A Literature Survey on Methods of Debiasing Gender Stereotypes in Natural Language Processing

Katrina Li

Abstract

With the rise in use of Machine Learning (ML) to develop Artificial Intelligence (AI) tools, Natural Language Processing (NLP) has assumed a particular importance in our everyday lives. As NLP has become a cornerstone in producing popular, successful systems such as voice recognition and machine translation, it is ever more important that we also acknowledge the role NLP has in amplifying and perpetuating gender biases and stereotypes found in text and speech corpora. Beyond recognizing the presence of these biases, it is crucial that we investigate methods of mitigating them as well. As such, this literature survey will review modern approaches to debiasing gender stereotypes in NLP and provide related avenues for future studies.

1 Introduction and History

Natural language processing (NLP) research dates back to the 1940s where the first computer-based application related to NLP was Machine Translation (MT) (Liddy, 2001). Since then, the field of NLP has grown rapidly, expanding to areas such as voice recognition software (e.g. voice assistants like Siri, Alexa, and Cortana), machine translation, abusive language detection, visual recognition tasks (e.g. captioning), and resume-reviewing (Hoy, 2018; Nimbekar et al., 2019; Olive et al., 2011; Park et al., 2018; Vinyals et al., 2015).

However, the data, training corpora, algorithms, and embedding techniques used in many NLP systems contain gender biases that can create detrimental scientific and ethical issues (Bender and Friedman, 2018; Bolukbasi et al. 2016; Caliskan et al., 2017; Garg et al., 2018; Sun et al., 2019). For example, as a result of biases present in training data, resume-review systems may rank female candidates as less qualified than their male counterparts for software engineering jobs even though the only distinguishing factor between the candidates is gender.

Although gender bias is unquestionably a complex issue, the increasing prevalence of AI involving NLP systems in our everyday lives compels the need to find methods to address the issue both in the short term and long term. This literature survey thus reviews several approaches to mitigating gender bias and stereotypes and presents issues that call for further study and investigation.

2 Review of the Literature

There are many methods that have been proposed to debias gender stereotypes in NLP. The ones explored in this literature survey are as follows: debiasing training corpora, debiasing gender in word embeddings, debiasing by adjusting algorithms, and data transparency.

2.1 Debiasing Training Corpora

Methods to debias training corpora include counterfactual data augmentation and bias fine-tuning.

2.1.1 Counterfactual Data Augmentation

While previous approaches focused on mitigating gender stereotypes in English, these approaches often produce grammatically incorrect sentences in morphologically rich languages where masculine-inflictions and feminine-inflections must be taken into account. Specifically, in languages like Spanish and Hebrew, if the gender of one word

changes, the others in the sentence must also change to maintain morpho-syntactic agreement. For example, if we wish to replace the "El" in Sentence 1 "*El ingeniero alemán es muy experto"* with "La" to create the feminine version of the sentence (Sentence 2), "*ingeniero*" must change to "*ingeniera*", "*aleman*" to "*alemana*" and "*experto*" to "*experta*". (Zmigrod et al., 2019).

Zmigrod et al. (2019) proposed to introduce a method in which they create a Markov random field with an optional neural parameterization that, when changing the grammatical gender of particular nouns, can infer how a sentence must change to maintain agreement between morphology and syntax. Specifically, they focus on a four-step process: 1) analyze the sentence (parsing, morphological analysis, lemmatization), 2) intervene on a gendered word, 3) infer the new-morphosyntactic tags (primarily through the proposed Markov random field), and 4) reinflect the lemmata to their new forms. The Markov random field, built on a labeled dependency tree with associated part-of-speech (POS) tags, specifies a joint distribution across morpho-syntactic tag sequences. Rather than considering the actual words themselves, the model considered only the labeled dependency tree and POS tags. It then can be used to infer how the remaining tags should be modified to preserve morpho-syntactic agreement after a gendered term is updated.

Despite not having access to existing annotated corpus of pairs sentences that can be used as a control, intrinsic (whether the approach yields correct reinflections and morpho-syntactic tags) and extrinsic (the extent to which using resulting transformed sentences reduces gender stereotyping in neural language models) evaluation of the approach proposed proved promising: the approach achieved accuracies of 90 percent and 87 percent for form-levels and F1 scores of 82 percent and 73 percent for tag-levels for Spanish and Hebrew. After evaluating the approach on four different languages, they also found that, on average, their approach decreases gender stereotyping by a factor of 2.5 without sacrificing grammaticality.

Although there is success in this approach, there are still drawbacks that have yet to be addressed. For instance, for languages like Spanish where conjoined nouns do not normally need to have the same gender, the lack of co-reference information in the Markov random field model cannot tell when both nouns refer to the same person. This results in a failing to convert nouns that are indirect objects of verbs or noun-modifiers. Furthermore, the model overall did not perform as accurately for Hebrew in terms of sacrificing less grammaticality, hence another area in which the approach could be improved.

2.1.2 Bias Fine-Tuning

Gender bias also extends to abusive language detection models containing imbalanced training datasets. For example, "You are a good woman" are considered "sexist" when trained on an existing dataset likely due to the sentence containing "woman" (Park et al., 2018). Because of the limited unbiased data sets for tasks done by abusive language detection models and the bias present in datasets that are used to train many models, Park et al. (2018) found a way to use unbiased data sets for a similar task may exist to help debias other datasets: a process known as bias fine-tuning. In essence, before fine-tuning a model on a highly biased data set used to train for the target task directly, bias fine-tuning uses transfer learning from an unbiased data set to verify that it contains minimum bias first. Not only does this allow models to adequately train to perform a task, but it also prevents them from learning and perpetuating biases from training sets.

Evaluation of their method showed that bias fine tuning can effectively reduce gender bias by 90-98 percent, boosting the robustness of the models used. Despite its effectiveness, however, bias fine-tuning can lead to a decrease in classification performance. As such, this leaves room to further explore how their method could improve to mitigate bias yet preserve (or even increase) classification performance.

2.2 Debiasing Gender in Word Embeddings

A popular framework to represent text data as vectors in many machine learning and NLP tasks is word embedding. Given the widespread use of word embeddings, biases present in those embeddings tend to amplify. Thus, two ways to mediate these biases include removing gender subspace in word embeddings and learning gender-neutral word embeddings.

2.2.1 Removing Gender Subspace in Word Embeddings

Since gender definition words are geometrically separable from gender neutral words in word embeddings, embeddings can be debiased by solely removing the gender component from gender-neutral words. Bolukbasi et al. (2016) explores this idea by making gender-neutral words orthogonal to the gender direction, debiasing embedding vectors yet avoiding removing gender altogether. This methodology helps eliminate gender stereotypes, such as the association between "secretary" and "female", but keeps desired associations between words like "princess" and "female". Ultimately, the modified algorithm significantly reduced gender bias in embeddings without disrupting key properties such as the ability to solve analogy tasks through clustering related concepts. In turn, this prevents the amplification of gender bias.

2.2.2 Learning Gender-Neutral Word Embeddings

Another method in which to remove gender bias from word embeddings is through modifying training procedures to learn gender-neutral word embeddings. Zhao et al. (2018) proposed GN-GloVe, a gender-neutral variant of GloVe that can compel certain dimensions of word vectors to be free of gender influence while preserving gender information in other dimensions. They achieve this by minimizing the negative difference (i.e. maximizing the difference) between the gender dimension in male and female definitional word embeddings, as well as maximizing the difference between the gender direction and the other neutral dimensions in the word embeddings. Because the gender dimensions are optional (e.g. can be used or ignored), their framework allows for greater flexibility to any definition and measurement (Zhao et al., 2018).

Although both Zhao et al. (2018) and Bolukbasi et al. (2016)'s methods have demonstrated success in debiasing word embeddings, it is not clear whether this applies to words relating to sentiment, debiasing beyond binary gender, and applications to other languages where most nouns carry gender markers. As such, further work is needed.

2.3 Debiasing by Adjusting Algorithms

Algorithm adjustment methods, such as adjusting predictions in NLP systems, have also shown to be effective in debiasing gender stereotypes. The examples discussed include constraining predictions through Reducing Bias Amplification (RBA) and adversarial learning.

2.3.1 Constraining Predictions

Bias amplification can happen when structured prediction models are employed to take advantage of correlations between co-occurring labels and visual data, but embed social biases prevalent in online corpora inadvertently (Zhao et al., 2017). In instances where the activity "shopping" is over 30 percent more likely to involve females than males in a training set, models trained on that set could amplify that bias to predict a disparity of 60 percent at the time of the test.

Zhao et al. (2017)'s RBA aims to mediate this by introducing corpus-level constraints so that gender indicators co-occur with elements of the prediction task as much as or less than the frequency present in the original training distribution. For collective inference, the calibration constraint was combined with the original structured predictor and reweighed bias-creating factors in the original model through Lagrangian relaxation (Korte and Vygen, 2008; Rush and Collins, 2012, Zhao et al., 2017).

The method is effective, resulting in a 47.5 percent decrease in bias amplification for multilabel classification and 40.5 percent for visual semantic role labeling without sacrificing performance loss (Zhao et al., 2017). However, since only one method for measuring bias was presented, future work investigating how different predictors affect bias and bias measurement and deamplification remain.

2.3.2 Adversarial Learning: Adjusting the Discriminator

Although debiasing gender in word embedding methods have shown to be helpful, they are not directly applicable to dialogue systems as they are likely to force dialogue models to produce similar responses even when genders are different. As a result, both the diversity of the generated responses

and the performance of the dialogue models suffer. To address this problem, Liu et al. (2020) propose Debiased-Chat, an adversarial learning framework for training discourse models without gender bias while maintaining their performance.

To separate the unbiased gender features and the semantic features of gender-related speech, Liu et al. (2020) designed a disentanglement model trained with a specially-designed bias-free dialogue model that produced responses with clear unbiased gender features and eliminated responses with biased gender features. Through using unbiased gendered utterances as the training set, the extracted unbiased gender features act as a discriminator and provide all of the information needed to infer the gender of the utterance. This thus allows the disentanglement model to automatically extract unbiased gender features from a biased utterance, leaving biased features in the remaining semantic features.

From experimenting on two human conversation datasets (Twitter conversation dataset[1], Reddit movie dialogue dataset[2]), Liu et al. (2020) demonstrate that their model produces more diverse, engaging, and gender-specific responses than baseline models, thereby successfully reducing gender bias in dialogue models while maintaining response quality.

2.4 Data Transparency

Data transparency is another method with which can be used to address issues rising from gender bias and stereotypes in NLP. Bender and Friedman (2018) propose that data statements be adopted and widely used as a design solution and professional practice for NLP technologies in both research and development.

Their proposed data statement schema follows two forms: long form and short form. The former, suggested to be included in academic papers presenting new datasets and in system documentation, is recommended to provide the following: curation rationale, language variety (described with a language tag from BCP-47 and matching description), speaker demographic, annotator demographic, speech situation/setting, text characteristics, recording quality, other relevant information (e.g. demographic characteristics of curators), and a provenance appendix for datasets built out of existing datasets. For short form data statements meant to be included in any publication testing, training, or tuning a system, or system documentation, they include a pointer to the long form one, but do not replace it. Instead, Bender and Friedman (2018) suggest that the data statements are 60-100 word summaries that describe most of the main points included in the long form one.

Data transparency through data statements is meant to be feasible and easy for NLP technologists to adopt, but can still be a challenge to implement given the amount of time required to write carefully crafted data statements. Despite the cost of time, it is nonetheless worthwhile as embracing this practice can lead to short-term and long-term benefits for the scientific community, for industry, and for the public at large. From foregrounding how data represents and do not represent the world so that researchers can make more precise claims, to encouraging conscientious software development so that companies can make sure to address bias and inclusion, to building practices in NLP system development that serve interests of users and indirect stakeholders equitably, Bender and Friedman demonstrate that data statements are undoubtedly integral to debiasing our systems.

3 Conclusion and Future Directions

This paper has focused on the recent methods used to mitigate gender bias and gender stereotypes in NLP. From debiasing training corpora to debiasing gender in word embeddings to debiasing by adjusting algorithms to adopting data transparency, the methods presented have all been crucial to resolving the complex issue of removing gender bias. However, it is important to acknowledge that the scope of this paper is limited: other approaches such as gender tagging (e.g. Bordia and Bowman, 2019), privacy preservation (e.g. He et al., 2021), and reducing biases in sentence-level representation (e.g. Liang et al., 2020) were not discussed, but are nonetheless worth looking into.

While all of these methods are certainly a step forward in mitigating gender bias, none of them are

[1] https://github.com/Marsan-Ma/chatcorpus/ (Liu et al., 2020)

[2] Dodge et al., 2015

perfect: there is still work to be done. Two directions in which future studies involving mitigating gender bias and stereotypes can move toward include expanding to non-binary genders and associated personal pronouns and modern-day colloquialism (slang).

Non-Binary Genders and Associated pronouns. Many of the studies presented focus on binary genders, but there is yet more work to be done involving debiasing non-binary genders in NLP. Outside of "male" and "female", there is still question whether bias mitigation methods work on non-binary and non-cis-heteronormative genders. In the same vein, the effectiveness of these methods in debiasing when it comes to data involving personal pronouns such as "they" is also brought to question. Thus, as we debias existing systems with regards to binary genders and pronouns of "he/him/his" and "she/her/hers", we need to also make sure that NLP systems reflect and represent identities beyond those as well.

Modern-Day Colloquialism (Slang). Language is ever changing and developing. As certain words are "retired" from common day speech and other words are used in new ways, previous and current models with which gender debiasing were built on may not be sufficient to address new issues that arise because of modern-day colloquialism, hence leaving a gap in the language-dependent models that we rely on daily. For example, words like "bro" and "bruh" may once have been associated more with males than females, but recently, these words have become increasingly gender neutral. As such, current methods to debias NLP systems should find ways to adapt to accommodate this change as to not perpetuate previous gender biases. Furthermore, with the increase in use of words like "Karen" (a term used to describe white women who are perceived as entitled and demanding) in everyday speech and the lack of a term for a non-female "Karen", debiasing methods in NLP developed using current speech datasets need to make sure to not perpetuate misogynistic biases. All in all, as our methods of communication evolve, our methods of debiasing language-based systems must evolve too.

References

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

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. In *Neural Information Processing Systems (NIPS'16).*

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics Derived Automatically from Language Corpora Contain Human-Like Biases. *Science*, 356(6334):183–186.

Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston. 2015. Evaluating prerequisite qualities for learning end-to-end dialog systems. *arXiv preprint arXiv:1511.06931.*

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.

Peiyang He, Charlie Griffin, Krzysztof Kacprzyk, Artjom Joosen, Michael Collyer, Aleksandar Shtedritski, and Yuki M. Asano. 2021. Privacy-preserving Object Detection. *arXiv preprint arXiv:2103.06587*.

Matthew B. Hoy. 2018. "Alexa, Siri, Cortana, and more: an introduction to voice assistants." *Medical reference services quarterly,* 37(1): 81-88.

Bernhard Korte and Jens Vygen. 2008. *Combinatorial Optimization: Theory and Application.* Springer Verlag.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*.

Elizabeth D., Liddy. 2001. Natural Language Processing.

Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Mitigating gender bias for neural dialogue generation with adversarial learning. *arXiv preprint arXiv:2009.13028.*

Rohini Nimbekar, Yoqesh Patil, Rahul Prabhu, and Shainila Mulla. 2019. "Automated Resume Evaluation System using NLP." In *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, pp. 1-4. IEEE.

Joseph Olive, Caitlin Christianson, and John McCary, eds. 2011. *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation*. Springer Science & Business Media.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Empirical Methods of Natural Language Processing (EMNLP'18).*

Alexander M. Rush and Michael Collins. 2012. A Tutorial on Dual Decomposition and Lagrangian Relaxation for Inference in Natural Language Processing. *Journal of Artificial Intelligence Research,* 45:305– 362.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pages 3156–3164.

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. In *Empirical Methods of Natural Language Processing (EMNLP'17).*

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and KaiWei Chang. 2018. Learning Gender-Neutral Word Embeddings. In *Empirical Methods of Natural Language Processing (EMNLP'18)*.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571.*