MMERB: A Multimodal Dataset for Measuring Marginalized Ethnicity Reporting Bias

Tom Södahl Bladsjö
Course: LT2314

Abstract

Human reporting bias, or the human tendency to omit unnecessary or obvious information while mentioning things that are considered relevant or surprising, is a phenomenon that affects the representations learned by language models. I present the MMERB dataset, to my knowledge the first dataset for measuring model reporting bias with regards to marginalized identities. MMERB contains two sets of images, depicting non-white and white people respectively, where each image has two contrasting captions; one that mentions the ethnicity of the person in the image, and one that does not. Thus, the dataset is designed to measure differences in how likely a model is to mention ethnicity in image captions, based on whether or not the people depicted in the images are white. The goal is to explore the relationship between reporting bias and racial bias, and to provide insights into what perspectives, and what norms, are being encoded in image captioning models.

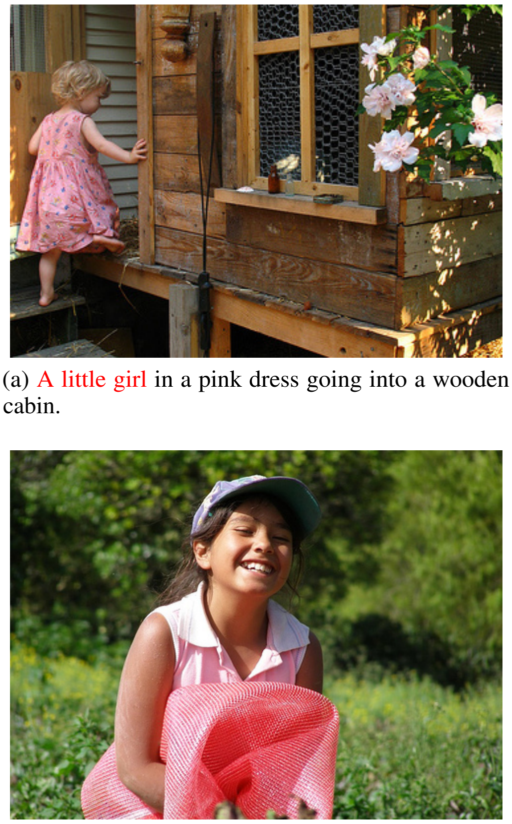
1 Introduction

As known by pragmaticians and NLP researchers alike, human language is notoriously underspecified. When making an utterance, a speaker (or writer, as it may be) needs to filter the available information and include only what is relevant and necessary for the communicative purpose (Grice, 1975). This fact of communication leads to something that is known as *human reporting bias*:

> The frequency with which people write about actions, outcomes, or properties is not a reflection of real-world frequencies or the degree to which a property is characteristic of a class of individuals.
> (Chang et al., 2019)

For example, Gordon and Durme (2013) note that the word *murdered* occurs almost 4 times as often as the word *breathed* in the Google Web 1T n-gram data (Brants and Franz, 2006). This, of course, does not mean that we murder more often than we breathe. It means that people are assumed to breathe, so we normally do not feel the need to mention it. A murder, on the other hand, is an unexpected event that will seldom pass unmentioned. This fact of human language has implications for models trained on natural language text data, since the distributions they learn are likely to correlate more strongly with those reported in text than those found in the real world (see e.g. Paik et al., 2021). The tendency to mention the unexpected and omit the obvious also affects how humans describe each other. Consider the two images and annotations in Figure 1, taken from the Flickr8k dataset (Hodosh et al., 2013). The annotator who wrote the caption for Figure 1a did not remark on the skin color or ethnicity of the child in the image, likely because they did not find it relevant or unexpected enough to be worth mentioning. The child in Figure 1b, on the other hand, is described as "Asian". Clearly, in this case the annotator considered her perceived ethnicity to be relevant or unexpected enough to warrant mentioning. She is thereby being defined by her ethnicity in a way that the white child is not. In fact, if we look at the most common skin colors and ethnicities mentioned in the Flickr8k and MS COCO (Lin et al., 2014) datasets, mentions of non-white ethnicities are noticeably more common than mentions of whiteness (see Table 1[1]). A visual inspection of the datasets, on the other hand, indicates that the majority of the people depicted in the images are in fact white (see §3.3.2). It is clear that these datasets represent a perspective where being white is the norm – not only in the images themselves, but also in the captions. Models trained on data like this are likely to encode a similar perspective, or even amplify the effects. The purpose of

[1] The counts were obtained using the same regular expression patterns used when creating the dataset. See §3.3.1 for more details.



(a) A little girl in a pink dress going into a wooden cabin.

(b) An Asian girl in a pink dress is smiling whilst out in the countryside.

Figure 1: Two examples from the Flickr8k dataset.

the MMERB dataset is to provide a way to test if, and to what extent, models encode a perspective where whiteness is considered the norm. This is achieved by comparing minimal pairs of image captions, where the only difference is whether or not ethnicity is mentioned, and then comparing these differences between different groups.

2 Background and Related Work

Earlier work on reporting bias (e.g. Gordon and Durme, 2013; Paik et al., 2021; Hagström and Johansson, 2022; Misra et al., 2016; Shwartz and Choi, 2020) has mainly focused on the phenomenon as a whole and its implications for learning common sense knowledge from text. To my knowledge no work has this far been done on how reporting bias affects model behavior with regards to marginalized groups. Conversely, much previous research in bias and fairness has focused on harmful stereotypes encoded in language models (Caliskan et al., 2017; Bolukbasi et al., 2016, among others), often using variations on the Implicit Association Test (Greenwald et al., 1998) from psychology, adapted for testing language models. Other work on bias and fairness in language models includes Felkner et al. (2023), who present a benchmark dataset for testing models for anti-queer bias. While their work also focuses on stereotypes and not reporting bias, it builds on similar basic assumptions about bias in language models; namely, that models encode the social perspective of the data they have been trained on. In fact, Felkner et al. (2023) demonstrate that by finetuning models on data produced by the affected communities themselves, they can shift the perspective encoded in the model and thus reduce the observed bias. By creating the MMERB dataset, I aim to begin to fill a gap in the existing research by providing a method to explore the relationship between reporting bias and racial bias.

Expression	Count	Expression	Count
black man	35	asian man	89
asian woman	28	asian woman	84
asian girl	28	black man	50
asian man	24	asian girl	32
asian girls	20	asian people	24
asian women	17	white man	23
asian boy	17	asian women	19
asian children	10	asian men	18
dark skinned man	9	white woman	15
white man	8	indian man	13

Flickr8k MS COCO

Table 1: The number of times each expression appears in each dataset, ordered by frequency. Only the ten most common expressions are included.

3 Data

The MMERB dataset consists of a total of 1313 images, of which 633 depict people interpreted as white and 680 depict people interpreted as non-white. Each image comes with two contrasting captions; one where the ethnicity of the person (or persons) in the image is mentioned, and one where it is not. When testing a model on this dataset, each image will be viewed twice (once with each version of the caption), resulting in a total of 2626 examples. The captions and metadata are divided into four separate CSV files, representing the four settings to be compared: Non-white people being described with ethnicity, non-white people being described without ethnicity, white people being described with ethnicity and white people being described without ethnicity. More details, as well as download- and usage instructions can be found on Github[2].

3.1 A Note on Terminology and Categorization

The word *ethnicity* is one that has been used in a lot of different, and sometimes conflicting, ways (Chandra, 2006). In this work, it is used very broadly to mean a set of visible attributes that can cause a person to be subject to, or exempt from, racist discrimination. The intuition behind this is that when a person in an image is described as "Black", the same basic mechanism is at work as when a person is described as "Asian", even though one expression refers to skin color and the other to perceived area of origin. For example, the annotator who described the child in Figure 1b as "Asian" likely chose to do so based on visual attributes such as skin color.

When categorizing examples in this dataset, I have chosen to limit it to two very broad categories: *non-white* and *white*. One reason for this is that the dataset is relatively small, and a more fine-grained categorization would result in very few examples per group. Another reason is that there is no one "true" way to categorize ethnicities, and to do so in a way that is connected in a meaningful way to how these groups experience racism differently would require a firm grounding in theory and the personal experiences of members of these groups, neither of which I have access to at the moment. Therefore, I chose a broad categorization based on my own intuitions about who is likely to be excempt from racism (white people). This makes MMERB a useful dataset for detecting differences between how models treat white people as opposed to other groups, but it will not capture differences in how the many non-white subgroups are affected. There is a considerable risk that this choice of broad categorization hides important differences between groups, and potentially disguises greater inequities that are specific to certain subgrups (Castillo and Gillborn, 2021).

3.2 Original Data Sources

The images and original captions in MMERB come from two sources: the Flickr8k dataset (Hodosh et al., 2013) and the MS COCO dataset (Lin et al., 2014)[3]. The images in both dataset were originally collected from Flickr and annotated by crowd workers at Amazon Mechanical Turk.

3.3 Creating the Dataset

3.3.1 Filtering the Captions

In order to filter the examples containing descriptions (and, consequently, images) of people in the original datasets, the original captions were filtered based on three regular expressions: one matching captions containing mentions of people, one matching captions explicitly describing people as non-white, and one matching captions explicitly describing people as white.[4] The filtered examples were then sorted according to if they contained ethnicity words and which group those words described (white or non-white). Images whose captions matched both regular expressions (that is, that mentioned both white and non-white people) were excluded. While the original datasets contain multiple captions for each image, only one caption per image was kept when creating MMERB. This resulted in three stacks of examples for each dataset: one with people described as white, one with people described as non-white, and one where ethnicity was not mentioned (see Table 2).

3.3.2 Balancing the Groups

As can be seen from the results of the first filtering, captions describing people as non-white were much more common than captions describing people as white. While this does support the hypothesis that ethnicity-based reporting bias is present in the datasets, it also introduces some difficulty in obtaining a balanced dataset with regards to the groups being tested. In order to solve this, an additional 1400 images (400 from Flickr8k and 600 from MS COCO) were inspected and manually sorted as depicting either non-white or white people[5] (images

[2] https://github.com/TomBladsjo/LT-Resources-project

[3] https://cocodataset.org/

[4] The specific search terms to use were based on preliminary exploration of the Flickr8k captions, where I collected a list of all words immediately preceding words for people, and then manually filtered all words describing ethnicity. For non-white ethnicities, all words that appeared more than 5 times in the original corpus were included, while all words for white people (including e.g. expressions like "fair-skinned" and "caucasian") were included. This is because captions describing people as non-white were significantly more common than captions describing people as white.

[5] The decision to manually sort images as depicting white or non-white people is by no means uncomplicated, since it entails labeling people according to perceived ethnicity regardless of how they self-identify. See Limitations and Ethics Statement for more in-depth discussion.

Dataset	Regex	n examples
Flickr8k	Non-white	238
	White	14
	No mention	24234
MS COCO	Non-white	442
	White	53
	No mention	209965

Table 2: **Result of first filtering.** The number of examples from each group extracted from the datasets at the first filtering step. As can be seen, while it is much more common for annotators to mention ethnicity when the person in the image is non-white, the most common thing in general is to not mention ethnicity at all.

depicting large groups of people or where the ethnicity of the people depicted was unclear were excluded). This resulted in an additional 269 images of white people from the Flickr8k dataset and 297 from the MS COCO dataset (the number of images sorted as clearly depicting non-white people was 47 for Flickr8k and 85 for MS COCO).

3.3.3 Creating the Contrasting Examples

Starting from the existing captions from the original datasets, contrasting examples were created by removing the ethnicity word (using the same regular expressions that were used for filtering) from the descriptions mentioning ethnicity, and adding the word "white" to the images of white people where the original caption did not mention ethnicity. If the original description included the indefinite article (a/an), this was modified when needed to avoid introducing grammatical errors ("An African-American woman" → "A woman"). A notebook containing the full code used when creating the dataset is available on Github.

3.4 Usage and Comparison to Existing Datasets

This dataset is similar to various other benchmark datasets that use minimal sentence pairs to test for variuos linguistic phenomena, whether bias related or otherwise (e.g Felkner et al., 2023; Warstadt et al., 2020). However, it differs from these existing datasets 1) in that it measures reporting bias with regards to marginalized groups (which to my knowledge has not been done before), and 2) in that it incorporates the visual modality. Another difference compared to existing minimal pair based benchmarks is that, because of the nature of the phenomenon (the difference is in whether or not something is being mentioned), the resulting sentence pairs will inevitably have different sequence length. Care must therefore be taken when using this dataset to not accidentally compare apples and pears. For example, if using metrics based on sequence probability, keep in mind that sequence length affects the probability of the sequence. Thus, the minimal pairs are not directly comparable to each other. On the other hand, the set of differences in probability for each pair can be compared across groups to obtain a measure of the difference in model performance for the different groups. See (my course paper for LT2318) for a usage example. Another related dataset is the CoDa dataset (Paik et al., 2021), which is used to compare the distribution of color captured by language models to how humans perceive color. It is similar to MMERB in that it specifically deals with reporting bias, and that it includes thevisual modality. Unlike MMERB, however, CoDa uses hand-crafted templates instead of examples created from naturally occurring language. Finally, while MMERB is primarily aimed at generative image captioning models, it can also be used to test other multimodal models, such as image-text matching models, and as a probing dataset for masked language-and-vision models.

4 Conclusion and Future Work

I have presented MMERB, a multimodal dataset for testing reporting bias with regards to marginalized ethnicities in language models. Specifically, this dataset is designed to measure differences in how likely a model is to mention ethnicity in image captions, based on whether or not the people depicted in the images are white. I believe that exploring these differences can provide valuable insights into what perspectives, and what norms, are being encoded in models – in other words, whose language we are modeling when we model language. Future work should aim to expand this approach to include other marginalized attributes, such as, for example, gender, queerness, dis- ability and religion. It would also be useful to investigate the effects of intersectionality on reporting bias (how does belonging to multiple marginalized groups affect model behavior), as well as develop a more fine-grained and theory-motivated categorization of the groups affected.

Limitations

The methods used when developing the MMERB dataset have a number of limitations. For one, since many images were collected based on captions (see §3.3.1), there may be context-specific reasons ethnicity was mentioned when it was.[6] Another limitation, as discussed in §3.1, is that the categorization of groups in this this dataset is extremely broad and somewhat naïve. It is possible that this categorization distorts results, particularly by hiding differences between subgroups.

Ethics Statement

This dataset contains sensitive data in the sense that it groups images of people by perceived ethnicity. The images have been sorted partly based on labels attributed to them by annotators based on visual appearance, and partly by me based on visual appearance. This was done without access to information about how the people in the images identify themselves. While this is in some sense inherent to the phenomenon being investigated (it deals with how people are described by others rather than how they would choose to describe themselves), it should be noted that it is problematic that the people depicted in the images had no say in how they were categorized, and that it is likely that this dataset contains incorrectly labeled images. It should also be noted that technically, this dataset could be used for malevolent purposes, such as intentionally training a model to behave differently for different groups based on skin color. On the whole, grouping people as datapoints based on attributes for which they are likely to face discrimination is problematic and potentially risky. Researchers using this dataset should keep this in mind and tread carefully.

References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1. Philadelphia: Linguistic Data Consortium. LDC2006T13. Web Download.

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

Wendy Castillo and David Gillborn. 2021. How to "QuantCrit:" practices and questions for education data researchers and users. (EdWorkingPaper: 22-546). Retrieved from Annenberg Institute at Brown University: `https://doi.org/10.26300/v5kh-dd65`.

Kanchan Chandra. 2006. What is ethnic identity and does it matter? *Annual Review of Political Science*, 9(1):397–424.

Kai-Wei Chang, Vicente Ordonez, Margaret Mitchell, and Vinodkumar Prabhakaran. 2019. Tutorial: Bias and fairness in natural language processing. Recorded presentation at EMNLP 2019, hosted at UCLA NLP `http://web.cs.ucla.edu/~kwchang/talks/emnlp19-fairnlp/`.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A community-in-the-loop benchmark for anti-LGBTQ+ bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140, Toronto, Canada. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Durme. 2013. Reporting bias and knowledge acquisition. *AKBC 2013 - Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, Co-located with CIKM 2013*, pages 25–30.

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74 6:1464–80.

H. P. Grice. 1975. *Logic and Conversation*, pages 41 – 58. Brill, Leiden, The Netherlands.

Lovisa Hagström and Richard Johansson. 2022. What do models learn from training on more than text? measuring visual commonsense knowledge. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 252–261, Dublin, Ireland. Association for Computational Linguistics.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

[6] For example, in the caption "A white man with dreadlocks", it is possible that the reason the annotator mentioned ethnicity was that they would not normally associate dreadlocks with white people.

Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. 2021. The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 823–835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.