The Effect of Debiasing Techniques in Large Language Models for Gender and Racial Bias in Occupation

Jennifer Mickel
jamickel@utexas.edu
The University of Texas at Austin
Austin, Texas, USA

ABSTRACT

Bias between gender and occupation has been found in many language models. In society, one's experience can differ by both their race and gender. Likewise, it is possible that biases between race, gender, and occupation exist. Cohere's language model was used to determine whether a pronoun was referring to an occupation or a participant. The results showcased that Cohere had utilized debiasing and fairness techniques that caused the data that resulted from querying their model to not have any bias in occupation towards males. But, it was biased towards females, and it seemed that Cohere did not debias some occupations with a higher percentage of females.

KEYWORDS

bias, large language models, natural langague processing, fairness

ACM Reference Format:
Jennifer Mickel. 2022. The Effect of Debiasing Techniques in Large Language Models for Gender and Racial Bias in Occupation. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The rise in the development and usage of language models without significant efforts to ensure these models are developed in a manner without unfair bias has led to the embedding of bias in large language models. Bender et al. discuss how the development of large language models negatively impact the environment and can contribute to the oppression faced by marginalized groups [1]. De-Arteaga et al. showcase how classifiers representing words using the bag-of-words, word embedding, and DNN methodologies demonstrate bias between occupation and gender regardless of if models were trained with gendered pronouns [4].

Buolamwini and Gebru showcased disparities in accuracy metrics for facial recognition software developed by Face++, IBM, and Microsoft. They found that these facial recognition systems worked better for males than females and for people with lighter skin than darker skin. Although the system was less accurate for lighter-skinned females and darker-skinned males, the system performed the worst for darker-skinned females [2]. This supports the theory of intersectionality introduced by Kimberle Crenshaw in 1989 [3]. Intersectionality is the idea that a person's experience, whether positive or negative, is influenced by their identities. For example, a Black woman may experience oppression similar to the oppression experienced by white women, Black men, but also, oppression unique to Black women that neither Black men nor white women experience. This can be seen in Bouwalmi and Gebru's work because the accuracy of the facial recognition systems for darker-skinned women is significantly worse than the accuracy for lighter-skinned women and darker-skinned men.

Differences in biases and oppressions faced by people of different identities are present within language. For example, the usage of the word "slave" to describe people who were enslaved decreases the personhood and humanity of these people. Recently, there has been a push in academia to refer to someone who experienced enslavement as an "enslaved person" rather than a "slave" [8]. Despite this shift, the language large language models train on may refer to enslaved people as "slaves" which can dehumanize these people. The language large language models are trained on also reflects societal biases, for the language the model sees is likely going to reflect the biases of the people who produced it and the biases behind any data the language model may see. These biases can present themselves in occupations and careers. For example, "women account for 25% of those working in computer occupations" [5]. [1] This underrepresentation of women within computer occupations leads language referring to people in computer occupations to primarily refer to men since they comprise a greater percentage. A large language model trained on this data may associate computer occupations such as "software engineer" or "data scientist" with male pronouns and may predict that whenever a "software engineer" or "data scientist" is mentioned, these positions are referring to men. Biases in occupation can also be found in female-dominated occupations. For example, women comprise 76.6% of public school teachers, and this may lead a model to automatically associate female pronouns with "teacher" [6]. Gender biases can be harmful because they can lead the language model to make assumptions about the characteristics of a particular occupation and may harm people by enforcing gender occupation stereotypes.

Although work has been done to locate bias between gender and occupation [4][7], significantly less work has been done to locate bias between gender, race, and occupation. This paper seeks to tackle this challenge.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] This study did not mention the percentage of non-binary folks in computer occupations or STEM

2 BACKGROUND INFO

Rudinger et al. created the Winogender schemas dataset which of sentences that contain an occupation, a participant, and a gendered pronoun. The task the system must do is determine whether the gendered pronoun is referring to the occupation or the participant. Rudinger et al. tested their dataset on three systems and found all of them were biased to predict male pronouns for occupation and that gendered pronoun resolution by occupations positively correlated with the percentage of females by occupation in text [7].

3 METHODOLOGY

To locate bias between gender, occupation, and race, I expanded the Winogender schemas dataset created by Rudinger et al. to include race. The gender pronouns I used were she/her/hers, he/him/his, and they/them/theirs. The categories of race and ethnicity I utilized were None, White, Black, Asian, Hispanic, and Indigenous. In the None case, no race was provided to accompany the occupation and/or race. Although these racial and ethnic categories do not encompass all of the racial and ethnic identities a person could have, they were simplified to locate correlations in bias between race, gender, and occupation. This was added such that if a pre-existing sentence such as "The paramedic performed CPR on the passenger even they knew it was too late" sentences such as "The Indigenous paramedic performed CPR on the Asian passenger even though they knew it was too late" would be created from the Winogender schemas dataset to include race and ethnicity.

From there, I queried the Cohere large off-the-shelf classification model with the data created by modifying the Winogender schemas dataset. The Cohere model was given 72 examples consisting of an equal amount of female, male, and neutral pronouns that alternated (male, female, neutral) to prevent bias from entering the model through the examples given. These examples contained no reference to race. From there, the model was queried with all of the examples, and the results were evaluated.

I proceeded to calculate the true and false occupation rates, the true and false participant rates, the occupation and participant PPVs (predictive precision value), the occupation and participant error rates, and the accuracy rate. These metrics are evaluated on each combination of race and gender with the addition of 'All'. 'All' refers to the summation of the results of all the genders i.e. the error rate for 'Hispanic All' would look at the total error rate across Hispanic male, female, and neutral pronouns.

4 RESULTS

Given that the exact same sentences were used for each combination of races and genders, an unbiased system would perform the same regardless of the race or gender presented. As seen in Figure 1, this is not the case. Accuracy and other performance metrics vary given the race and gender of the participant and occupation. Surprisingly, the true occupation rate and the false occupation rate are higher for females than males. This contradicts the findings found in Rudinger et al. because this implies that occupation is predicted more often for females than males. Surprisingly, the highest false occupation rates seem to exist for Hispanic and Indigenous women. Given societal biases and systems of oppression, I would accept males to have higher true and false occupation rates and that white and Asian people would have higher true and false occupation rates which contradict the data. The language model may not have seen 'Hispanic' or 'Indigenous' in front of occupations often unless that particular person was successful in their career. This is seen in society for the media and people tend to refer to people of color by their race and/or ethnicity. For example, a white filmmaker would be referred to as a filmmaker, yet a Hispanic filmmaker would be referred to as a Hispanic filmmaker. The occupation PPV (positive predictive value) which measures how likely a pronoun predicted to refer to occupation is correct is lower for females than males. This is interesting because, in a system biased towards females, I would accept PPV to be high suggesting that the system performs well for females. Since this is not the case, it seems that Cohere attempted to correct for bias within their large language model against females, but this led the model to be more likely to predict occupation when it sees a female pronoun resulting in a lower PPV and higher occupation error rate.

Racial and ethnic bias within the model seems to exist because the occupation PPV is lowest for Hispanic females and Indigenous females meaning that when the model predicts occupation for Hispanic females or Indigenous females it is less accurate than when the model predicts occupation for other demographics and genders.

Interestingly, the true and false participant rate is highest for males and lowest for females which contradicts the findings of Rudinger et al. because Rudinger et al. found that participation was more likely to be predicted for females and males. Similar to the lower PPV for occupation and females, males have a lower PPV meaning that when the model predicts participant for male pronouns, it is less likely to be accurate than it is for women.

Although variation exists in the performance across races, it doesn't seem to be significant which seems to suggest the racial bias that may exist within the model does not seem to impact the aggregate result of occupation and gender. Despite this, it seems the combination of race and gender does impact the model's performance because the error rate for occupation and Hispanic females is higher than for females of other races. This supports intersectionality because it showcases how certain combinations of gender and racial identity can lead to discrepancies in the language model's performance.

Although variation exists in the performance of the model across gender, performance discrepancies are greatest between male and female pronouns. Neutral pronouns perform similarly to 'All' (the aggregate of male, female, and neutral pronouns). This seems to imply that the model treats neutral pronouns as an average between male and female pronouns.

The model provided confidence levels when making a prediction and, as expected, increased the confidence level threshold led to greater model accuracy.

5 ANALYSIS

Upon analyzing the data by observing the model's predictions for each gender pronoun and race it seems Cohere chose a set of occupations to ensure their model would not be biased against women. Such occupations include "programmer", "engineer", etc. Despite this effort, it seems occupations where women comprise a large percentage did not receive such efforts. For example, there is a

higher percentage of females employed as teachers than males, and this bias remains in the model.

Figure 2 showcases how the percent difference in occupation between females and males correlates with the percentage of women in certain occupations, and the results differ drastically from Rudinger et al. There does not seem to be an obvious line of best fit and the occupations seem to vary in the percent difference in occupation between female and male in a non-orderly manner. This suggests that Cohere chose occupations to ensure the number of women predicted is equal to the number of men. Interestingly, some of the occupations with a low percentage of women are biased toward women. This result may be due to Cohere's efforts to debias their model, leading to an overcorrection that causes this bias.

Although it is evident Cohere employed debiasing and/or fairness techniques in an effort to make their system fairer it did not completely remove all bias within gender, race, and occupation. Although some occupations seem to predict

6 IMPACT

Cohere claims to have made efforts to debias their language model [10], and although they do not seem to have removed all bias and unfairness capable of being located in this task, they have removed some bias from the model. This is evident because Cohere's language model performs better on the Winogender schema dataset than the previous models Rudinger et al. tested the dataset on. Despite these efforts, Cohere's language model continues to be biased toward women in female-dominated occupations. Historically, female-dominated occupations have less prestige than male-dominated occupations, so the harm of predicting male-dominated occupations as male seems to be larger than the harm of predicting female-dominated occupations as female. Although predicting the pronouns associated with female-dominated occupations as female can be harmful because it means the model may associate these occupations with women when it should not make these associations. Furthermore, this bias towards female pronouns in female-dominated occupations may cause the model to perceive men as holding these occupations as female which would be inaccurate. For example, imagine a scenario in which a male teacher is communicating with a chatbot. In their conversation, it comes up that he is a teacher, but he does not specify his pronouns. If the model used by the chatbot is biased towards women, the chatbot might refer to the teacher as 'she' even though those are not his pronouns. Thus, even though the societal implications of having a model predict female pronouns for female-dominated occupations may be lower than a model predicting male pronouns for male-dominated occupations, the potential for harm still exists.

In some instances, Cohere seems to overcorrect for gender bias found in male-dominated occupations. Although this isn't ideal because ideally, a model would not have discrepancies in gender regarding occupation, having a model that predicts female for some male-dominated occupations could help combat that bias within society.

7 FUTURE WORK

Perform the task of determining whether Cohere's language models have bias between race, gender, and occupation using a dataset that is more challenging to perform well on. Utilizing a more challenging dataset may illuminate biases within the model that did not appear because of the simplicity of the task I gave.

Future work could also include expanding the number of occupations in the dataset and conducting more research to determine which occupations have disparities in what pronoun the model assigns. Given this additional understanding of how occupation and gender correlate, it may become more clear why these discrepancies in how the model assigns pronouns exist.

8 LIMITATIONS

A limitation of this work is it is possible the despite utilizing methods to mitigate the training examples from biasing the data, the manner in which I presented the training examples to the model affected the model's performance on the task.

Another limitation of this type of work in language models is it is very hard to rule out bias within the model because it is hard to have an understanding of the model's associations. Thus, even in instances where the model does not seem to be biased and is performing in a manner we expect an unbiased model would perform, we still cannot rule out that the model has bias.

REFERENCES

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[2] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[3] Kimberlé Williams Crenshaw. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics.

[4] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 120–128. https://doi.org/10.1145/3287560.3287572

[5] Richard Fry, Brian Kennedy, and Cary Funk. 2021. STEM Jobs See Uneven Progress in Increasing Gender, Racial and Ethnic Diversity. *Pew Research Center* (April 2021).

[6] Richard M. Ingersoll, Elizabeth Merrill, Daniel Stuckey, and Gregory Collins. 2018. Seven Trends: The T ends: The Transformation of the T ansformation of the Teaching F eaching Force – Updated October 2018. *Consortium for Policy Research in Education* (October 2018), 1–28.

[7] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, New Orleans, Louisiana.

[8] Katy Waldman. 2015. Slave or Enslaved Person? *Slate* (May 2015).

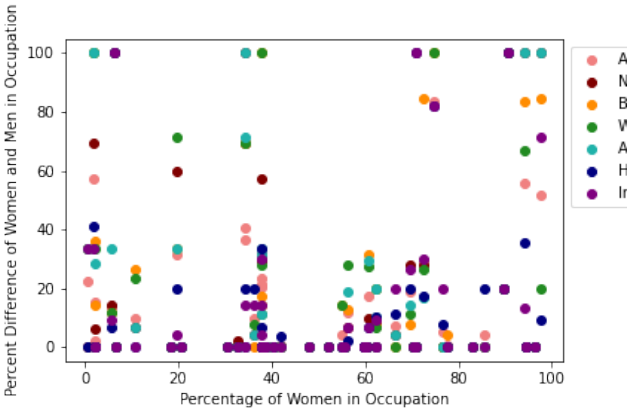

Figure 2: This plot shows how gender bias in occupational gender statistics from the U.S. Bureau of Labor Statistics correlates with the extent female pronouns are preferred over male pronouns i.e. % Female - % Male. A score of 100% correlates with maximum female bias and a score of -100% correlates with maximum male bias.

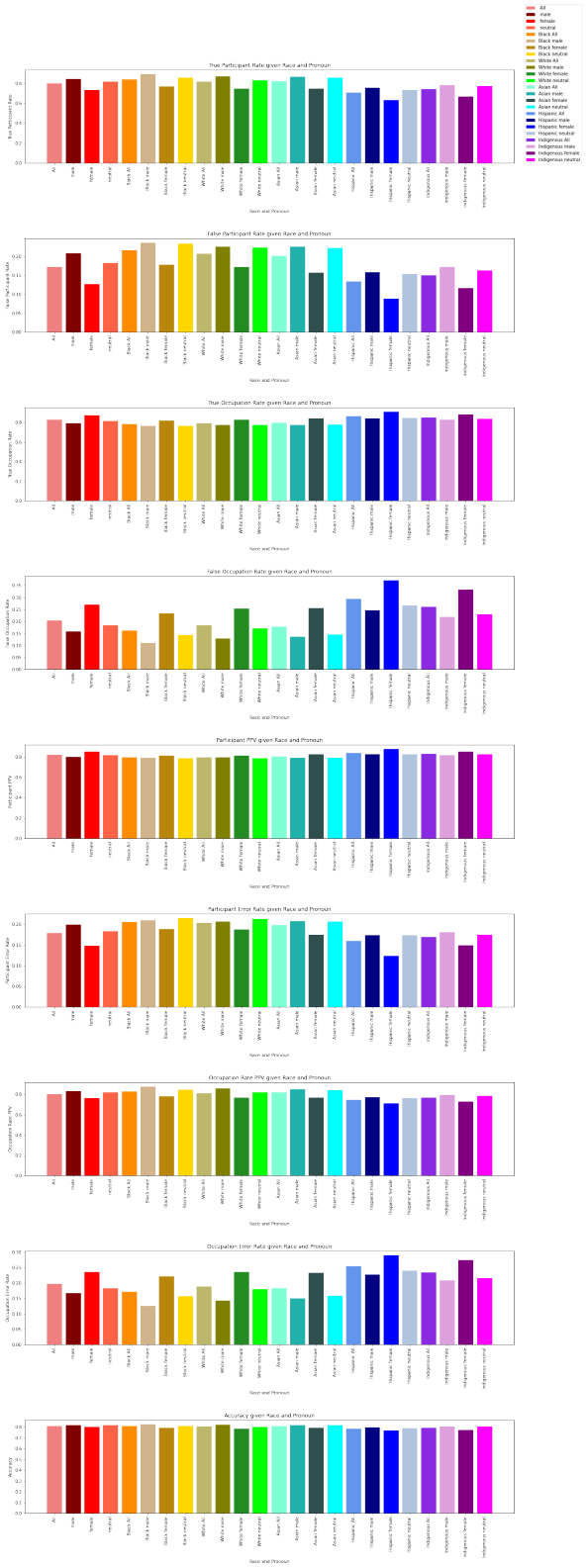

Figure 1: These graphs showcase how various accuracy and fairness metrics differ depending on race and gender.

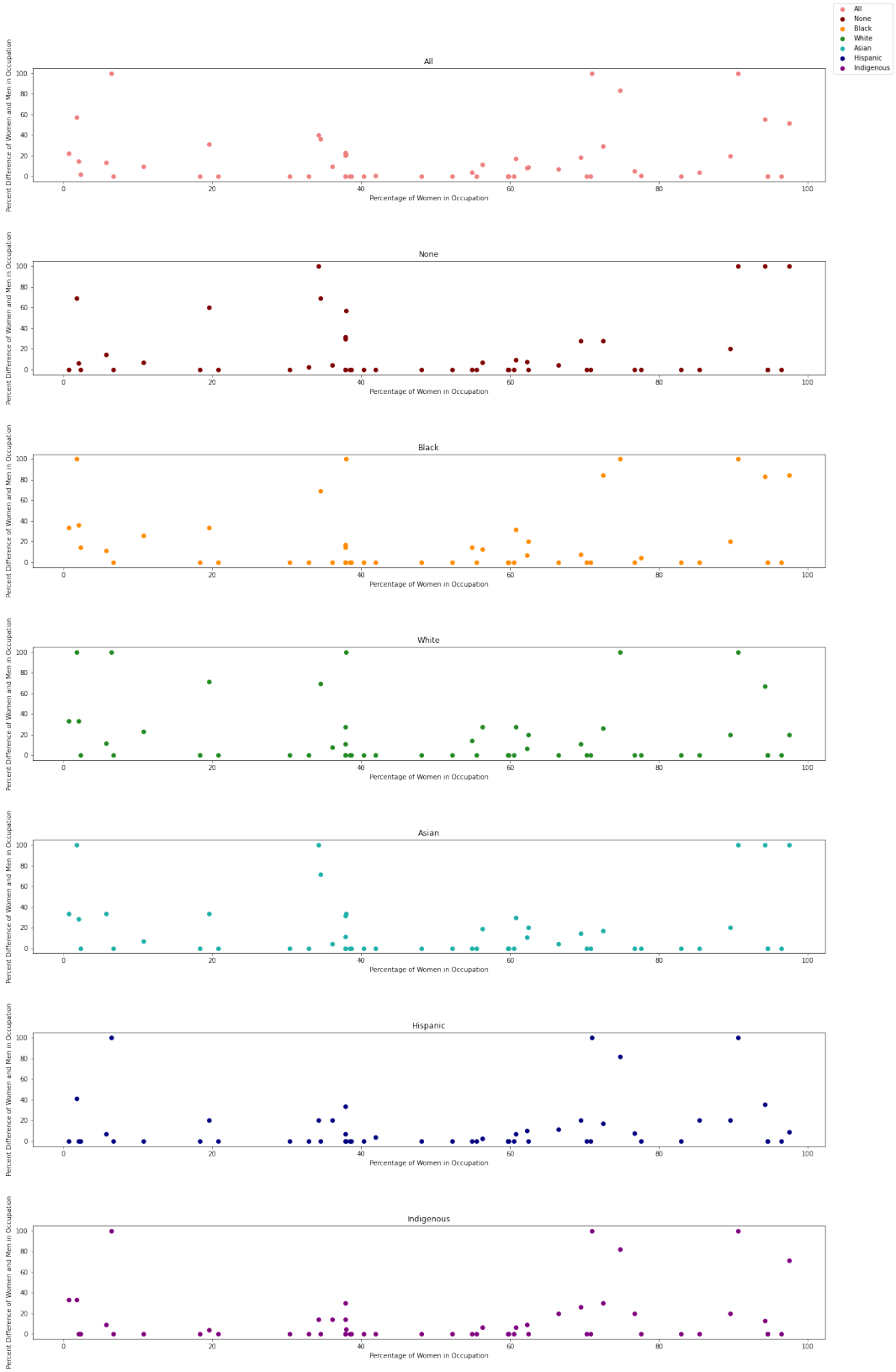

Figure 3: These plots show how gender bias in occupational gender statistics from the U.S. Bureau of Labor Statistics correlates with the extent female pronouns are preferred over male pronouns i.e. % Female - % Male. A score of 100% correlates with maximum female bias and a score of -100% correlates with maximum male bias.