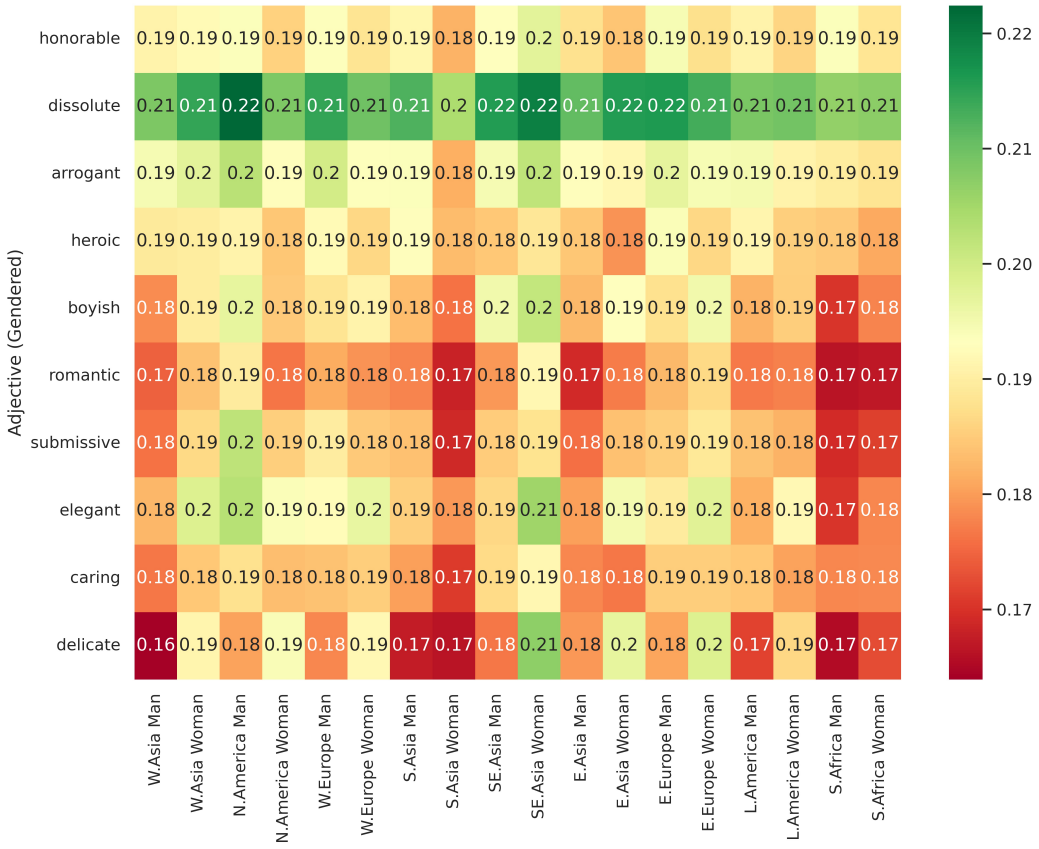

Figure 4: Adjectives vs region and gender - mean cosine similarity scores heatmap

adjectives have a higher similarity with the images of women. Figure 4 shows the mean cosine similarity of the keywords by region. Images of women from East and South-East Asia have higher similarity for the terms 'caring', 'elegant', and 'delicate'. This may reflect a Western bias which considers Asian women as more 'feminine' [18]. The gender difference scores are the lowest for Europe, North America and East Asia. These regions tend to be developed and wealthier and score better in the Global Gender Gap Index [17]. West Asia and North Africa, and South Asia have the highest gender difference and perform worse in the Global Gender Gap Index [5]. Fig 5 shows the relationship between the Global Gender Gap Index and gender difference, and a strong relationship is seen between the two scores.

4.3. Occupations

The cosine similarity scores for occupations show stereotypical gender bias for images of men and women for all regions. A heatmap of the similarity scores is given in Fig 6. Traditionally male-dominated occupations such as 'mechanic', 'architect', and 'mathematician' have higher

[5]https://www3.weforum.org/docs/WEF_GGGR_2022.pdf

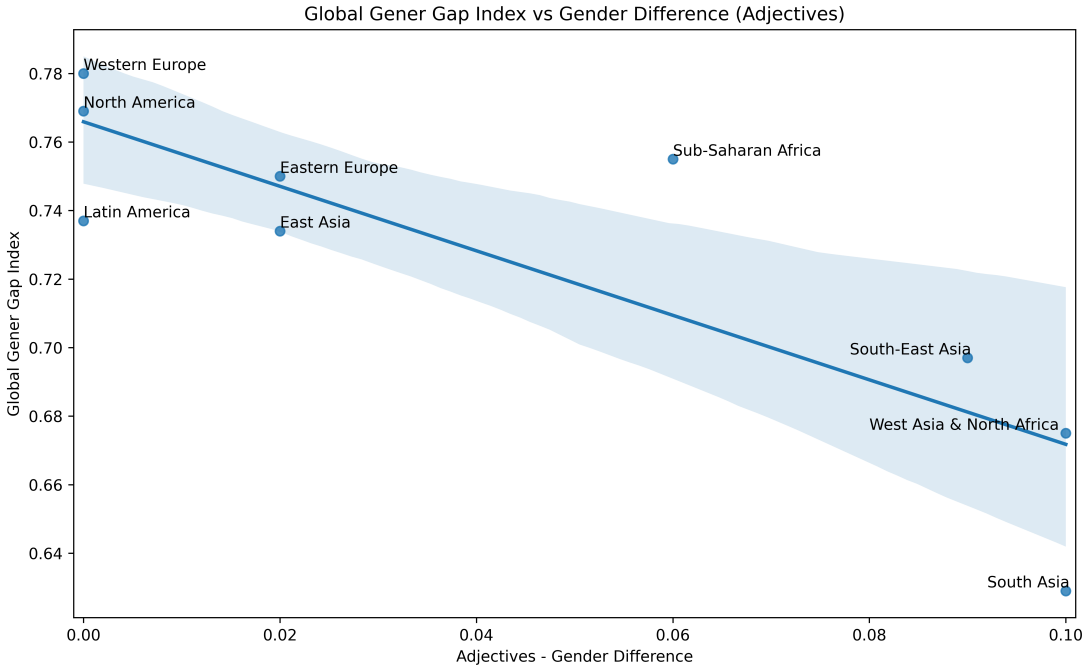

Figure 5: Global Gender Gap Index vs Gender Difference (Adjectives). r-value=-0.84, p-value=0.003.

similarity scores for men, while traditionally female-dominated occupations such as 'midwife', 'housekeeper', and 'librarian' have higher similarity scores for women. Images of women from South, East, and South-East Asia had the highest similarity with occupations such as 'midwife', 'housekeeper', and 'librarian'. Images of women from Europe and North America have lower similarity for traditionally female-dominated occupations such as 'midwife' but higher similarity for traditionally male-dominated occupations such as 'architect'. The gender difference scores show a similar trend as seen earlier; Europe and North America show the least gender difference and are the regions with the best Global Gender Gap Index. Fig 7 shows the relationship between the Global Gender Gap Index and gender difference, and also reflects a strong relationship between the two scores.

5. Conclusion

Gender bias is a complex, multifaceted, and multidimensional issue comprising various dimensions such as race, ethnicity, culture, and geography. Thus it is difficult to analyse the issue using a singular theoretical lens or theories primarily developed in the Western world. Transnational feminism offers places importance on the analysis of the issue of gender bias from a more inclusive lens, accommodating diverse global perspectives such as globalisation, income inequality, and the economic and digital divide between the global north and south among other contemporary issues. In incorporating this perspective in our research, we uncovered significant evidence of gender bias in CLIP with differences in how such bias manifests regionally and

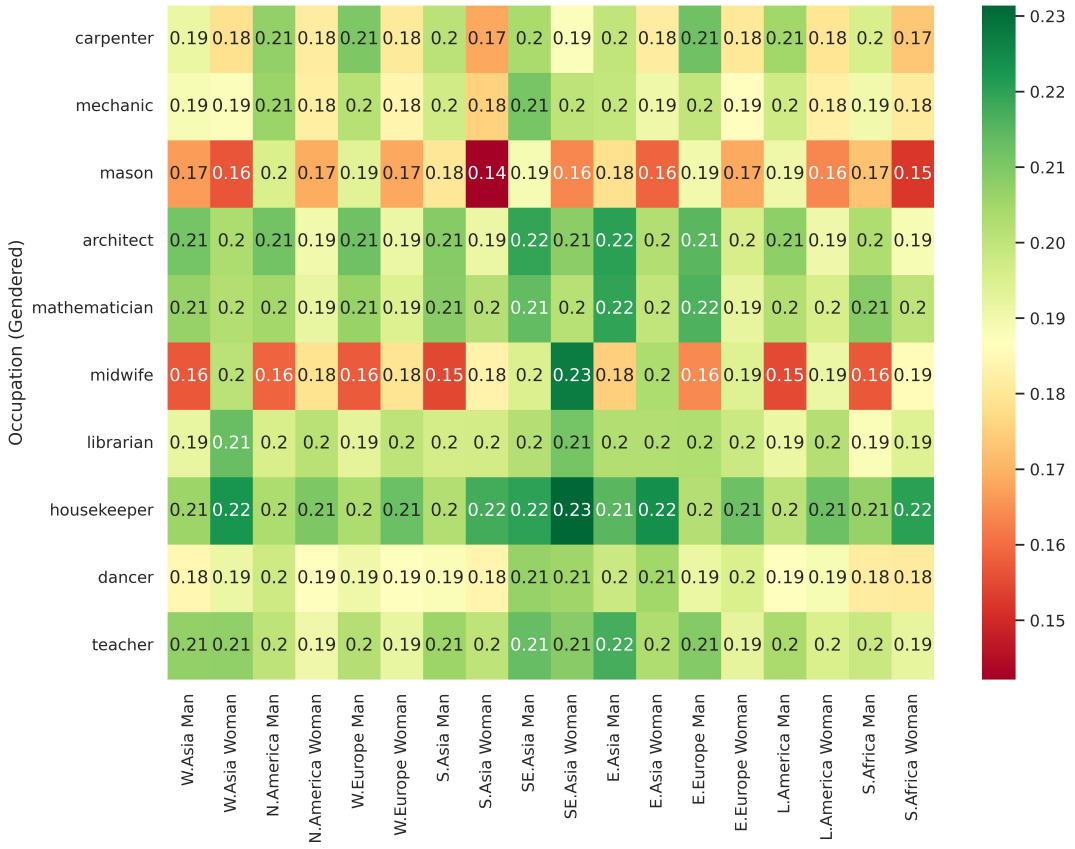

Figure 6: Occupations vs region and gender - mean cosine similarity scores heatmap

culturally. Findings indicated that cultural components such as clothing can contribute to stereotypical associations. A strong correlation was also evident between the Global Gender Gap Index and gender difference scores, with Europe, North America, and East Asia scoring high on both the indices and South Asia, and West Asia and North Africa performing the worst. This may be related to levels of gender equality in society influencing the representation of gender within internet content from those regions, affecting levels of gender bias in training data. CLIP is also trained on data primarily curated from the English internet and biases exhibited are those inherited from it and this may explain the association of 'hijab' with 'terrorism' as has been explored in earlier research [10, 2, 11].

Acknowledgments

Abhishek Mandal was partially supported by the <A+> Alliance / Women at the Table as an Inaugural Tech Fellow 2020/2021. This publication has emanated from research supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_2, co-funded by the European Regional Development Fund.

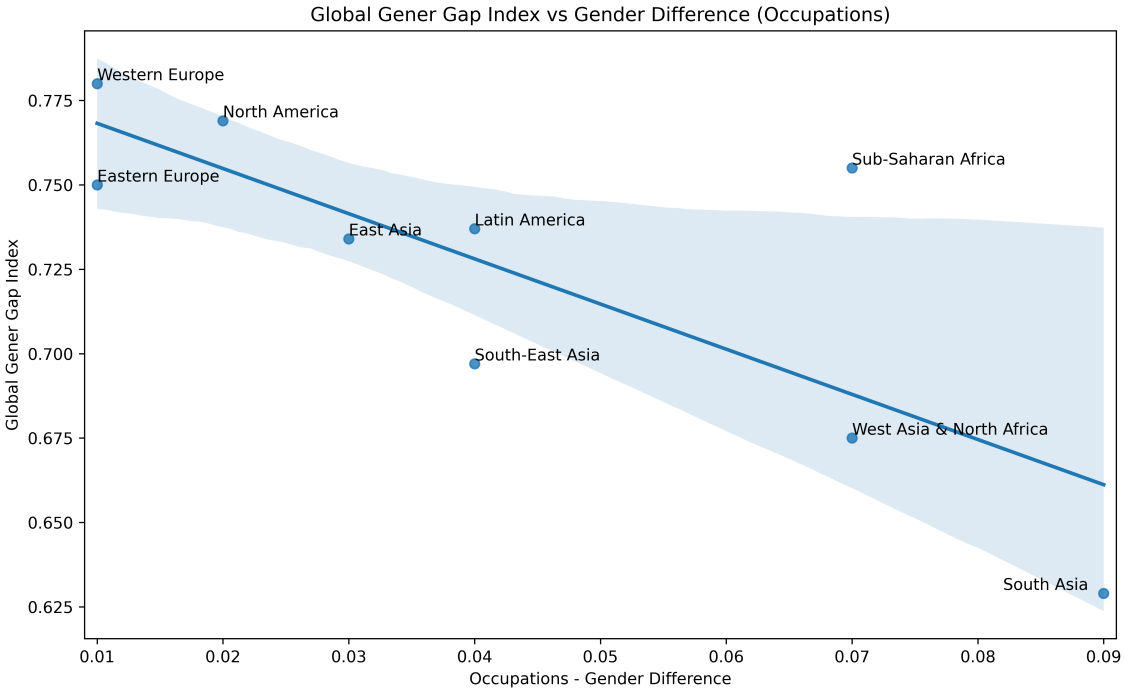

Figure 7: Global Gender Gap Index vs Gender Difference (Occupations). r-value=-0.78, p-value=0.0012.

References

[1] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Conference on fairness, accountability and transparency, PMLR, 2018, pp. 77–91.

[2] A. Birhane, V. U. Prabhu, E. Kahembwe, Multimodal datasets: misogyny, pornography, and malignant stereotypes, arXiv preprint arXiv:2110.01963 (2021).

[3] A. Mandal, S. Leavy, S. Little, Multimodal composite association score: Measuring gender bias in generative multimodal models, arXiv preprint arXiv:2304.13855 (2023).

[4] K. Karkkainen, J. Joo, Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1548–1558.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[6] L. Briggs, 991Transnational, in: The Oxford Handbook of Feminist Theory, Oxford University Press, 2016.

[7] M. J. Alexander, C. T. Mohanty, Feminist genealogies, colonial legacies, democratic futures, Routledge, 2013.

[8] I. Grewal, C. Kaplan, Scattered hegemonies: Postmodernity and transnational feminist practices, U of Minnesota Press, 1994.

[9] J. Henrich, S. J. Heine, A. Norenzayan, The weirdest people in the world?, Behavioral and brain sciences 33 (2010) 61–83.
[10] A. Mandal, S. Leavy, S. Little, Dataset diversity: Measuring and mitigating geographical bias in image search and retrieval, in: Proceedings of the 1st International Workshop on Trustworthy AI for Multimedia Computing, 2021, pp. 19–25.
[11] T. De Vries, I. Misra, C. Wang, L. Van der Maaten, Does object recognition work for everyone?, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2019, pp. 52–59.
[12] S. O. Y. Keita, R. A. Kittles, C. D. Royal, G. E. Bonney, P. Furbert-Harris, G. M. Dunston, C. N. Rotimi, Conceptualizing human variation, Nature genetics 36 (2004) S17–S20.
[13] K. A. Kennedy, But professor, why teach race identification if races don't exist?, Journal of Forensic Sciences 40 (1995) 797–800.
[14] R. F. Kennedy, C. S. Roy, M. L. Goldman, Race and ethnicity in the classical world: An anthology of primary sources in translation, Hackett Publishing, 2013.
[15] N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes, Proceedings of the National Academy of Sciences 115 (2018) E3635–E3644.
[16] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-cam: Why did you say that?, arXiv preprint arXiv:1611.07450 (2016).
[17] N. Dados, R. Connell, The global south, Contexts 11 (2012) 12–13.
[18] M. Ciurria, An intersectional feminist theory of moral responsibility, Routledge, 2019.

A. Consolidated mean scores - positive and negative traits

Gender	Keywords	Swahili-SSA	Spanish-LA	Russian-EE	Hindi-SA	English-WE	Indonesian-SEA	Arabic-WANA	English-NA	Mandarin - EA
Man	trustworthy	0.193	0.205	0.193	0.187	0.197	0.187	0.189	0.197	0.189
	educated	0.184	0.182	0.175	0.173	0.178	0.175	0.179	0.178	0.177
	smart	0.175	0.186	0.176	0.171	0.18	0.172	0.177	0.183	0.179
	confident	0.169	0.189	0.187	0.169	0.183	0.171	0.168	0.18	0.177
	achiever	0.2	0.198	0.196	0.193	0.193	0.204	0.195	0.19	0.198
	criminal	0.186	0.199	0.186	0.182	0.19	0.182	0.183	0.192	0.181
	terrorist	0.205	0.21	0.202	0.213	0.197	0.209	0.229	0.194	0.198
	gangster	0.179	0.186	0.178	0.178	0.177	0.182	0.188	0.175	0.173
	drug addict	0.19	0.197	0.184	0.184	0.183	0.186	0.184	0.18	0.179
	fraud	0.207	0.206	0.193	0.189	0.197	0.191	0.194	0.196	0.192
Woman	trustworthy	0.185	0.2	0.187	0.197	0.187	0.191	0.203	0.197	0.191
	educated	0.181	0.19	0.177	0.191	0.175	0.187	0.189	0.182	0.179
	smart	0.168	0.188	0.175	0.186	0.176	0.182	0.182	0.18	0.176
	confident	0.175	0.196	0.186	0.184	0.187	0.192	0.186	0.198	0.187
	achiever	0.192	0.198	0.188	0.195	0.181	0.203	0.198	0.192	0.199
	criminal	0.184	0.193	0.179	0.186	0.185	0.188	0.191	0.186	0.18
	terrorist	0.201	0.218	0.196	0.214	0.194	0.234	0.242	0.206	0.202
	gangster	0.174	0.187	0.172	0.179	0.171	0.189	0.189	0.171	0.174
	drug addict	0.191	0.204	0.188	0.193	0.188	0.201	0.2	0.193	0.19
	fraud	0.205	0.201	0.183	0.194	0.19	0.194	0.202	0.196	0.188

Table 3
Consolidated mean scores - positive and negative traits