Towards Region-aware Bias Evaluation Metrics

Anonymous ACL submission

Abstract

When exposed to human-generated data, language models are known to learn and amplify societal biases. While previous work has introduced benchmarks that can be used to assess the bias in these models, they rely on assumptions that may not be universally true. For instance, a bias dimension commonly used by these metrics is that of *family–career*, but this may not be the only common bias in certain regions of the world. In this paper, we identify topical differences in gender bias across different regions and propose a region-aware bottom-up approach for bias assessment. Our proposed approach uses gender-aligned topics for a given region and identifies gender bias dimensions in the form of topic pairs that are likely to capture gender societal biases. Several of our proposed bias dimensions are on par with human perception of gender biases in these regions in comparison to the existing ones, and we also identify new dimensions that are more aligned than the existing ones.

1 Introduction

Human bias refers to the tendency of prejudice or preference towards a certain group or an individual and can reflect social stereotypes with respect to gender, age, race, religion, and so on.

Bias in machine learning (ML) refers to prior information which is a necessary prerequisite for intelligence (Bishop, 2006). However, biases can be problematic when prior information is derived from *harmful precedents* like prejudices and social stereotypes. Early work in detecting biases includes the Word Embedding Association Test (WEAT) (Caliskan et al., 2017) and the Sentence Encoder Association Test (SEAT) (May et al., 2019). WEAT is inspired by the Implicit Association Test (IAT) (Greenwald et al., 1998) in psychology, which gauges people's propensity to unconsciously link particular characteristics—like family versus career—with specific target groups—like female (F) versus male (M). WEAT measures the distances between target and attribute word sets in word embeddings using dimensions similar to those used in IAT.

Biases toward or against a group can vary across different regions due to the influence of an individual's culture and demographics (Grimm and Church, 1999; Kiritchenko and Mohammad, 2018; Garimella et al., 2022). However, existing bias evaluation metrics like WEAT and SEAT follow a "one-size-fits-all" approach to detect biases across different regions. As biases can be very diverse depending on the demographic lens, a fixed or a small set of dimensions (such as family–career, math–arts) may not be able to cover all the possible biases in society. In this paper, we address two main research questions about gender bias: (1) Is it possible to use current NLP techniques to automatically identify gender bias characteristics (such as family, career) specific to various regions? (2) How do these gender dimensions compare to the current generic dimensions included in WEAT/SEAT?

The study makes two main contributions: (1) An automatic method to uncover gender bias dimensions in various regions that uses (a) topic modeling to identify dominant topics aligning with the F/M groups for different regions, and (b) an embedding-based approach to identify F-M topic pairs for different regions that can be viewed as gender bias dimensions in those regions; and (2) An IAT-style test to assess our results of automatic bias detection with humans. To the best of our knowledge, this is the first study to use a data-driven, bottom-up methodology to evaluate bias dimensions across regional boundaries.

2 Data

We use GeoWAC (Dunn and Adams, 2020a), a geographically balanced corpus that consists of web pages from Common Crawl. Language samples are geo-located using country-specific domains, such as a *.in* domain suggesting Indian origin (Dunn and Adams, 2020b). GeoWAC's English corpus spans 150 countries. We select the top three coun-

Target words - Attribute words	Region	WEAT
Male names vs Female names - career vs family	Africa	1.798
	Asia	1.508
	North America	**1.885**
	Europe	1.610
	Oceania	1.727
Math vs Arts - Male terms vs Female terms	Africa	**1.429**
	Asia	1.187
	North America	0.703
	Europe	0.334
	Oceania	1.158
Science vs Arts - Male terms vs Female terms	Africa	**1.247**
	Asia	0.330
	North America	0.036
	Europe	-0.655
	Oceania	0.725

Table 1: Region-wise WEAT scores using word2vec.

tries with the most examples per region: Asia, Africa, Europe, North America, and Oceania as in (Garimella et al., 2022). Psychological studies and experiments that demonstrate human stereotypes vary by continental regions (Damann et al., 2023; Blog, 2017) and even larger concepts like western and eastern worlds (Markus and Kitayama, 2003; Jiang et al., 2019) serve as an inspiration for the use of regions to determine differences across cultures. Dataset details are included in Appendix A.

3 Variations in Gender Bias Tests Across Regions

We investigate the differences in existing gender bias tests across different regions using WEAT. WEAT takes in *target words* such as male names and female names, to indicate a specific group, and *attribute words* that can be associated with the *target words*, such as "Math" and "Art". It computes bias by finding the cosine distance between the embeddings of the target and attribute words. We compute WEAT scores using word2vec embeddings (Mikolov et al., 2013) trained on five regions separately. Table 1 shows the region-wise scores for the three gender tests.

Although we see a positive bias for most gender bias dimensions, the scores vary across regions. For example, the *family–career* dimension is the most predominant one for North America, Africa, and Oceania, whereas in Asia, *math–arts* is predominant. Europe has a negative bias on *science–arts* (indicating a stronger F-science and M-arts association). These results provide preliminary support to our hypothesis that gender bias dimensions vary across regions, thus propelling a need to come up with further bias measurement dimensions to better capture gender biases in these regions in addition to the existing generic ones in WEAT.

4 A Method to Automatically Detect Bias Dimensions Across Regions

We propose a two-stage approach to automatically detect region-aware bias dimensions that likely capture the biases in specific regions in a bottom-up manner. In the first stage, we utilize topic modeling to identify prominent topics in each region, and the second stage involves using an embedding-based approach to find pairs of topics among those identified in the first stage that are likely to represent prominent gender bias dimensions in each region.

4.1 Identifying Region-wise Bias Topics

We use topic modeling to identify dominant topics in the male and female examples in each region.

We build Female- and Male-aligned datasets (F-M datasets) using the examples from GeoWAC for each region. We use 52 pairs of gender-defined words that are non-stereotypically F/M (e.g., wife, brother, see Appendix E) from (Bolukbasi et al., 2016), and find examples that contain these words. These datasets are used to find gender-aligned topics from GeoWAC. The dataset statistics are specified in Table 5 in Appendix B.

For topic modeling, we use Bertopic (Grootendorst, 2022), which identifies an optimal number of topics n for a given dataset (see Appendix F.1 for implementation details). We further refine the resulting topics using Llama 2 (Touvron et al., 2023) to label and better understand the topic clusters identified by Bertopic. The prompting mechanism for Llama2 is provided in Appendix G.

We next compute the alignment of topics to either F/M groups. We first compute the topic distribution of a data point, which gives the probability p_{it} of an example i belonging to each topic t. For a topic t, we take n examples that dominantly belong to topic t: $i_1, i_2, ..., i_n$. If m out of n data points belong to the F group in the F-M dataset, and the other (n - m) belongs to the M group, we compute the average of topic probabilities for both groups separately: $p_{Ft} = \frac{(p_{i_1 t} + p_{i_2 t} + ... + p_{i_m t})}{m}$ and $p_{Mt} = \frac{(p_{i_{m+1} t} + p_{i_{m+2} t} + ... + p_{i_n t})}{(n-m)}$, where p_{Ft} and p_{Mt} refer to the average probability by which a topic dominantly belongs to the F and M groups respectively. If $p_{Ft} > p_{Mt}$, we say the topic is a *bias topic* that aligns with the F group and vice-versa.

Region	Female	Male
Africa	Credit cards and finances, Royalty and Media, Trading strategies and market analysis, Dating and relationships guides, Parenting and family relationships	Fashion and Lifestyle, Male enhancement and sexual health, Nollywood actresses and movies, Nigerian politics and government, Essay writing and research
Asia	Hobbies and Interests, Healthy eating habits for children, Social media platforms, Royal wedding plans, Online Dating and Chatting	DC comic characters, Mobile Application, Phillippine Politics and Government, Sports and Soccer, Career
Europe	Pets and animal care, Fashion and Style, Education, Obituaries and Genealogy, Luxury sailing	Political developments in Northern Ireland, Christian Theology and Practice, Crime and murder investigation, EU Referendum and Ministerial Positions, Criminal Justice System
North America	Pets, Cooking: culinary delights and chef recipes, Fashion and style, Family dynamics and relationships, Reading and fiction	Civil War and history, Middle East conflict and political tensions, Movies and filmmaking, Political leadership and party dynamics in Bermuda, Rock Music and songwriting
Oceania	Cooking and culinary delights, Romance, Weight loss and nutrition for women, Water travel experience, Woodworking plans and projects	Harry Potter adventures, Art and Photography, Superheroes and their Universes, Music recording and Artists, Football in Vanuatu

Table 2: Top 5 topics for F and M for each region

4.2 Finding Topic Pairs as Region-wise Bias Dimension Indicators

We use an embedding-based approach to identify F-M topic pairs from the pool of topics identified in the previous stage, to generate topic pairs (bias dimensions) that are comparable to IAT/WEAT pairs.

We use BERT-large (`stsb-bert-large`) from SpaCy's (Honnibal and Montani, 2017) `sentencebert` library to extract contextual embeddings for topic words for each region. For a topic t consisting of topic words $w_1, ..w_n$, the topic embedding is given by the average of embeddings of the top 10 topic words in that topic.

We identify topic pairs from the embeddings taking inspiration from (Bolukbasi et al., 2016): let the embeddings of the words she and he be E_{she} and E_{he} respectively. The embedding of a topic t_i be E_{t_i}. A female topic F_{t_i} and a male topic M_{t_j} are a topic pair if: $cos(E_{F_{t_i}}, E_{she}) \sim cos(E_{M_{t_j}}, E_{he})$ and/or $cos(E_{F_{t_i}}, E_{he}) \sim cos(E_{M_{t_j}}, E_{she})$, where $cos(i, j)$ refers to the cosine similarity between embeddings i and j, given by $cos(i, j) = \frac{i.j}{||i||||j||}$. The threshold for the difference between the cosine similarities we consider for two topics to be a pair is 0.01, i.e., two topics $(t1, t2)$ are considered a pair if the difference of cosine similarities cos($t1$, she)/cos($t1$, he) and cos($t2$, he)/cos($t2$, she) respectively is < 0.01. We manually choose 0.01 since differences close to 0.01 are almost $= 0$.

5 Results & Discussion

Region-wise Bias Topics. Table 2 displays the top topics based on u_{mass} (Mimno et al., 2011) coherence for each region.

Region	F-M topic pair
Africa	Parenting and family relationships-Nollywood Actress and Movies Marriage and relationships - Sports and Football Womens' lives and successes - Fashion and Lifestyle Music - Social Media Dating and relationships advice - Religious and Spiritual growth
Asia	Hotel royalty - Political leadership in India Healthy eating habits for children - Sports and Soccer Royal wedding plans - Social Media platforms for video sharing Royal wedding plans - Religious devotion and spirituality Marriage - Bollywood actors and films
Europe	Education - Music Comfortable hotels - Political decision and impact on society Luxury sailing - UK Government Taxation policies Obituaries and Genealogy - Christian Theology and Practice Fashion and style - Christian theology and practice
North America	Online Dating for Singles - Religion and Spirituality Fashion and Style - Reproductive Health Education and achievements - Reinsurance and capital markets Family dynamics and relationships - Nike shoes and fashion Reading and fiction - Cape Cod news
Oceania	Family relationships - Religious beliefs and figures Woodworking plans and projects - Music record and Artists Weight loss and nutrition for women - Building and designing boats Exercises for hormone development - Superheroes and their Universes Kids' furniture and decor - Building and designing boats

Table 3: Top 5 topic pairs for F and M for each region.

Several topics are exclusive to certain regions. Some topics like *family* and *parenting*; *cooking*; *pets* and *animal care* are common across some regions for F. Similarly we have *movies*; *politics* and *government*; *sports*; and *music* for M. Finally, there are differences between regions in terms of *education*, *reading*, and *research* (F-Europe, NA, and M-Africa), and *fashion* and *lifestyle* (F-Europe, NA, and M-Africa). Some other popular topics across regions are *religion and spirituality*, *Christian theology* in M; *obituaries and genealogy*, *online dating*, *travel*, and *sailing* in F (see Appendix H).

Region-wise Bias Dimensions. Table 3 shows the top five topic pairs per region, chosen based on the u_{mass} score from the top 10 topics each for F and M from the topic modeling scheme.

As expected, topic pairs differ by region, and we also note new topic pairs that do not appear in the WEAT tests. Among the top ones, there are recurring topics in F such as *dating and marriage*, *family and relationships*, *luxury sailing*, and *education*, whereas in M, we have *politics*, *religion*,

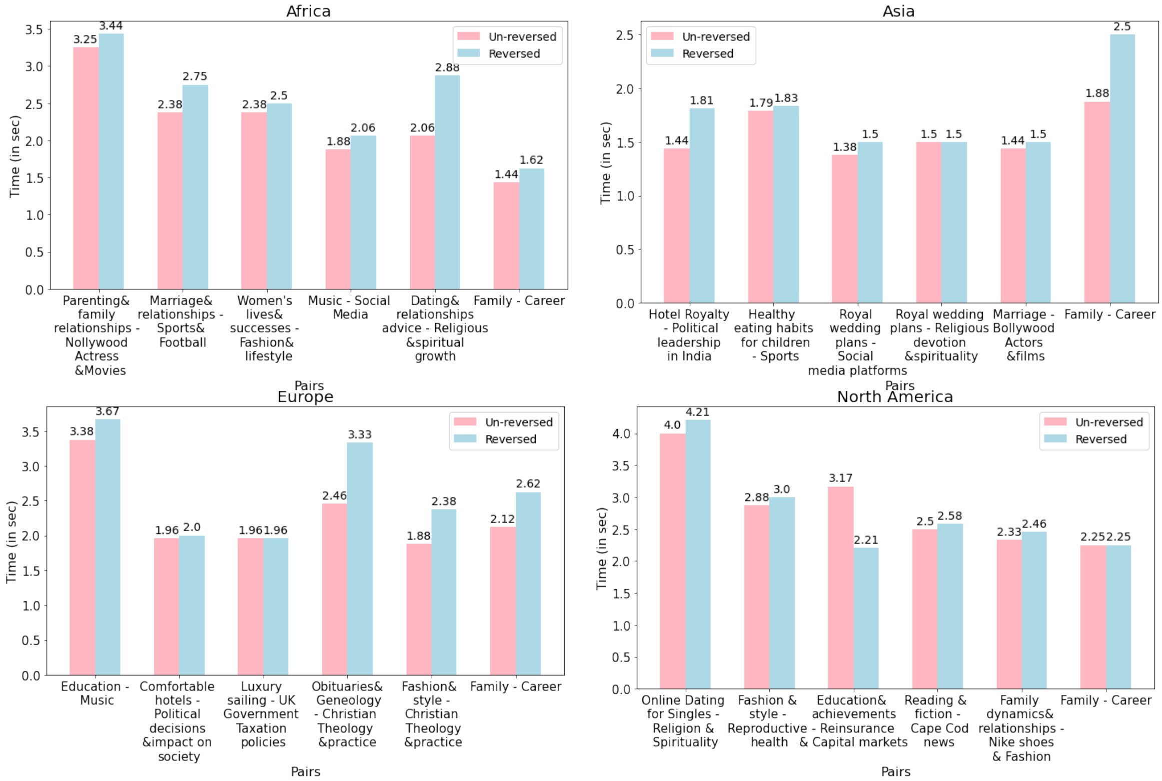


Figure 1: Human validation results across regions

sports, and *movies*. These region-specific pairs may supplement generic tests to detect regional biases. We validate this by conducting IAT-style human surveys for top regional topics.

Human Validation. Each topic pair test form contain two tasks. In one, annotators match a female face f with a female topic T_f and a male face m with a male topic T_m, timing responses as r_1 and r_2. In the reverse task, they pair T_m with f and T_f with m, timing these as r_3 and r_4. We average r_1 and r_2 for the 'un-reversed' case and r_3 and r_4 for the 'reversed' case. To avoid bias, the form order is randomized. We conducted this survey with 3 annotators each from Africa, Asia, Europe, and North America, also including a family-career topic pair, a standard WEAT bias dimension.

Human Validation Results. Fig 1 shows response times for top five topic pairs in each region for un-reversed and reversed scenarios. Larger time differences indicate more bias, suggesting that the pair could be a potential gender bias dimension for that region. If un-reversed time is lower, it suggests a stronger association of T_f with the F group and T_m with the M group. The family-career pair was also surveyed as a standard WEAT bias dimension.

As expected, the Family-Career pair has differences across most regions; it is interesting that the difference is zero in the case of North America, indicating that American annotators in our study suggested almost no biases in this dimension for the two genders.[1] We also note that some pairs, such as *dating and relationships advice–religious and spiritual growth* for Africa, *obituaries and geneology–Christian theology* for Europe, and *online dating–religion and spirituality* for North America have differences higher than those for *family–career* in the respective regions, indicating that the participants associated more biases on our uncovered bias dimensions than the existing one in WEAT. These findings support our hypothesis that gender bias dimensions vary across regions and also bring preliminary evidence that the region-aware bias dimensions we uncover are in line with the human perception of bias in those regions.

6 Conclusion

In this paper, we proposed a bottom-up approach to identify topic pairs that capture gender biases across different regions. Our human evaluation results demonstrated the validity of our proposed dimensions. Future work includes incorporating region-specific bias dimensions into existing tests, testing different model/dataset combinations, and surveying a larger population for more accurate results. We also aim to explore region-aware bias mitigation techniques.

[1] In our surveys, all the American participants happen to be males, as we do not control for gender. It would be interesting to study how these responses vary with equal participation from female and males, which will be part of our future work.

7 Limitations

Our preliminary human validation tests consist of three annotators per region (namely, Africa, Asia, Europe, and North America), which is low in comparison to existing human bias evaluation tests like IAT, which had 32 participants for their bias evaluation test. Therefore, we plan to survey a wider population to generalize our findings further. We used only one corpus, GeoWAC to obtain data from different regions and perform our experiments. Also, we only control for the regional backgrounds and not the genders of the participants. It would be interesting to study how these responses vary with equal participation from female and males, which will be part of our future work. We plan to incorporate more datasets in the future to investigate if your hypothesis holds across different domains.

8 References

References

Christopher Bishop. 2006. Pattern recognition and machine learning. *Springer google schola*, 2:5–43.

National Geographic Education Blog. 2017. What continent do you think they are from? drawing humans to reveal internalized bias.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Taylor J Damann, Jeremy Siow, and Margit Tavits. 2023. Persistence of gender biases in europe. *Proceedings of the National Academy of Sciences*, 120(12):e2213266120.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Dunn and Ben Adams. 2020a. Geographically-balanced gigaword corpora for 50 language varieties. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2528–2536.

Jonathan Dunn and Ben Adams. 2020b. Mapping languages and demographics with georeferenced corpora. *arXiv preprint arXiv:2004.00809*.

Aparna Garimella, Rada Mihalcea, and Akhash Amarnath. 2022. Demographic-aware language model fine-tuning as a bias mitigation technique. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 311–319, Online only. Association for Computational Linguistics.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Stephanie D Grimm and A Timothy Church. 1999. A cross-cultural study of response biases in personality measures. *Journal of Research in Personality*, 33(4):415–441.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Mengyin Jiang, Shirley KM Wong, Harry KS Chung, Yang Sun, Janet H Hsiao, Jie Sui, and Glyn W Humphreys. 2019. Cultural orientation of self-bias in perceptual matching. *Frontiers in Psychology*, 10:1469.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Claudia Malzer and Marcus Baum. 2020. A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE.

Hazel Rose Markus and Shinobu Kitayama. 2003. Models of agency: sociocultural diversity in the construction of action.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

A GeoWAC dataset details

Table 4 contain the total number of examples per country in a region. We consider the top three countries with the highest number of examples per region.

B F-M Dataset statistics

Table 5 displays the total number of examples from female and male groups per region for the region-specific F-M dataset.

C Cultural differences in biases using WEAT

Table 6 shows the WEAT scores for all WEAT dimensions defined in (Caliskan et al., 2017). We see several differences in WEAT scores across regions for different dimensions.

Region	Country	#Examples
Africa	Nigeria	3,153,761
	Mali	660,916
	Gabon	645,769
Asia	India	12,327,494
	Singapore	6,130,047
	Philippines	3,166,971
Europe	Ireland	8,689,752
	United Kingdom	7,044,434
	Spain	465,780
North America	Canada	7.965,736
	United States	8,521,094
	Bermuda	244,500
Oceania	New Zealand	94,476
	Palau	486,437
	Vanuatu	165,355

Table 4: Region-specific details in GeoWAC

Region	Total	#Female	#Male
Africa	57895	20153	37742
Asia	56877	21400	35477
Europe	59121	21049	38072
North America	70665	27627	43038
Oceania	62101	25951	36150

Table 5: F-M dataset statistics for regions

D Region specific BERTs to identify top words in F/M direction

To motivate our case to investigate differences in biases across regions, we use BERT to compute the top words corresponding to the *she-he* axis in the embedding space. BERT is a pre-trained transformer-based language model that consists of a set of encoders. As a motivation experiment to identify differences in the contextual embedding space for different regions, we fine-tune BERT with the masked language modeling task (no labels) for each region separately. We then compute embeddings of each word in our dataset by averaging out all sentence embeddings where the word occurs across the dataset. To compute the embeddings, the tokenized input goes through the BERT model and we take the hidden states at the end of the last encoder layer (in our case, BERT-base, i.e. 12 encoder layers) as sentence embeddings. We identify the top words with the highest projection across the *she-he* axis in the region-specific datasets. If we find differences in the top words across regions, it is possible that dominating bias topics vary by

Target words - Attribute words	Region	WEAT
flowers vs insects - pleasant vs unpleasant	Africa	0.312
	Asia	0.869
	North America	0.382
	Europe	0.332
	Oceania	0.660
young people names vs old people names - pleasant vs unpleasant	Africa	0.855
	Asia	0.917
	North America	1.325
	Europe	0.917
	Oceania	0.947
instruments vs weapons - pleasant vs unpleasant	Africa	1.443
	Asia	1.001
	North America	1.202
	Europe	1.21
	Oceania	0.951
European American names vs African American names - pleasant vs unpleasant	Africa	0.008
	Asia	-0.453
	North America	1.29
	Europe	0.617
	Oceania	0.492
Male names vs Female names - career vs family	Africa	1.798
	Asia	1.508
	North America	1.885
	Europe	1.610
	Oceania	1.727
Math vs Arts - Male terms vs Female terms	Africa	1.429
	Asia	1.187
	North America	0.703
	Europe	0.334
	Oceania	1.158
Science vs Arts - Male terms vs Female terms	Africa	1.247
	Asia	0.330
	North America	0.036
	Europe	-0.655
	Oceania	0.725
Mental disease vs Physical disease - temporary vs permanent	Africa	0.835
	Asia	1.201
	North America	0.692
	Europe	1.382
	Oceania	1.620

Table 6: Region-wise WEAT scores across all dimensions specific in WEAT using word2vec

region as well.

D.1 Top words from region-specific BERTs

The top words across the she-he projection space per region are displayed in Figures 3 and 4. We find many differences in the top F (close to she) and M (close to he) words across regions.

Some top F words are soprano, archaeological (Africa); graduate, secretary (Asia); innovative, graphics (Europe); poets, sentiments (NA); and arts, sleep (Oceania). Some top M words are history, leading (Africa); astronomer, commissioners (Asia); honorary, songwriters (Europe); owner, hospital (NA); and wrestlemania, orbits (Oceania). Gender-neutral words such as poets, secretaries, astronomers, commissioners, songwriters, owners, and so on are closer to either the she or he axes. Although comparable to the findings of (Bolukbasi et al., 2016), the variances among regions inspire us to look deeper into the data to arrive at culture-specific bias themes.

E Paired-list for F-M datasets

Here is the list of the 52 pairs used to create the F-M datasets per region:
[monastery, convent], [spokesman, spokeswoman], [Catholic priest, nun], [Dad, Mom], [Men, Women], [councilman, councilwoman], [grandpa, grandma], [grandsons, granddaughters], [prostate cancer, ovarian cancer], [testosterone, estrogen], [uncle, aunt], [wives, husbands], [Father, Mother], [Grandpa, Grandma], [He, She], [boy, girl], [boys, girls], [brother, sister], [brothers, sisters], [businessman, businesswoman], [chairman, chairwoman], [colt, filly], [congressman, congresswoman], [dad, mom], [dads, moms], [dudes, gals], [ex girlfriend, ex boyfriend], [father, mother], [fatherhood, motherhood], [fathers, mothers], [fella, granny], [fraternity, sorority], [gelding, mare], [gentleman, lady], [gentlemen, ladies], [grandfather, grandmother], [grandson, granddaughter], [he, she], [himself, herself], [his, her], [king, queen], [kings, queens], [male, female], [males, females], [man, woman], [men, women], [nephew, niece], [prince, princess], [schoolboy, schoolgirl], [son, daughter], [sons, daughters], [twin brother, twin sister].
Each pair in the above is denoted as a [male, female] pair.

F Implementations details

For training our Bertopic model, we use Google Colab's Tesla T4 GPU, and it takes 15 min to run topic modeling for a region-specific F-M dataset. Region-specific BERTs are run on

```
[ ] # System prompt describes information given to all conversations
    system_prompt = """
    <s>[INST] <<SYS>>
    You are a helpful, respectful and honest assistant for labeling topics.
    <</SYS>>
    """
```

```
[ ] # Example prompt demonstrating the output we are looking for
    example_prompt = """
    I have a topic that contains the following documents:
    - Traditional diets in most cultures were primarily plant-based with a little meat on top, but with the rise of industrial style meat
    production and factory farming, meat has become a staple food.
    - Meat, but especially beef, is the word food in terms of emissions.
    - Eating meat doesn't make you a bad person, not eating meat doesn't make you a good one.

    The topic is described by the following keywords: 'meat, beef, eat, eating, emissions, steak, food, health, processed, chicken'.

    Based on the information about the topic above, please create a short label of this topic. Make sure you to only return the label and nothing more.

    [/INST] Environmental impacts of eating meat
    """
```

```
[ ] # Our main prompt with documents ([DOCUMENTS]) and keywords ([KEYWORDS]) tags
    main_prompt = """
    [INST]
    I have a topic that contains the following documents:
    [DOCUMENTS]

    The topic is described by the following keywords: '[KEYWORDS]'.

    Based on the information about the topic above, please create a short label of this topic. Make sure you to only return the label and nothing more.
    [/INST]
    """
```

Figure 2: Llama 2 prompt

NVIDIA RTX2080 GPUs. Each BERT training experiment takes 1 GPU hour.

F.1 Bertopic

We use Bertopic's default models: SBERT (Reimers and Gurevych, 2019) to contextually embed the dataset, UMAP (McInnes et al., 2020) to perform dimensionality reduction, HDBSCAN (Malzer and Baum, 2020) for clustering to perform topic modeling. We choose the embedding model $BAAI/bge-small-en$ from *Huggingface* (Wolf et al., 2019). We set top_n_words to 10 and `verbose` as True and set the `min_topic_size` to 100 for the Bertopic model. Finally, we use Bertopic's official library to implement the model.

F.2 Llama2

We use Llama to finetune the topics to give shorter labels for each topic. We set the temperature to 0.1, max_new_tokens to 500 and `repetition_penalty` to 1.1. We utilize Bertopic's built-in representation models to use Llama2 in our topic model.

F.3 Region-specific BERT

We use the uncased version BERT (Devlin et al., 2019) for our region-specific BERT model trained for the MLM objective. We use a batch size of 8, a learning rate of $1 \cdot 10^{-4}$, and an AdamW optimizer to train our BERT models for 3 epochs.

G Llama 2 prompt for topic modeling

The prompt scheme for Llama 2 consists of three prompts: (1) System Prompt: a general prompt that describes information given to all conversations, (2) Example Prompt: an example that demonstrates the output we are looking for, and (3) Main Prompt: describes the structure of the main question, that is with a given set of documents and keywords, we ask the model to create a short label for the topic. Fig 2 displays the three prompts as used in the code.

H Topic lists for different regions

Table 7 displays a comprehensive list of topics for female and male groups across all regions.

I Human Validation

Students from a college campus were recruited as annotators in the study. Screenshots of the form are displayed in Fig 5.

J Reproducibility

We open-source our codes, which are uploaded to the submission system. We include commands with hyperparameters in our codes. This would help future work to reproduce our results.

Region	Female	Male
Africa	Credit cards and finances, Royalty and Media, Trading strategies and market analysis, Dating and relationships guides, Parenting and family relationships, Fashionable Ankara Styles, women's lives and successes, online dating	Fashion and Lifestyle, Male enhancement and sexual health, Nollywood actresses and movies, Nigerian politics and government, Essay writing and research, Medical care for children and adults, Journalism and Media Conference, Music industry news and releases, Football league standing and player performances, Academic success and secondary school education, Religious inspiration and spiritual growth, Economic diversification and Socio-economic development
Asia	Hobbies and Interests, Healthy eating habits for children, Social media platforms, Royal wedding plans, Online Dating and Chatting, Adult Services, Gift ideas for Valentine's Day	DC comic characters, Mobile Application, Philippine Politics and Government, Sports and Soccer, Career, Bike enthusiasts, Artists and their work, Youth Soccer Teams, Career in film industry, Political leadership in India, Bollywood actors and films, Religious devotion and spirituality, Phone accessories
Europe	Pets and animal care, Fashion and Style, Education, Obituaries and Genealogy, Luxury sailing, Traveling, Energy and climate change, Family and relationships, Pension and costs, Tech and business operations, Dating, Comfortable hotels, Government transportation policies	Political developments in Northern Ireland, Christian Theology and Practice, Crime and murder investigation, EU Referendum and Ministerial Positions, Criminal Justice System, Israeli politics and International relations, Cancer and medications, UK Government Taxation policies, Art Exhibitions, Political decision and impact on society, Music Gendres and artists, Medical specialties and university training, Political discourse and parliamentary debates
North America	Pets, Cooking: culinary delights and chef recipes, Fashion and style, Family dynamics and relationships, Reading and fiction, Scheduling and dates, Life and legacy of Adolf Hitler, Gender roles and inequality, Education and achievements, Online dating for singles, Luxury handbags, Footwear and Apparel brands, Essay writing and literature	Civil War and history, Middle East conflict and political tensions, Movies and filmmaking, Political leadership and party dynamics in Bermuda, Rock Music and songwriting, Wartime aviation adventures, Religion and Spirituality, Reproductive health, Reinsurance and Capital markets, Nike shoes and fashion, Cape Cod news, NHL players
Oceania	Cooking and culinary delights, Romance, Weight loss and nutrition for women, Water travel experience, Woodworking plans and projects, Time management and productivity, Inspiring stories and books for alleges, Sexual violence and abuse, Car insurance, Exercises for hormone development, kid's furniture and decor	Harry Potter adventures, Art and Photography, Superheroes and their Universes, Music recording and Artists, Football in Vanuatu, Pet care and veterinary services, Building and designing boats, Religious beliefs and figures, Fashion, Classic movie stars, Men's hairstyle and fashion, Male sexual health and supplements

Table 7: Region-wise topics for female and male.

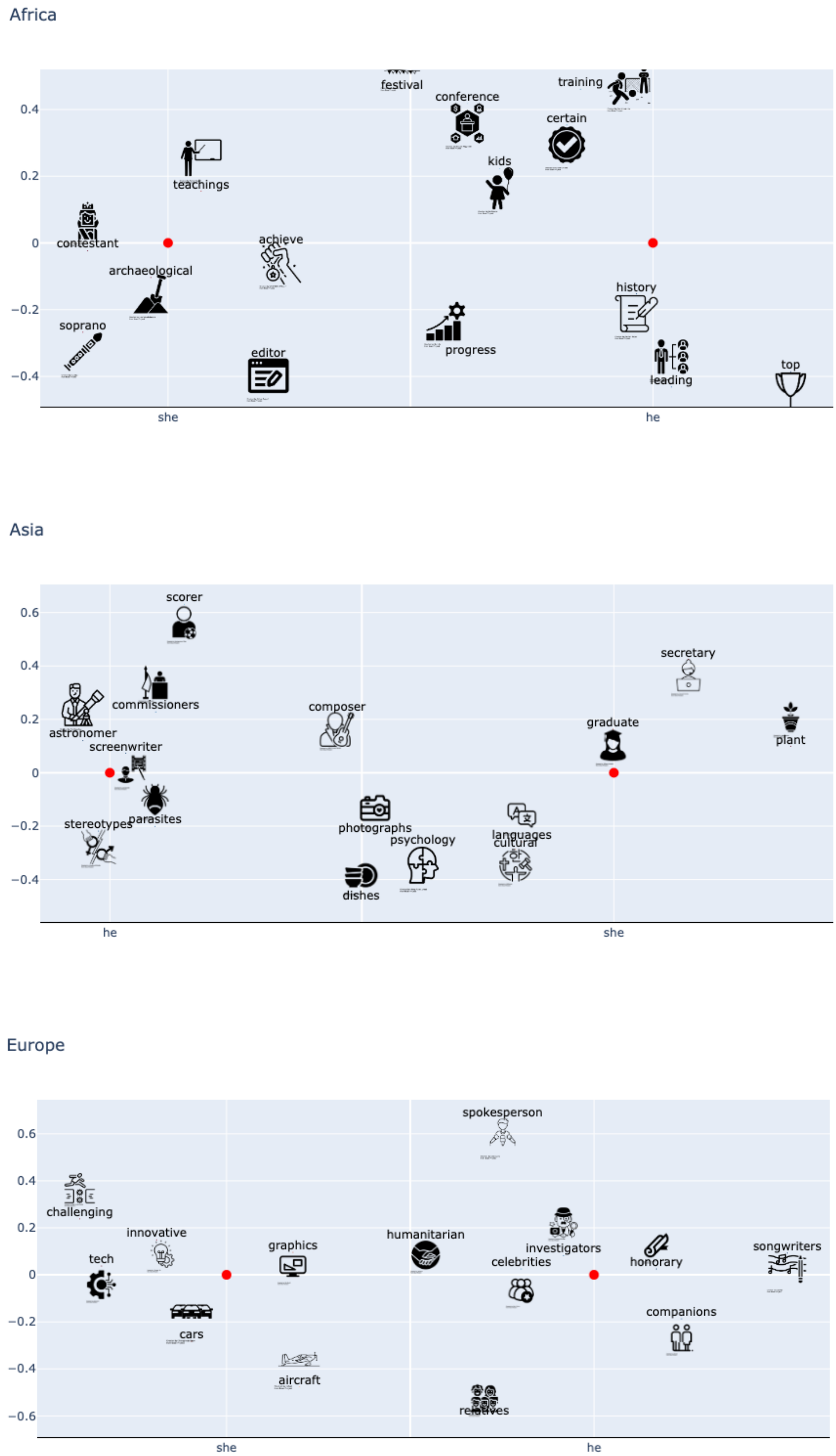
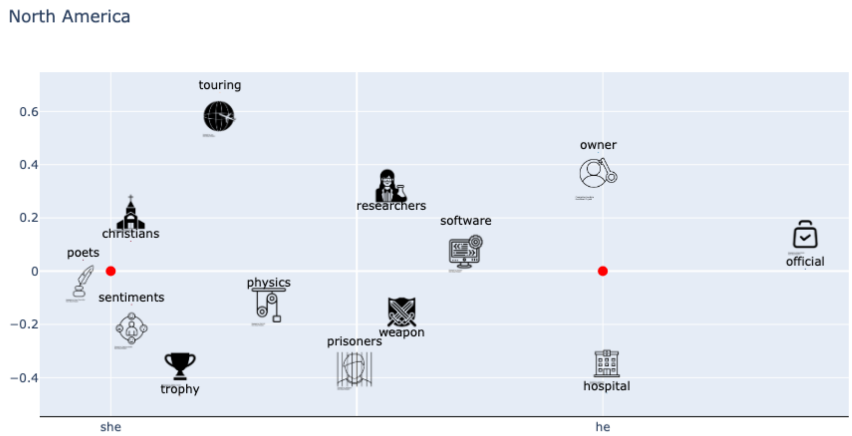


Figure 3: Top words for each region(Africa, Asia, and Europe) using region-specific BERTs



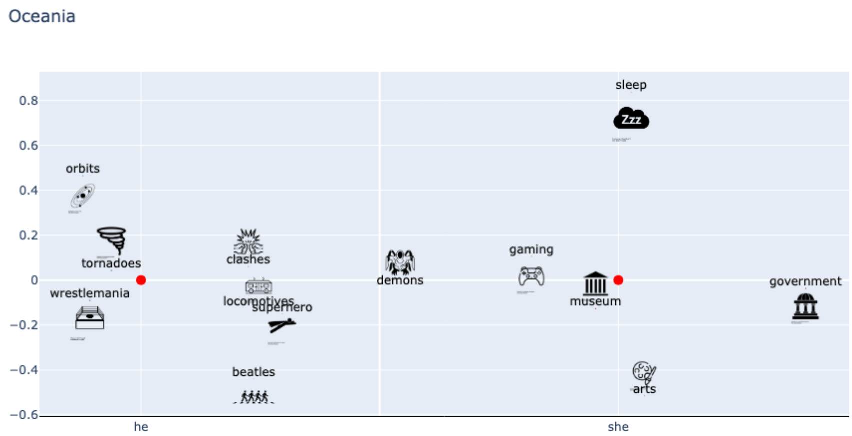
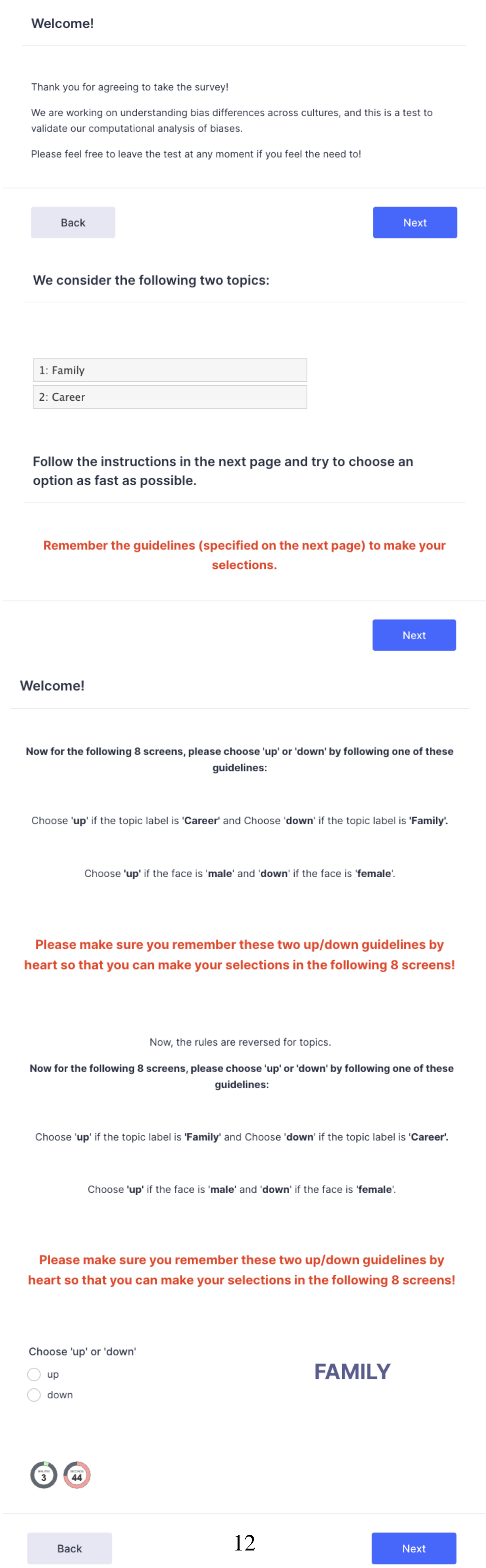


Figure 4: Top words for each region(North America, Oceania) using region-specific BERTs



Welcome!

Thank you for agreeing to take the survey!

We are working on understanding bias differences across cultures, and this is a test to validate our computational analysis of biases.

Please feel free to leave the test at any moment if you feel the need to!

Back | Next

We consider the following two topics:

1: Family

2: Career

Follow the instructions in the next page and try to choose an option as fast as possible.

Remember the guidelines (specified on the next page) to make your selections.

Next

Welcome!

Now for the following 8 screens, please choose 'up' or 'down' by following one of these guidelines:

Choose '**up**' if the topic label is '**Career**' and Choose '**down**' if the topic label is '**Family**'.

Choose '**up**' if the face is '**male**' and '**down**' if the face is '**female**'.

Please make sure you remember these two up/down guidelines by heart so that you can make your selections in the following 8 screens!

Now, the rules are reversed for topics.

Now for the following 8 screens, please choose 'up' or 'down' by following one of these guidelines:

Choose '**up**' if the topic label is '**Family**' and Choose '**down**' if the topic label is '**Career**'.

Choose '**up**' if the face is '**male**' and '**down**' if the face is '**female**'.

Please make sure you remember these two up/down guidelines by heart so that you can make your selections in the following 8 screens!

Choose 'up' or 'down'

up

down

FAMILY

3 44

Back | Next

Figure 5: Annotation Form Screenshots