Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/52/20241234

Exploration, detection, and mitigation: Unveiling gender bias in NLP

Chunxiao Zhang

Department of Mathematics, University of California, San Diego, 9500 Gilman Dr, La Jolla, CA 92093, United States

chz024@ucsd.edu

Abstract. Natural Language Processing (NLP) systems have a mundane impact, yet they harbour either obvious or potential gender bias. The automation of decision-making in NLP models even exacerbates unfair treatment. In recent years, researchers have started to notice this issue and have made some approaches to detect and mitigate these biases, yet no consensus on the approaches exists. This paper discusses the interdisciplinary field of linguistics and computer sciences by presenting the most common gender bias categories and breaking them down with ethical and artificial intelligence approaches. Specific methods for detecting and minimizing bias are shown around biases present in raw data, annotator, model, and the linguistic gender system. In this paper, an overview of the hotspots and future perspectives of this research topic is presented. Limitations of some detection methods are pinpointed, providing novel insights into future research.

Keywords: Gender Bias, Bias Detection, Bias Mitigation, Ethics in NLP.

1. Introduction

A branch of artificial intelligence called Natural Language Processing (NLP) aims to teach computers how to interpret natural languages [1]. Though AI has always been viewed as an impartial force, it is made by humans, so it exhibits our preconceptions, including gender bias [2] [3]. When translating from English to Spanish, the word “nurse” is translated as female and the word “doctor” is translated as male, implying word embeddings identify linkages between certain professions with certain genders [1]. This is one instance of sexism in the NLP field. In recent years, as NLP has become increasingly utilized, the impact of its bias on society has also amplified. Stereotypical information encoded in NLP models can even lead to the deprivation of women’s working opportunities. Because of the gender imbalance in certain occupations in real life, AI-based machines from Amazon once automatically penalized resumes that included “women’s” and downgraded these resumes [4]. Gender bias, as one of the serious emerging issues in NLP, has led to intense discussions in academic fields concerning the origins of bias and the approaches to tackling it.

This paper explores the interdisciplinary domain of linguistics and computer science by presenting the prevailing categories of gender bias and examining them through ethical and artificial intelligence perspectives.

DOI: 10.54254/2755-2721/52/20241234

2. Identifying Gender Bias in NLP

Gender bias can be categorized into specific types, including under-represented gender bias, translation bias, gendered word embeddings bias, and gendered sentiment analysis bias. Identifying gender bias in NLP into distinct types helps researchers to analyze the reasons leading to them and perform tailored mitigations on them.

2.1. Under-represented Gender Bias

Firstly, under-represented bias is when the population of specific identity groupings is not proportional to the number of people in those groups [2]. According to Zhao et al., ELMo embeddings perform completely differently on female and male pronouns, as predictions on male entities are 14% more accurate than those on female entities from occupation words [5]. Underrepresentation can make certain groups experience exclusion from the mainstream, causing them to be isolated and hindering them from various opportunities.

2.2. Translation Bias

Secondly, language systems have differences in gender indication and cause translation bias. Languages can be notional-gendered if they are limited to pronominal gender, or they can be grammatical-gendered if they have a higher level of gender inflection [6] [7]. According to Piazzolla et al., the Italian equivalent of the English word "sexy dancer" would be feminine, while the Italian equivalent of the English phrase "top dancer" would be masculine [6]. Moreover, the resulting Italian term for “sexy dancer” would still be feminine even if there is a clue suggesting that the dancer is a boy [6]. Gender bias across languages can lower translation accuracy and may lead to misunderstandings that reduce the effectiveness of cross-lingual communications. This distortion of reality could skew gender imbalance in occupations and lead to misperceptions towards them.

2.3. Gendered Word Embeddings Bias

Thirdly, gendered word embeddings associate biased gender representations of vectors. These embeddings are word representations learned from large corpora, which draw analogies between words by vectors in a numerical space. One example of gender bias in word embeddings is that the word association between “men” to “computer programmer” is almost the same as “women” to “homemaker”, inferring certain associations of genders and occupations [8]. One of the consequences of indulging this bias is that targeted recruitment and content recommendations frustrate women in their job applications. For example, Facebook's targeted hiring algorithms have resulted in female job seekers being unable to see job ads from companies that frequently employed males in the past [9].

2.4. Gendered Sentiment Analysis Bias

Fourthly, gendered sentiment analysis bias causes sentiment intensity predictions for one gender to be consistently higher [10]. For instance, gendered sentiment analysis would prioritize a call from an angry male over a call from an angry female [10]. This biased analysis would wrongly assess users’ sentiments, leading to unfair treatment towards one gender, and reinforces norms about how different genders express emotions.

All types of gender bias mentioned above entrench gender stereotypes and suppress diversity. The perpetuation of gender norms can lead machines to generate biased results, making them function in unfair ways. Machines could also sway people's views and discourage certain gender groups from pursuing some fields, adversely affecting people’s conception of the inclusiveness of gender. To avoid the continual effects of these biases, researchers have been striving to find out what leads machines to incur biases. Understanding the reasons behind these biases would ultimately benefit researchers by proposing mitigation strategies on specific factors, thereby eliminating biases outright.

DOI: 10.54254/2755-2721/52/20241234

3. Factors Causing Gender Bias in NLP

Gender bias in NLP is a multifaceted issue which can be caused by various intersecting factors. Factors causing gender bias in NLP can be generalized to data imbalance, annotator bias, model's self-learning, and language systems' gender structures. After dissecting these factors, the formation of gender bias in NLP becomes more apparent. As a result, strategies to address bias can be implemented more deliberately and effectively.

3.1. Data Imbalance

Data imbalance can be shown in restricted training data, where problems of quality, representativeness, and fairness occur. For example, the previously introduced difference in ELMo embeddings' accuracies is caused by different number of entities in the corpus. According to Zhao et al., the number of male entities in the corpus used to train ELMo is almost three times that of female entities [5]. Data imbalance is not only reflected in underrepresentation due to differences in the amount of data, but also in gendered sentiment analysis due to bias in the data itself. Kiritchenko et al. mentioned that their training data contains bias that associates specific genders and emotions [10]. When data imbalance exists, while the language model performs well in some overrepresented groups, it performs badly in underrepresented groups; thus, further leading to underrepresentation and unfair results. To reduce bias due to data imbalance, it is critical to ensure that the amount of data is equal for all groups and that random sampling techniques are used. In addition, automated algorithms need to be implemented to detect possible biases in the data collected.

3.2. Annotator Bias

Annotator bias encompasses two critical aspects. Firstly, annotators may introduce their own bias to the model consciously or subconsciously, leading to the reinforcement of gender stereotypes [11]. Secondly, the lack of annotators from minority groups causes data to learn more about the biases held by annotators from major groups. Some annotators may not be sensitive to the issue of gender bias, so informing and educating them on this could reduce bias that they are unaware of. Experts from related fields, such as sociology and gender studies, can discuss ethical implications with annotators. In addition, if more people from diverse groups become annotators, they could also supervise each other to ensure that the data produced is unbiased and inclusive.

3.3. Model's Self-learning Bias

During the self-learning process, models may further amplify existing biases in the training data [12]. According to Nemani et al., this amplification is caused by the models' choice of loss objects during training [2]. Models exploit gender features since they prioritize prediction accuracy, and this bias amplification is hard to detect until it becomes consistent, resulting in a trade-off between reducing gender bias and achieving high accuracy [2]. Models' self-learning process may lead to discriminatory outcomes, perpetuate existing gender inequalities, and raise ethical concerns. To mitigate this, model structures need to be modified and re-trained by adjusting components of models, such as weights.

3.4. Language Systems' Gender Structures

Languages with different gender systems can also be a factor leading to gender bias in NLP. Many highly gender-encoded languages lack gender-neutral words, making it a challenge to be translated from less gender-encoded languages without gender bias. Not only nuance is lost in translation, but it would also output stereotypical results. To circumvent this, engineers should work on developing translation models that are more sensitive to gender-related nuances. They may also work with linguists to incorporate gender-neutral language alternatives and evaluation metrics that assess gender bias in translations.

The occurrence of gender bias in NLP is influenced by all four of these factors, which are interrelated and tend to exacerbate gender stereotypes and biases. In addition to researchers, society and government may play a crucial role by monitoring and supervising NLP development. For example, they can

DOI: 10.54254/2755-2721/52/20241234

establish regulations and a code of ethics for the NLP system. Additionally, continuing efforts in education can help reduce social biases embodied in data and everyday language.

4. Detecting and Quantifying Gender Bias in NLP

Detecting and quantifying gender bias in NLP is important as it evaluates the existence of bias from different aspects. The detection and quantification process involves bias from annotators and in word embeddings, bias in models, and bias in language expressions. Individual metrics apply to each of these aspects, yet they can only deal with a specific one. To better detect and quantify all existing gender biases in NLP, researchers should combine multiple metrics to have a more comprehensive and accurate assessment.

4.1. Bias from Annotators and in Word Embeddings

Psychological tests can be used to detect and quantify annotator biases. For example, the Implicit Association Test (IAT) is a psychological test which evaluates human biases by seeing how quickly they pair words up. According to Caliskan et al., when humans pair two concepts they consider to be similar, their response times differ from that of pairing two dissimilar concepts [13]. They found that female words such as "woman" and "girl" are more associated with the arts than mathematics by conducting IAT [13]. Human biases also exist in word embeddings. Building on IAT, Caliskan et al. introduced the Word Embedding Association Test (WEAT). First, they defined two sets of words related to two genders and another two sets of words related to occupation and family. Then, if the absolute value of the WEAT score between one gender group and one attribute group is high, then there is a strong association between certain a gender group to a certain attribute, indicating possible bias. The limitation of both IAT and WEAT is that they mainly focus on word associations and that they may not be able to capture nuances in context. In addition, the sets of words are predefined, so they may not cover all possible words related to a particular attribute.

4.2. Bias in Models

Despite bias by annotators and bias in word embeddings, bias in models is also a significant component of gender bias in NLP. To detect model-related bias, researchers may apply counterfactual evaluation. They would swap gender-related words, and quantify the difference between the original text and gender-swapped text to see any potential bias. One of the metrics used in the counterfactual evaluation is "inconsistency across genders" (Iacross), which measures inconsistency in instance pairs referring to different genders [14]. This inconsistency can be assessed by relevant distance metrics such as calculating the cosine similarity between vectors, which is a method inspired by previous work [14]. The counterfactual evaluation also has some limitations. For example, the gender-swapped context may contradict the fact, such as the gender of a historical person. This intervention of changing gender in sentences may also cause semantic changes that are unrelated to bias itself.

4.3. Bias in Language Expressions

In addition, bias in language expressions can be evaluated by gender attribution analysis. This assesses the deviation of gender-related information in the translation process. One well-known gender-related element is name, as its varied form across distinct language systems can represent different genders. Languages with grammatical gender systems would require an adjustment based on the gender of the names to assign genders for other words in the same sentence, so biased assumptions of genders in names may result in poor-quality translations. To detect bias in language expressions, Wang et al. have associated names with certain genders based on the actual frequency of association in the U.S. birth data [15]. They choose names that are associated with a specific gender more than three times as often as they are associated with the other gender [15]. Then, they calculate the absolute difference between the mean accuracy for male and female names to see if gender bias exists in the translations [15]. In this case, the potential disparity in the treatment between male and female names can be seen from the translation output. One limitation here is that names can be influenced by cultural differences, and names

DOI: 10.54254/2755-2721/52/20241234

being considered feminine in one language may not be feminine in another, as they may be arbitrary due to the randomness of languages themselves.

By rigorously analyzing the biases from annotators, data, word embeddings, and language expressions, researchers may be better at achieving their goal of mitigating the gender bias.

5. Bias Mitigation and Challenges

Recognizing the importance of addressing these biases, some debiasing methods and approaches have been employed to mitigate gender bias in NLP. Targeting on bias by annotators, the model, word embeddings, and language structures, NLP researchers are actively using innovative and composite methods to make sure that the diversity of the world is reflected, which contributes to building a more impartial society. Along with introducing the bias mitigation methods, some potential challenges are also posed for future researching purposes.

5.1. Bias by Annotators

To address gender bias related to annotators, diverse recruitment should be considered. For example, active recruitment of annotators of different demographic backgrounds would ensure the annotations and developments of models are more inclusive. According to Abbasi et al., "fairness by design" is a set of five rules that emphasizes the importance of always considering fairness when making annotations to prevent bias from penetrating into NLP [16]. Except for traditional machine metrics, fairness by design underscores the significance of annotator ethics for the NLP field and suggests that they need to self-prompt for potential bias while also having sociologists alert them [16]. In this way, the demographic bias people bring to the NLP field can be reduced.

5.2. Bias from models

For models' self-learning processes, regularization techniques can be implemented to address bias. One common way is to modify loss functions which indirectly promote fairness in NLP. They would penalize bias perceived. For example, Li et al. regulate their model using pairwise difference loss function and T-statistics loss function, which both evaluate the differences between groups and can let the model minimize the disparities between gender groups [17]. However, the changing of loss functions may harm models' performance on running speed, and there may also exist a trade-off between fairness and accuracy.

5.3. Bias from Word Embeddings

For word embeddings, debiasing algorithms focus on post-processing word embeddings, which modify the embeddings that may contain gender bias in pre-trained word embeddings. This process preserves useful information and necessary distinctions between words. Bolukbasi et al. first identified gender subspace, and then applied hard debiasing or soft bias correction [8]. During the hard debiasing process, they neutralize any gender-neutral words to let them have no effects in the gender subspace, and they make words outside the subspace have the same distance too all other words [8]. During the soft bias correction process, there is a trade-off between reducing gender bias and preserving the original embeddings so both semantic information and reduction of gender bias can be balanced [8]. One potential challenge here is that debiasing word embeddings may also unrelate some important information according to gender. For example, in medical research, the amount of medicine applied to biological genders may vary, so it is important to pertain information such as methods of treatment referring to gender while eliminating gender bias.

5.4. Bias from Language structures

Considering the variation of gender systems in different languages, many researchers have proposed their guiding principles for achieving gender-neutral machine translation. Piergentili et al. suggested three principles on the matter. Firstly, gender should not be expressed in the output if it cannot be assumed from the input [18]. Secondly, gender information from the source should be recognized and

DOI: 10.54254/2755-2721/52/20241234

preserved in the output translation [18]. Thirdly, masculine generics should not propagate from the source to the output [18]. Using named entities to tag genders is also an approach to identifying gender-neutral entities, which allows NLP systems to generate content using gender-neutral language when appropriate [19]. One challenge is that there exist cultural differences that are highly related to gender norms and language structures, so using the method described here might still be difficult to solve culture-related gender bias.

The core idea is to create an augmented dataset that reflects the original dataset but favors underrepresented genders. This augmented dataset is generated with gender swapping and aims to reduce bias by training the model on a gender-balanced dataset. During the data augmentation process, for each sentence in the original dataset, a gender-swapped version is created following the method outlined previously. Name anonymization can also be applied to the original sentence and the gender-swapped sentence to de-emphasize individual identities. By training both datasets, the NLP system would no longer be influenced by the existing gender bias in the original training data; thus, mitigating gender bias successfully.

6. Conclusion

This study delves into the issue of gender bias in NLP and reveals the limitations of current research. To reduce gender bias, annotators should be educated on ethics by sociologists, and more advanced techniques are needed to better deal with gender bias in word embedding and modeling. In addition, attention should be paid to gender systems in different languages and specialized methods should be implemented to reduce gender bias in translation. Not only are technical approaches critical when dealing with bias, but ethical and cross-disciplinary oversight is also essential. By critically selecting and evaluating previous studies, I have pointed out some future research areas for further study.

The issue of gender bias must be brought to the fore as it can result in individuals in certain gender groups being severely undermined, creating an unbalanced social situation. This study may have limitations because all the works referenced previously in this paper have focused on the gender binary system, which does not account for non-binary identities and other gender-diverse identities. In the future, as more data is collected from these identities, research on gender-related issues in NLP should incorporate these identities as well. In this way, fairness to gender in the NLP field can be better realized.

References

[1] Font J and Costa-jussà M 2019 Equalizing gender biases in neural machine translation with word embeddings techniques Preprint arXiv:1901.03116
[2] Nemani P, Joel Y, Vijay P and Liza F 2023 Gender bias in transformer models: a comprehensive survey Preprint arXiv:2306:10530 (Yericherla Deepak Koel, Farhama Ferdousi Liza)
[3] Nadeem A, Sbedin B and Marjanovic O 2020 Gender bias in AI: a review of contributing factors and mitigating strategies ACIS 2020 Proceedings 27
[4] Dastin J 2022 Amazon Scraps Secret AI Recuiting Tool that Showed Bias against Women vol 1
[5] Zhao J, Wang T, Yakstar M, Cotterell R, Ordonez and Chang K 2019 Gender bias in contextualized word embeddings Preprint arXiv:1904.03310
[6] Piazzolla S, Savoldi B and Ventivogli L 2023 Good, but not always fair: an evaluation of gender bias for three commercial machine translation systems Preprint arXiv:2306.05882
[7] Cabrera L and Niehues J 2023 Gender lost in translation: how bridging the gap between languages affects gender bias in zero-shot multilingual translation Preprint arXiv:2305.16935
[8] Bolukbasi T, Chang K, Zou J, Saligrama V and Kalai A 2016 Man is to computer programmer as woman is to homemaker? debiasing word embeddings Preprint arXiv:1607.06520
[9] Ali M, Sapiezynski P, Bogen M, Korolova A, Mislove A and Rieke A 2019 Discrimination through optimization: how Facebook's ad delivery can lead to skewed outcomes Preprint arXiv:1904.02095
[10] Kiritchenko S and Mohammad S 2018 Examining gender and race bias in two hundred sentiment analysis systems Preprint arXiv:1805.04508

DOI: 10.54254/2755-2721/52/20241234

[11] O'Neil C 2016 Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy vol 1
[12] Cho W, Kim J, Yang J and Kim N 2021 Towards cross-lingual generalization of translation gender bias ACM. pp 449-57 (Won Ik Cho, Nam Soo Kim)
[13] Caliskan A, Bryson J, Narayanan A 2017 Semantics derived automatically from language corpora contain human-like biases Science vol. 356 no. 6334 pp 183-6 (Joanna J. Bryson)
[14] Xie Z, Kocijan V, Lukasiewicz and Camburu O 2023 Counter-GAP: counterfactual bias evaluation through gendered ambiguous pronouns Preprint arXiv:2302.05674
[15] Wang J, Rubinstein B and Cohn T 2022 Measuring and mitigating name biases in neural machine translation ACL. Vol. 1 2576-90 (Benjamin I. P. Rubinstein)
[16] Abbasi A, Li J, Clifford G and Taylor H 2018 Make "fairness by design" part of machine learning Harvard Business Review
[17] Li Y, Wei X, Wang Z, Wang S, Bhatia P, Ma X and Arnold A 2022 Debiasing neural retrieval via in-batch balancing regularization Preprint arXiv:2205.09240
[18] Piergentili A, Fucci D, Savoldi B, Bentivogli L and Negri M 2023 Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges Preprint arXiv:2301.10075
[19] Saunders D and Olsen K 2023 Gender, names and other mysteries: towards the ambiguous for gender-inclusive translation Preprint arXiv:2306.04573