Adversarial Examples Generation for Reducing Implicit Gender Bias in Pre-trained Models

Fei Xu
jx1260@nyu.edu

Chenye Xu
cx650@nyu.edu

Abstract

Over the last few years, Contextualized Pre-trained Neural Language Models, such as BERT, GPT, have shown significant gains in various NLP tasks. In the meanwhile, gender bias embedded in the models seems to be a serious problem in practical applications. Many researches have covered the gender bias produced by word-level information(e.g. gender-stereotypical occupations), while few researchers have investigated the sentence-level cases and implicit cases.

In this project, we proposed a method to automatically generate implicit gender bias samples at sentence-level and a way to measure gender bias. Samples generated by our method will be evaluated in terms of accuracy. Finally, we discussed the evaluation efficacy of our generated examples on reducing gender bias for future research.

1 Introduction

The attention paid on the field of machine learning systems has kept increasing in recent years. However, many machine learning models are shown to be vulnerable to adversarial attacks; namely, these models would give incorrect predictions by feeding them intentionally-modified inputs (aka adversarial examples). For example, Belinkov and Bisk (2018) showed that swapped characters and typos lower the BLEU score significantly in neural machine translation models. On the other side, the adversarial attacks generated on the model developers' side will make the trained models more robust.

Most recently, with the increasing focus on large-scale deep learning systems, a plethora of pre-trained models have emerged extensively and shown significant gains in various NLP tasks with fine-tuning. Such remarkable progress intrigued researches on evaluating whether these models are robust by conducting adversarial attacks and, in the meantime, provided suggestions (e.g. data augmentation, adversarial learning) to improve the performance of existing pre-trained models. In practice, the propagation of gender bias in pre-trained models has raised the public awareness, and it may pose the danger of reinforcing damaging stereotypes in downstream applications. Models automating resume screening have also been proved to have a heavy gender bias favoring male candidates (Lambrecht and Tucker (2019)). Such data and algorithmic biases have become a growing concern.

The state-of-the-art techniques, like the one proposed by Bolukbasi et al. (2016), have already covered word-level gender bias detection using the gender subspace distance. Also, from the result of Bordia and Bowman (2019), it has shown the possibility to accurately measure the word-level gender bias in a text corpus and reduce such bias by proposing a regularization loss term for the language model. However, these word-level techniques cannot generalize well to phrase-level or sentence-level cases. For existing techniques on the sentence-level bias like May et al. (2019a), they also relied on the word-level techniques and have a strict limitation on the sentences. In this project, we proposed a generalized method to generate implicit gender bias samples in sentence level and we proposed a metric to measure such bias. The metric can guide the generation of bias examples from Pre-trained models. And the creation of such examples allows for com-

parison of its impact on different Pre-trained models. Finally, we discussed our plan to evaluate the efficacy of our generated examples on reducing gender bias, we left this to our future investigation due to the limited time.

2 Methods

2.1 Problem Statement

Most previous work on gender bias in the NLP community has focused on identifying the bias and debiasing existed models. To tackle the problem at its root and allow for a comparison of its impact on various language models, we believed that an open dataset must be created.

To that end, the first step is to quantify the qualitative definition of implicit gender bias. Inspired by the gender bias taxonomy proposed by Hitti et al. (2019), we focused on two main gender bias issues: structural bias and contextual bias, because they have not been well studied and needed contextual information to understand it. In general, we defined implicit gender bias in the text as the whole sentence that connotes, prejudices, or implies an inclination against one gender. Implicit gender bias cannot be indicated only by the words used in the sentence. We also introduced a metric to quantify it. The details are shown below.

Structural bias. Followed by the definition of Hitti et al. (2019) as well as the examples in Macaulay and Brice (1997), we defined structural bias as one type of implicit gender bias. This kind of gender bias occurs when bias could be traced down from a particular syntactical construction. This includes searching for any syntactic patterns that enforce gender assumptions in a gender-neutral context. In another word, this type of gender bias appears when a gendered pronoun is linked to a gender-neutral term in a gender-free context. Gendered pronouns included: *he, his, him, himself, she, her, hers and herself.*

- "Someone winds up his right arm and knocks the fighter down with a haymaker." - shows an action of a person and assumes the person to be a man by referring to

"his".

- "The belly dancer dances on stage shaking her hips and body." - gives a fact about an arbitrary dancer and assumes a woman to be the dancer by referring to "her".

Counter example:

- "A girl will always want to play with her Barbie doll."- although describing a stereotype, it is not assuming the gender for a gender neutral word because the word girl (gendered - female) is linked to a female pronoun. Therefore, it is not gender bias.

Contextual bias. On the other side, contextual gender bias does not have a pattern-based definition. It often requires the learning of the link between gender obvious keywords and contextual information. Unlike above type, this type of bias cannot be observed through syntactical structure but requires contextual knowledge and some degree of human perception.

- "Presidential candidates need their wives to support them throughout their campaign." - the word "wife" is described as a supporting role when we do not know the gender of the Presidential candidates and the supporting role can be a male partner, a husband.

- "Your sister could go to College, but would she get a degree?" - it clearly casts doubt on the possibility of academic success for this woman.

- "John must love football because all boys like playing with it." - assumes that football are only liked by boys.

Implicit gender bias metric. Caliskan et al. (2017) first introduced Word Embedding Association Test (WEAT) to quantify gender bias in English embedding, which is a permutation test. The test considered two sets of target

words and two sets of attribute words, and measured the association between them. Formally, let X and Y be two sets of target word vectors of equal size (such as word embeddings related to different professions) and let A and B be sets of attribute word vectors (e.g., gender marked definition word vectors,i.e, "man", "woman"). Let $cos(\vec{a}, \vec{b})$ denotes the cosine similarity between word vectors $\vec{a}$ and $\vec{b}$. The test objective function is a difference between sums over the each target words,

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

where $s(\vec{w}, A, B)$ measures the difference of the word w with the attributes,

$$s(\vec{w}, A, B) =$$
$$\frac{1}{|A|} \sum_{\vec{a} \in A} cos(\vec{w}, \vec{a}) - \frac{1}{|B|} \sum_{\vec{b} \in B} cos(\vec{w}, \vec{b}).$$

Followed by this work, we extend the target word sets by adding a kind of phrase sets. We considered gender-neutral behavioral phrases but usually depicted in stereotypes(like "knocks the fighter down" and "wear a revealing dress"). This metric will be used to compute the bias score of our produced examples by our filtering method and we discussed using this as a guidance to generate gender bias samples from Pre-trained models in the future work.

2.2 Approach

Between the two types of gender bias proposed in section 2.1, we mainly focused on judging Structural gender bias. Our method of examining whether a given pre-trained model exhibit implicit gender bias by feeding the model sentences where gendered pronouns are masked and comparing the probabilities it predicts for male words and female words. We designed a two-stage method that contains a filter stage and a bias computation stage. In the filter stage, we perform a 4-step filtering to get desired sentences. We propose such algorithm to ensure a sentence contains only one gendered pronoun and no other sex indicators so that the

pre-trained model we try to investigate will not be misled by irrelevant words in the sentence.

3 Experiments

3.1 Initial Datset

We chose SWAG, a large-scale adversarial dataset for grounded commonsense inference (Zellers et al., 2018), to be our starting point. Each row of SWAG dataset has the format as shown in Table 1, where the fields in the first column are the column headers in SWAG.

video-id	lsmdc1005_Signs-5092
fold-ind	6072
startphrase	Someone walks over to the radio. Someone
sent1	Someone walks over to the radio.
sent2	Someone
gold-source	gold
...	...

Table 1: **The Format of The Training Data of SWAG.** The rows represents the original columns in SWAG and only the first 6 columns are shown here.

We extracted **sent1** from the dataset because we only need one complete sentence for each sample. We eventually obtained 73547 lines of sentences.

3.2 Filter Stage

The dataset we got includes various types of sentence, but not all of the sentences exhibit gender bias. For example, the sentence "someone walks over to the radio" in Table 1 does not contain any words related to gender. Therefore, we proposed a filtering process to get sentences that could be modified to exam gender bias in pre-trained models. The filtering steps are shown in Figure 1.

We defined a group of gendered identities such as "sister", "boy", "actress", etc. The words in gendered pronouns and gendered identities are collectively called sex indicators. Then, we selected sentences that contain only one of the gendered pronouns and no other sex indicators. By doing so, we eliminated
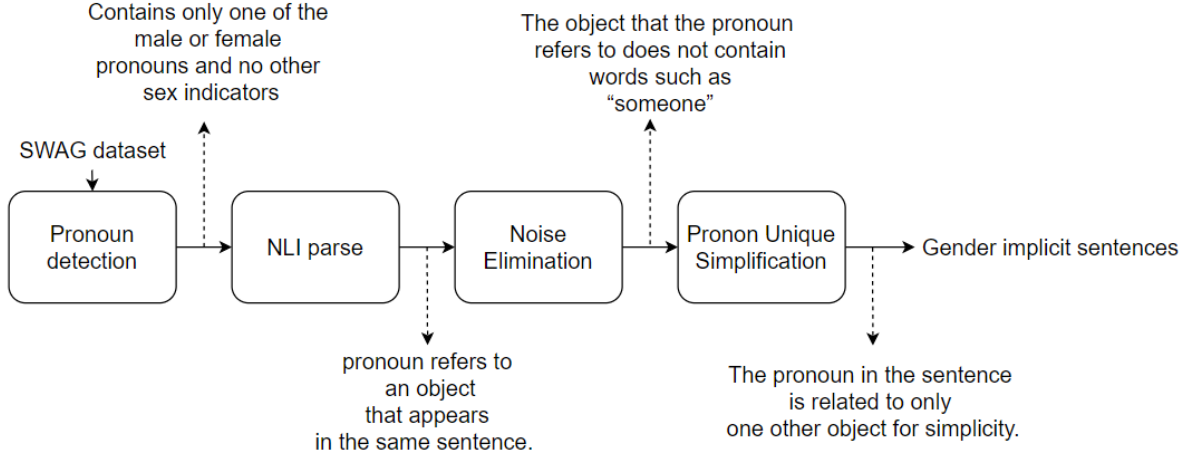
Figure 1: **Design diagram for filtering the sentences with implicit gender bias.**

sentences with no gendered pronoun and sentences containing multiple sex indicators such as "the boy met his friends in his house." The latter is not wanted because a model can infer the gender of one word from the other sex indicators.

We used the model proposed by Lee et al. (2017), which achieves an F1 score of 78.87 on the Ontonotes 5.0 dataset, to do coreference resolution on the sentences. We obtained one or more groups of associated expressions for each sentence. For example, the correct result for "the nurse is looking after her patients" is a group containing "the nurse" and "her". For each sentence, we checked whether there exists a bag that satisfies the following criteria:

- the number of expressions in the bag is 2,

- one expression is a gendered pronoun and the other is a word other than "someone".

These criteria are not mandatory as they only serve to formulate our dataset in a more restrictive fashion. The sentences not fulfilling the above criteria were filtered out from the dataset.

Finally, we masked the gendered pronouns in the sentences.

3.3 Bias Computation Stage

After acquiring the masked sentences that cannot be explicitly inferred by word level techniques, we used them to test how gender biased the Pre-trained model is. Figure 2 shows the procedure to do such job. We use Pre-tranied model to do a masked model prediction on those sentences to get top $k = 10$ predictions and their corresponding probabilities. From those predictions, we picked the highest probability for the word that refer to a female gender identity as $P(s|f)$. In the same way, we picked the highest probability for the word that refer to a male gender identity as $P(s|m)$. We define the function to compute bias score $bias(s)$ as:

$$\text{bias}(s) = \frac{P(s|m)}{P(s|m) + P(s|f)}, \quad (1)$$

where s is the masked sentence.

An online Masked Language Modeling tool provided by AllenNLP(for Artificial Intelligence) is used here to show an short example. The model used here is BERT, the first large transformer to be trained on this task. The following masked sentence is used as an input for this mode:

- "[MASK] stands up next to someone."

The model predicts [MASK] to be 'He' with a probability of 43.5%, and 'She' with a probability of 19.5%. According to our metric for computing bias, the score of the prediction for this sentence is 0.69, which is biased to male on this scenario.

We fed a masked sentence to the model and calculated the score for the prediction. If the

score does not fell in the range $[0.5-\delta, 0.5+\delta]$, in which δ is the threshold we defined, we considered the model is genderbiased on the input sentence.
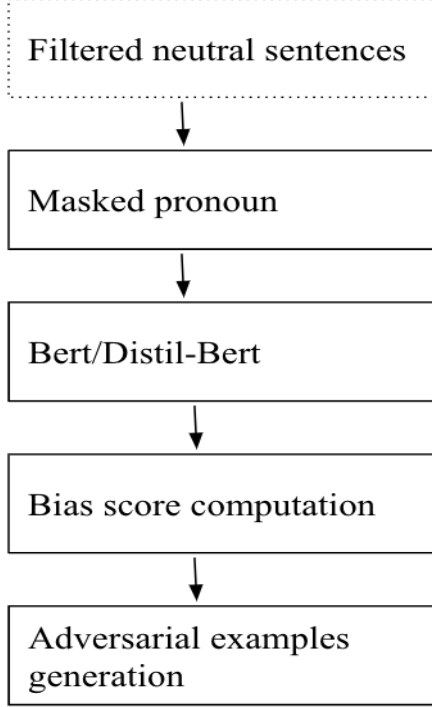

Figure 2: **Design diagram for computing bias and generating adversarial examples**

4 Results and Analysis

By filtering the SWAG dataset, we got 640 sentences that could be used to examine possible implicit gender bias in pre-trained model. We fed those sentences into BERT-uncased model, BERT-cased model and DistilBERT to do further validation.

4.1 Evaluation and Case study

To prove our method to filter sentences with structural gender bias is reasonable, we first examined the filtered sentences by ourselves in terms of accuracy,

$$Acc = \frac{\text{\# of gender bias samples}}{\text{\# of all filtered samples}}.$$

We sampled 640 examples from the original dataset, and we labeled 581 out of 640 filtered results as gender-bias samples. The filtered

Statistic	Numbers
No. of original examples	73547
No. of filtered examples	640
No. of gender-bias examples	581
Accuracy	90.78%

Table 2: Accuracy of our filtered method to produce gender-bias samples.

results were reviewed by both of our group members and we took agreements both as truly gender-biased for reliability. The details was shown in Table 2.

We further gave some case studies of our results. We found that the non-structural bias sentence, like the counter example from section 2.1: "A girl will always want to play with her Barbie doll." was successfully filtered out by our method. The two correct examples with structural bias from Section 2.1 was picked from the result we got after the filtering stage. We masked one of the sentence like the following:

- "Someone winds up [MASK] right arm and knocks the fighter down with a haymaker."

The masked sentence was then fed into two pre-trained models. The BERT-uncased model predicted [MASK] to be 'his' with a probability of 14.2%, and 'her' with a probability of 11.3%. The DistilBERT model predicted [MASK] to be 'his' with a probability of 54.2%, but 'her' with only probability of 3.05%. According to our metric for computing bias, BERT-uncased model get a score of 0.56 on this sentence, while DistilBERT model got a score of 0.95. This result indicates that for a sentence produced by our method, the two pre-trained models showed different level of implicit gender bias, with DistilBERT model having a more severe bias.

4.2 Efficacy for reducing gender-bias Analysis

The score distributions of distil-bert before and after finetune are shown in Figure 3.

Before finetune, the result shows that, among those sentences, 531 of them are male
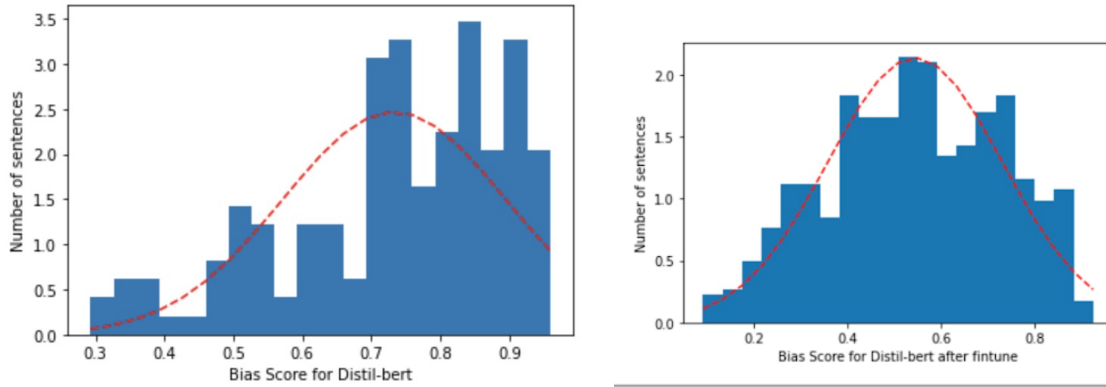
(a) **Distribution of the result from distil-bert before fine-tune** (b) **Distribution of the result from distil-bert after finetune**

Figure 3: **Distributions of the results obtained from distil-bert without and with finetune respectively**. After finetune, the distribution has a Guassain shape and the peak of the distribution move right to middle around 0.5, indicating the model becoming less biased to gender on the dataset after finetune.

biased, with an average bias score of 0.768. The rest sentences are female biased, with an average bias score of 0.396.

After finetune, among 640 filtered sentences, the result shows that among those sentences, 528 of them are male biased, with an average bias score of 0.671. The rest sentences are female biased, with an average bias score of 0.357.

The results shows that the bias ratio is reduce for male by 10% while female increased like 4% absolutely. The female increase may because our samples covers less samples on this gender, which in future we think should make our data set more balance on two genders to improve this issue.

5 Related Works

Robustness of pre-trained models. A lot of works have been proposed to study the robustness of pre-trained models in recent years. One spectrum of works approaches the attack in a black-box manner, focusing on variation of adversarial example generation (Jin et al. (2020); Kaushik et al. (2019)). Another spectrum of works attacks the model in a white-box way which uses the model information to generate adversarial examples(Liu et al., 2020; Tu et al., 2020). Our work mainly focused on evaluate the robustness of Pre-trained models on a particular aspect:gender-bias.

Gender bias in Language models. Neural language models are becoming prevalent in real world applications, and embedded bias will have a great societal impact(Buolamwini and Gebru, 2018; Tonry, 2010). Many types of research have studied the gender bias in neural language models such as word embedding models (Bolukbasi et al., 2016; Caliskan et al., 2017; Wang et al., 2020) and NMT models(Saunders and Byrne, 2020). Zhao et al. (2020) investigated the Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer. May et al. (2019b) measured gender bias in sentence encoders such as ELMo and BERT. Zhao et al. (2017) showed the model can amplify the bias without appropriate adjustments. In Shah et al. (2019) 's work, they summarized the previous works and proposed a unifying predictive bias framework for NLP. Different from the previous work, our work aim at understanding the implicit bias in phrase-level or sentence-level and proposed a model to generate bias examples to attack the pre-trained models.

6 Conclusion and Future work

In this paper, we investigated implicit gender-bias text at sentence-level. We defined the implicit gender-bias as the whole sentence that connotes, prejudices against one gender. In particular, we focused on two main gender-

bias issues which have not been well studied and needed contextual information to identify it. In the meantime, a metric was introduced to compute the gender-bias score of a sentence. The purpose of our project is to create an open dataset allows for a comparison of its impact on different language models and to reduce gender-bias if they existed in those models. To the end, we presented a two-stages method to produce such implicit samples from an dataset. Finally, we evaluated the results given by our method. All filtered results were reviewed by both of us. A detailed analysis of the results was also presented to see the two types of gender-bias studies by us. There are also four other directions need to further investigate,

- We should test our approach on more dataset to get a more comprehensive validation of results. Because our method to test gender bias so far on a pre-trained model depends on the dataset we used to filter sentences.

- We should test our approach on more pre-trained models like the variants of transformers, ALBERT or RoBERTa.

- With the help of proposed metric, automatic generating more implicit gender-bias examples using various contextualized Pre-trained models.

7 Acknowledgements

References

The Allen Institute for Artificial Intelligence. Masked language modeling.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.

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*, 29:4349–4357.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91.

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

Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. 2019. Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *arXiv preprint arXiv:1907.11932*.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Anja Lambrecht and Catherine Tucker. 2019. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*, 65.

Kenton Lee, Luheng He, M. Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *ArXiv*, abs/1707.07045.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

Monica Macaulay and Colleen Brice. 1997. Don't touch my projectile: Gender bias and stereotyping in syntactic examples. *Language*, pages 798–825.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019a. 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.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019b. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. *arXiv preprint arXiv:2004.04498*.

Deven Shah, H Andrew Schwartz, and Dirk Hovy. 2019. Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv preprint arXiv:1912.11078*.

Michael Tonry. 2010. The social, psychological, and political causes of racial disparities in the american criminal justice system. *Crime and justice*, 39(1):273–312.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2020. Double-hard debias: Tailoring word embeddings for gender bias mitigation. *arXiv preprint arXiv:2005.00965*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. *arXiv preprint arXiv:2005.00699*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.