Identifying and Neutralizing Gender Bias from Text

Maya Bedge
Department of Computer Science
Stanford University
mbedge@stanford.edu

Dante Danelian
Department of Computer Science
Stanford University
danelian@stanford.edu

Liam Smith
Department of Computer Science
Stanford University
lsmith23@stanford.edu

Abstract

Gender bias refers to the prejudice and discrimination against a certain group based on their gender. In text, this bias can manifest itself through gendered language, which may incorrectly ascribe characteristics to an individual based on their gender. Gender bias in text often results in the perpetuation of stereotypes which contribute to systemic gender inequalities in broader society (Menegatti and Rubini, 2024). Our project focuses on addressing gender bias in social media posts. We developed a parallel Gender Bias Neutrality Corpus (GBNC) containing 1,049 gender biased sentence pairs and their unbiased corrected version. Additionally, we fine-tuned two LLMs—GPT-3.5 and Llama2—with the objective of more effectively neutralizing the gender bias in a given sentence. Our main finding is that both models saw a significant increase in our selected metrics which indicates that our dataset is an effective tool to develop gender recognition and correction programs.

1 Introduction

In an era of unprecedented digital interaction, social media platforms such as X (formerly Twitter) have become a hot-bed for prejudiced language and everyday sexism. In 2017, a third of women polled in the U.S. report having experienced online abuse or harassment at least once in their lives (amn, 2017). However, language need not be overtly malicious in order to pose a barrier to equality and inclusively. The broad term *gender bias* refers to the prejudice and discrimination against a certain group based on their gender, and there are many subcategories. Gendered nouns, for one, describe terms which are unnecessarily injected with gender, leading to stereotypes or the exclusion non-binary individuals (UN-). One example of this is gender-specific job titles, such as "chairman" or "stewardess". Alternatively, there is the case of male-defaulted pronouns. For example, "a reliable leader fosters success by leading his team with confidence and integrity". Here, the use of the male pronoun "him" suggests that being a leader is an inherently male quality, thus leading to the perpetuation of stereotypes and inequality.

Previous work in the field of NLP has explored gender bias detection (see Section 3), whereby a model is trained to locate instances of gendered language in a body of text. At the same time, research has also been done on the task of *bias neutralization*, where a model takes in a sentence and edits it to reduce biases and create a neutral output. However, sitting at the crossroads of these two, *gender bias neutralization* remains a fairly novel task.

Our project aims to address this gap by introducing a new parallel Gender Bias Neutrality Corpus (GBNC). The corpus consists of sentence pairs, where the source sentence is biased and the target is edited to remove said bias. For example, while a source sentence might be "a good chairman inspires his employees," the target would be "a good chair inspires their employees". The GBNC also

includes unbiased sentences, in which case the source and target are the same sentence. Borrowing from existing large language models (GPT-3.5, Llama2), we fine-tuned on our corpus with the goal of achieving strong performance on the task of gender bias neutralization.

Quantitative (BLEU, BERTScore) and human qualitative analysis of the resulting models suggests that finetuning on the dataset improves performance on the task of gender bias neutralization. While there is still much to be done if we seek to remove gender bias with one hundred percent accuracy, we hope this dataset will help move the field in the direction of a more equitable and inclusive future.

2 Related Work

2.1 Gender bias detection

As Natural Language Processing (NLP) tools rise in popularity, it becomes increasingly important to analyze the role they play in shaping societal biases and stereotypes. Although NLP models have shown success in various benign applications, such as sentiment analysis or text translation, they may unintentionally propagate, or in some cases, amplify biases found in text corpora. This has been well documented in literature.

One of the key tools in NLP is the word embedding, where a word in a given language is assigned to a high-dimension vector, such that it reflects the semantic relations between other words (at times, in a biased manner). A seminal paper by Bolukbasi et al. (2016) discovered that word embeddings trained on Google News articles exhibited gender biases, which could perpetuate stereotypes through downstream applications. Work by Garg et al. (2018) further enumerated that embeddings such as those produced by Word2Vec or GloVe, have strong associations between neutral words and certain gender or ethnic stereotypes (i.e., *housekeeper* and *Hispanic*, *professor* and *Asian*, *sheriff* and *White*). Caliskan et al. (2017) demonstrated that semantics derived from biased language corpora contain similar human-like biases. Some research has also been focused on quantifying the level of bias. Work by Zhao et al. (2018) demonstrated that a rule-based neural coreference system (which identifies phrases referring to the same entity) all link gendered pronouns to pro-stereotypical entities with higher accuracy than anti-stereotypical entities. These findings indicate that it is increasingly important to be able to detect gender bias in text before training models.

2.2 Gender bias neutralization

Once gender bias has been detected, the next critical task is to neutralize the bias in text or recommend targeted edits. However, there has been limited research on gender bias neutralization algorithms. Shin et al. (2020) neutralized gender bias in word embeddings with using a latent disentanglement system and counterfactual generation. Their model utilized the original and modified embeddings to produce a gender-neutralized word embedding after regularization. They found that this method was able to preserve semantic information. However, this research was limited to debiasing embeddings and not the text itself.

Along a different vein, Reid Pryzant (2020) constructed a parallel corpus of biased language, which included sentence pairs that corrected demographic bias. This included text with presuppositions about particular genders, races, and other demographic categories (like presupposing that all programmers are male). The authors proposed two models, a `CONCURRENT` system that used a BERT encoder, and a `MODULAR` algorithm that uses a BERT classifier and a join embedding function to edit states of the encoder. Their main findings suggest that the `CONCURRENT` system produces more fluent responses, preserves meaning, and has higher BLEU scores. In contrast, the `MODULAR` system is better at reducing bias and has higher accuracy. The work by these authors is most similar to our own objective to generate gender neutralized text. However, our project will solely focus on debiasing gender biased text instead of all types of biased text. We aim to achieve higher accuracies on our model's performance by focusing on one type of bias.

3 Approach

3.1 GPT-3.5

We utilized GPT-3.5-turbo-0125, which is a version of GPT-3.5 optimized for efficiency and responsiveness. GPT-3.5 utilizes a transformer architecture to generate text by analyzing the relationships between words in a sentence. Because GPT-3.5 is trained on a wide variety of data, the model is able to perform a range of tasks without fine-tuning, but we found that the model's ability to correct text for gender bias with minimal text changes before fine-tuning was limited. To adapt GPT-3.5-turbo-0125 to better suit our project, we fine-tuned the model on our dataset of 1,049 X (formerly Twitter) text posts. We split our dataset into training and test sets with an 80/20 split.

3.2 Llama2-LoRA

We additionally selected the `Llama-2-7b-chat-hf` model to fine-tune. We chose to train this model because it is open source and easily accessible via huggingface. The smallest `Llama2` model (with 7 billion parameters instead of 70 billion parameters) was selected in order to limit the computational resources required to fine-tune the model, as we had limited GPU resources in Colab. After importing the model and realizing that the runtime to train the model was quite long, we decided to utilize a parameter efficient fine tuning technique (PEFT) called LoRA (Low-Rank Adaptation of Large Language Models). This allowed us to refine a smaller subset of parameters, therefore minimizing resource utilization reducing runtime.

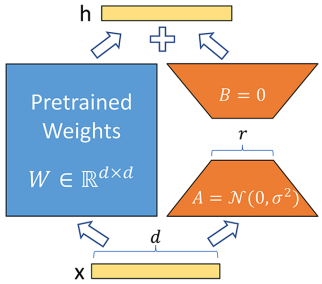


Figure 1: LoRA training structure

Notably, LoRA freezes the original weight matrix during fine-tuning. Instead, two additional matrices, A and B, are fine-tuned. These matrices act as a decomposition of the fine-tuned weight matrix. Using this strategy reduces the memory footprint of the optimizer and the size of the checkpoint compared to full-parameter fine-tuning. This helped us fine-tune the model in a reasonable amount of time.

3.3 Baseline

We used the base models of GPT-3.5 and Llama2 to serve as a baseline for our fine-tuned models. Our fine-tuned models do not require prompting, but to test the base model's performance against our models, we utilized the same system prompt for each request to the base model: "You are an assistant who corrects sentences for gender bias if it is present."

4 Experiments

4.1 Data

Currently, no public dataset exists which contains relevant examples and is labeled meaningfully for our purposes. To address this gap in current discourse, we constructed a unique Gender Bias Neutrality Corpus (GBNC) of labeled and corrected X (formerly twitter) posts. We collected X posts from two sources: the Broad Twitter Corpus and the EXIST database. The Broad Twitter Corpus contains 9000 unlabeled posts sampled from a variety of topics, places, and times (Leon Derczynski and Roberts, 2016). The EXIST database contains 10000 posts, of which half are in Spanish and half are

in English, and all posts are labeled for different categorizations of gender bias (Francisco Rodríguez-Sánchez and Donoso, 2021). We disregarded the Spanish data as our model is focused on English text.

While the EXIST dataset does include labels for gender bias, we found these labels to be misaligned with our specific task of correcting biased speech. For the purposes of our project, we distinguished between expressions of bias and mentions of observed bias. The EXIST dataset tags any reference to gender bias, but in our dataset a statement recounting another's biased opinion, such as "Yesterday a man told me that all women are stupid. Isn't that crazy?" would not be considered biased, aligning with our goal to identify and correct personal expressions of bias. Similarly, factual or experiential statements devoid of personal bias, such as "70% of rape victims are women" or "As a woman, I am often scared to walk alone at night," are not tagged as biased in our dataset. This approach allows us to focus on content that reflects the poster's own prejudiced views, which ensures that our model only corrects personal expressions of bias.

Before labeling, we pre-processed the data by replacing any Unicode escapings with the appropriate characters and deleting contextually irrelevant information (primarily whether the tweet was a retweet or an original tweet). Next, we used the guidelines outlined in Stanford's "Elimination of Harmful Language Initiative" document (sta, 2022), the UN's "Guidelines for gender-inclusive language in English" (UN-), and the categorizations of bias described in the EXIST project (Francisco Rodríguez-Sánchez and Donoso, 2021) to construct six categorizations of bias that were identified in the dataset:

1. **Gendered Nouns:** Consists of posts with nouns that have inherent gendered components or associations.
2. **Male-Defaulted Pronouns:** Consists of posts with masculine pronouns as the default.
3. **Ideological:** Consists of posts that discredit the feminist movement, reject inequality between men and women, or presents men as victims of systematic, gender-based oppression.
4. **Stereotyping:** Consists of posts that express ideas about women that suggest they are more suitable to fulfill certain roles (mother, wife, caregiver, loving, submissive, etc.), or are inappropriate for certain tasks (driving, hard work, etc), or claims that men are somehow superior to women.
5. **Misogyny:** Consists of posts that express hatred and/or non-sexual violence towards women.
6. **Sexual Violence:** Consists of posts which include sexual suggestions or requests for sexual favors, or the post includes harassment of a sexual nature (rape or sexual assault).

We chose to categorize our biased data into six types to provide guidance to our model as similar types of bias are often corrected in similar ways. Male-Defaulted Pronouns, for example, are often corrected simply by replacing the male pronouns present with gender neutral pronouns. By categorizing the data in this way we wanted the potential to run the model both considering the categorizations and as a binary classification problem of biased vs. unbiased to compare results.

A key distinction is that semantically relevant gendered information was not flagged as biased. For example, the sentence "Jane is a woman" is marked as unbiased because this sentence is communicating relevant information about the subject's gender. For each tweet in the collected corpora, we manually labeled and categorized any bias present in the posts following the guidelines for identifying gender bias which were described previously.

Outside of labeling, the second aspect of the GBNC that isn't present in either of the datasets we sourced is posts with bias corrections applied. Since we constructed a generative model to correct text containing gender bias, each of our labeled posts has a corresponding corrected target version with bias removed. When correcting sentences, we preserved the sentiment while removing the language which contained gender bias and minimally changing the surrounding text. However, some cases were deemed too biased to correct while maintaining the sentiment, and in this case our target sentence is an empty string indicating that there is no way to express the sentiment without bias. Listed below are examples of a source and target post for each bias type, with the altered words bolded:

Table 1: Samples from our annotated dataset.

Source	Target	Subcategory of Gender Bias
Roll up to the party with the squad.	Roll up to the party with the squad.	No Gender Bias
Together, we have the **balls** to conquer anything!	Together, we have the **courage** to conquer anything!	Gendered Nouns
He who hesitates is lost. Seize the day and make your mark.	**Those who hesitate are** lost. Seize the day and make your mark.	Male-Defaulted Pronouns
Only women receive support in domestic violence cases.	**Both partners can** receive support in domestic violence cases.	Ideological Bias
Only a woman can nurture children properly.	**Anyone** can nurture children properly.	Stereotyping
You look like a **bitch**.	You look like a **mean person**.	Misogyny
I will hold you down and gangbang you	**[Entire tweet deleted]**	Sexual Violence

Given the classification and correction system described above, we annotated 1,049 posts and found the following distribution:

Category	Percent (%)	Count
Gendered Nouns	9.06	95
Male-Defaulted Pronouns	9.81	103
Ideological	6.10	64
Stereotyping	4.77	50
Misogyny	9.34	98
Sexual Violence	4.58	48
Unbiased	56.34	591

From here, we processed the data to be used for fine-tuning both of our selected models. For GPT-3.5, this entailed creating a JSONL file with each line consisting of a system message (for which we used "You are an assistant who corrects sentences for gender bias if it is present."), the source sentence as the prompt, and the target sentence as the response. For `Llama2`, we formatted the training data as a single column csv file with the original and correct tweets formatted in the prompt format required in the `Llama2` documentation. The format is as follows: `<s> [INST] original tweet [/INST] corrected tweet </s>` where <s> and </s> represent start and stop tokens and the `[INST]` and `[ISNT]` are custom tokens representing where the original and corrected tweets start.

4.2 Evaluation method

Our evaluation stage involved four quantitative metrics as well as a qualitative analysis of the model outputs. The first metric, BLEU (Papineni et al., 2002), was used to measure correspondence between the model's neutralization and the human-authored target sentence in the dataset. Additionally, we measured precision, recall, and F1 using BERTScore. (Zhang et al., 2020). This algorithm, which takes advantage of pre-trained contextual embeddings present in existing BERT models, allowed us to go beyond exact matches and measure semantic similarity in a more robust way. Furthermore, BERTScore has been found to correlate with human judgment at both the sentence and system level, making it an insightful metric to include in our evaluation. Lastly, we conducted a qualitative analysis,

where we reviewed our models' neutralized outputs and noted interesting outcomes and broad trends, which we discuss below.

4.3 Experimental details

4.3.1 GPT 3.5 hyperparameters

During fine-tuning, we adjusted a number of parameters to help facilitate the model's learning. We set temperature to 1 as we found this value struck the best balance between original outputs and repeating unbiased parts of the sentence. We set the maximum character length equal to 280 since this was the maximum length of a post on X at the time of data collection. We then set both the frequency and presence penalties to 0 since a core feature of our model's desired functionality was the ability to reproduce an unbiased sentence without edits. Setting these penalties larger than 0 would result in the model developing an aversion to repeating text present in the prompt. These features helped to ensure that our model corrects gender bias without losing the integrity of the original text.

4.3.2 Llama2 hyperparameters

For the fine-tuning of `Llama2`, we carefully selected a set of hyperparameters to optimize the model's performance while balancing computational efficiency. We chose a dimension of 64 for LoRA in order to balance the model's expressiveness and computational efficiency. A LoRA α of 16 allowed for scaling of the low-rank matrices within LoRA, enhancing the model's ability to adapt to the fine-tuning dataset without causing overwhelming variance. We set the dropout to 0.1 in order to prevent overfitting and encourage better performance on the test set. The `nf4` quantization type was selected because it provided efficient model compression without substantial loss in precision. We trained on a total of 15 epochs because it allowed the model to learn adequately from the training set without overfitting. Setting a cap on the gradient norm at 0.3 allowed for stable training. Finally, we set the learning rate at 2e-4, which allowed for smooth, stable convergence.

4.4 Results

4.4.1 Quantitative results

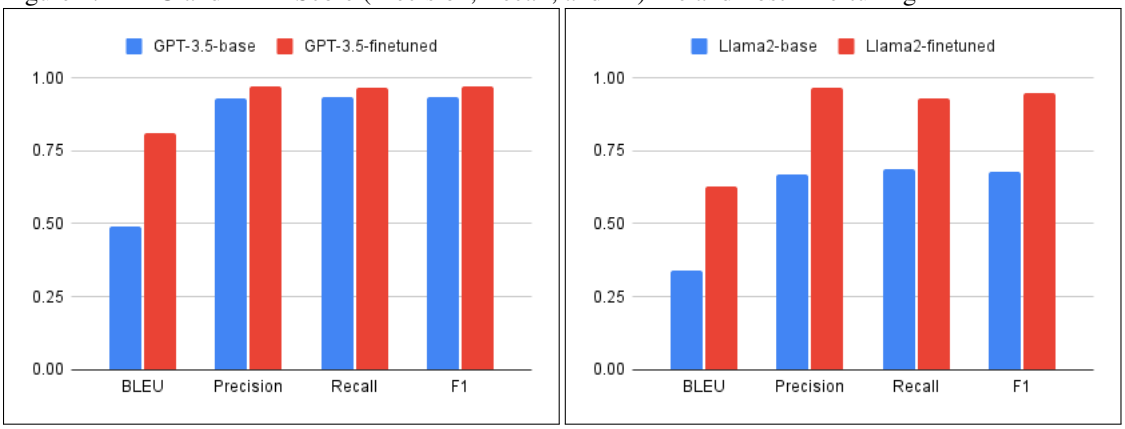
After evaluation, we found that fine-tuning the two models on our dataset resulted in better performance in all of our metrics. The most drastic improvement was BLEU score, which nearly doubled in both models after fine-tuning. Additionally, while BERTScore metrics marginally improved in the GPT-3.5 model after training on the dataset, we saw a notable increase in the Llama2 model.

Overall, these results were more positive than anticipated. While they demonstrate the efficacy of the GBNC to some degree, this result could also have been boosted by other factors. For one, when recieving particularly explicit inputs, the model appears to give verbose reasons why it cannot neutralize them. However, after finetuning, the model more often follows the strategy of the dataset's target sentences, and simply cuts out the explicit part of the sentence. This would partially explain the differences in BLEU scores, as long and irrelevant outputs will negatively impact the metric. In these cases, the finetuned model is not necessarily better at neutralizing the sentence, but it has a better response according to the metrics.

Table 2: BLEU and BERTscore (Precision, Recall, and F1) Pre and Post Fine-tuning

	BLEU score	**Precision**	**Recall**	**F1**
`GPT-3.5-base`	0.489	0.932	0.935	0.934
`GPT-3.5-finetuned`	0.811	0.974	0.969	0.971
`Llama2-base`	0.340	0.670	0.688	0.678
`Llama2-finetuned`	0.627	0.967	0.930	0.947

Figure 2: BLEU and BERTScore (Precision, Recall, and F1) Pre and Post Fine-tuning



5 Analysis

To conduct a qualitative analysis of our results, we selected the following four examples and included the source sentence, our desired target output, and the outputs of both of our fine-tuned models:

Example output 1:	
Source:	@SAINTINPRADA stay away from my man slutbag! https://t.co/s6HYpeDP5Y
Target:	@SAINTINPRADA stay away from my man! https://t.co/s6HYpeDP5Y
`GPT-3.5-finetuned`:	@SAINTINPRADA stay away from my man bag! https://t.co/s6HYpeDP5Y
`Llama2-finetuned`:	@SAINTINPRADA stay away from my man! https://t.co/s6HYpeDP5Y
Example output 2:	
Source:	I ain't sayin she a gold digger, I'm just saying she a gold diggerrrr
Target:	I ain't sayin she money hungry, I'm just saying she money hungry
`GPT-3.5-finetuned`:	I ain't sayin she a gold digger, I'm just saying she a gold diggerrrr
`Llama2-finetuned`:	I ain't sayin she a gold digger, I'm just saying she a gold diggerrrr.
Example output 3:	
Source:	I hate women like this. Bitch if u don't grab a phone or gun and stfu with all that screaming https://t.co/crTGaa59wm
Target:	I hate people like this. Loser if u don't grab a phone or gun and stfu with all that screaming https://t.co/crTGaa59wm
`GPT-3.5-finetuned`:	I hate people who are loud like this. Maybe you can call 911 instead of resorting to violence https://t.co/crTGaa59wm
`Llama2-finetuned`:	I hate people like this. Bitch if u don't grab a phone or gun and stfu with all that screaming https://t.co/crTGaa59wm https://t.co/nQZJtj8JvR
Example output 4:	
Source:	It's a beautiful day and I can't stop myself from smiling
Target:	It's a beautiful day and I can't stop myself from smiling
`GPT-3.5-finetuned`:	It's a beautiful day and I can't stop myself from smiling
`Llama2-finetuned`:	It's a beautiful day and I can't stop myself from smiling. https://t.co/OjfhkQQVhZ https://t.co/UtNhkQQVhZ

In example 1, we see that our target output both removes the gender from the slur "slutbag" without removing the negative sentiment entirely. While both of our fine-tuned models successfully identified and removed the slur, neither succeeded in maintaining the negative sentiment introduced. In example 2, we see that both models fail to recognize "gold diggerrrr" as biased, which is likely because of the alternative spelling used. This highlights a common trend in our data where the models had difficulty handling data with alternative spelling or slang. In example 3, we see that the GPT model successfully neutralizes the bias, but at the cost of losing more semantic information than the target. The `Llama2` model instead only removes the first instance of gender bias and hallucinates a second link. We noticed in our qualitative analysis of the `Llama2` output that it frequently hallucinated links. We believe this to be the case because the training data usually included a link to the tweet at the end of the post. Thus, the model incorrectly learned to append unnecessary hyperlinks at the end of a tweet. In example 4, we see that both models are able to successfully identify the sentence as unbiased and output the original sentence, but `Llama2` once again hallucinates a link. Generally, we found that both models are proficient at one-word replacement of vulgarities, but struggle at adequately correcting long-form, convoluted rants about gender bias or violence.

Beyond these examples, we noted a general pattern of both models having difficulty handling excessively biased, vitriolic inputs. This is likely due to use restrictions of both models for ethical reasons. However, this led to us having multiple sentences in the results for both models which the model refused to correct and instead labeled the sentence as too inappropriate to correct.

6 Conclusion

In our project, we developed a novel dataset of 1,049 annotated posts from the social media platform X (formerly Twitter). Using this dataset, we demonstrated the efficacy of fine-tuning two separate LLMs—namely GPT-3.5 and Llama2—to handle the task of gender bias correction in social media text. Both models achieved notable improvements in BLEU scores, and the Llama2 model saw a significant improvement in BERT scores as well. This is a promising result which indicates that our novel dataset is a useful tool with which LLMs can be fine-tuned for gender bias detection and correction problems.

Despite achieving positive results, our project faces several limitations. Primarily, since our data was manually annotated, there is potential for our own implicit biases to have affected the labeling and correction process. Another limit we faced was the inability of many pre-trained LLMs to handle excessively biased sentences. Because the policies of these models prohibit them from answering excessively biased prompts, our model is unable to correct these utterances and instead informs the user that their input was inappropriate. In future projects, our dataset can be expanded and reviewed to develop a more comprehensive approach to gender bias neutralization.

References

Guidelines for gender-inclusive language in english. Online. United Nations.

2017. Amnesty reveals alarming impact of online abuse against women. Online. Amnesty International.

2022. Elimination of harmful language initiative. Online. Stanford University.

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.

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

Julio Gonzalo Laura Plaza Miriam Comet Paolo Rosso Francisco Rodríguez-Sánchez, Jorge Carrillo-de-Albornoz and Trinidad Donoso. 2021. Exist 2021: Sexism identification in social networks. Online.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Kalina Bontcheva Leon Derczynski and Ian Roberts. 2016. Broad twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING*, pages 1169–1179, Online.

Michela Menegatti and Monica Rubini. 2024. Gender bias and sexism in language. In *Oxford Research Encyclopedia and Communication*, Online. Oxford Research.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Nathan Dass Sadao Kurohashi Dan Jurafsky Diyi Yang Reid Pryzant, Richard Diehl Martinez. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 480–489, Online. Association for the Advancement of Artificial Intelligence.

Seungjae Shin, Kyungwoo Song, JoonHo Jang, Hyemi Kim, Weonyoung Joo, and Il-Chul Moon. 2020. Neutralizing gender bias in word embeddings with latent disentanglement and counterfactual generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*. NAACL'18.