CS447 Literature Review: Gender Bias in NLP

NaveenKumar ConjeevaramBaskaran,
nc42@illinois.edu

December 9, 2024

Abstract

This review explores gender bias in natural language processing (NLP). Specifically, it focuses on large language models (LLM), reviewing four articles and expressing the findings to dig deeper into the issue of gender bias and discrimination.

1 Introduction

Gender bias has long been studied in computer systems and other domains. Large Language Models, one of the recent disruptive innovations, are typically trained on a large amount of language texts. Any human language text, like English and Chinese, in its literal sense can exhibit gender bias which means generating sentences with intentional or unintentional preference for one social group over another. LLMs are typically trained on large amounts of data collected from search engines, online forums, websites, and so on which are very very large amounts of human language texts. The training data collected from different cultures can demonstrate different types and levels of biases. This review attempts to explore the answer to the below question: "*Are LLMs capable of detecting gender bias and can they tailor the word generations without compromising on the context?*

To answer the above question, four papers related to this topic of gender bias were chosen. The relevant sections are summarized and explained accordingly. Finally, a conclusion is provided based on how each paper contributed to answering the above question and what future efforts would be needed for the same.

Gender bias in LLMs such as ChatGPT could harm half the global population if it is widely adopted. Understanding public perceptions toward gender bias in LLMs is crucial to ensuring policies and regulations are relevant and effective in meeting people's needs. In the meantime, LLMs are trained on data collected from search engines, online forums, websites, and so on. Thus, LLMs can reflect and even amplify existing biases in human language. Social biases exist in varying forms in different cultures. LLMs trained on data collected from different cultures may also demonstrate different types and levels of biases.

2 Background

The initial intent with the list of the chosen four papers was to explore the topic of gender bias with the surface-level idea of considering the explicit bias type in sentences considering men versus women gender types. After reviewing the papers, it educates us on the term gender, other social groups besides men and women, and interestingly various bias types like implicit bias, allocative

bias, representational bias, and so on. Reading the paper further will expose these findings which can be considered very informative and eye-opening in the context of LLMs.

3 Paper 1: Dissecting Biases in Relation Extraction

Relation Extraction (RE) in NLP is a technique that helps understand the connections between entities mentioned in texts. In a world brimming with unstructured textual data, RE is an effective technique for organizing information, constructing knowledge graphs, aiding information retrieval, and question-answering. There could be biases exposed by these RE systems and this paper proposes a method to detect these biases. The bias can occur at any stage of the NLP pipeline. This paper introduces two types of biases namely allocative and representational. *Allocative biases* occur when resource distribution decisions are made that disproportionately favor or disadvantage particular groups (e.g., job application screening or loan approval disparity). *Representational bias* often reflects and perpetuates stereotypes and societal prejudices between groups and certain features (e.g., women and lexicon about marriage and parenthood). These biases underline the importance of ethical considerations in NLP model design and deployment to prevent perpetuating societal inequities. Blodgett et al. (2020) show that existing works in NLP mainly focus on *representational* biases while the *allocative* ones are often overlooked.

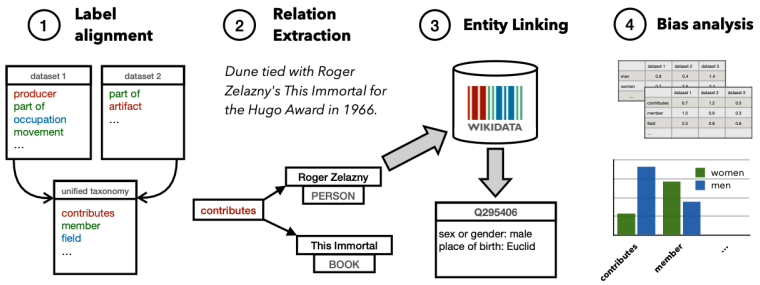

Figure 1: Overview of proposed RE Methodology

3.1 Relation Extraction Methodology

The paper Stranisci et al. (2024) proposes a Relation Extraction (RE) pipeline (as shown in Figure 1) which uses a four-step procedure to detect biases related to gender and place of birth. In this procedure, the method extracts triplets (subject, object, relation) in the texts where there is at least one person involved in the entities. Then the socio-demographic data about the person is collected from the available Wikidata. With these two elements available, the bias analysis is conducted by investigating any imbalance in the distribution of relations across different social groups (e.g., men versus women).

3.2 Experiment

The four steps described in Section 3.1 are used in three commonly adopted RE datasets SRED, CrossRE, and NYT, and predictions of the popular ReLiK model. Two categories of experiments were performed: 'Zero-shot', where Re- LiK is pre-trained on SREDFM and directly evaluated on

CrossRE and NYT; and 'fine-tuning', where ReLiK is both pre-trained on SREDFM and fine-tuned on the target dataset.

3.3 Social Bias Analysis

The allocative and representational biases are determined in the training datasets and predictions. The reporting is performed based on the analysis of two biases: 'gender' bias (men versus women) and additionally on 'place of birth' bias (Global North versus Global South). The findings reveal that the social group of women is *underrepresented* versus men and the same is applicable for the individuals from Global South versus Global North. This underrepresentation though is a concerning factor some limitations need to be considered. The gender considered was only binary based on the Wikidata but this approach is bypassing the gender who identify them as non-binary. Another limitation that may skew the result towards the Global South population is due to the data about them available at the time of this paper's research. Considering these limitations the bias statement about the under-representation needs to be carefully reviewed and mitigated with future work efforts to collect more Wikidata on those populations.

4 Paper 2: Discrimination Detection on Actor Level using Linguistic Discourse Analysis

This paper Urchs et al. (2024) explores the aspect of gender bias and especially discrimination using automated discrimination detection in text. One such method is referred to as linguistic discourse analysis (LingDA) and is explored further. The paper starts by defining bias, gender, and discrimination and various perspectives on the same. Bias against a particular gender creates harmful notions and discrimination against that gender. The LLM training data can contain stereotypes, biases, and discriminatory patterns. These can be reproduced by the models when used on production data. The paper further talks about two approaches to mitigate the stereotypical bias and discriminatory patterns from the existing training data which the LLMs are typically trained on. One approach is to clean the data upstream by removing these biased data before training the model. The alternate approach is to remove the biased data downstream from the model's output.

4.1 Gender - a different definition

The term gender can be considered from three different perspectives: linguistic gender, sex, and social gender. The paper argues that the notion of binary gender and sex is flawed as intersex humans exist. So, the scope of the study is targeted at social gender which has three groups: *man* for those who represent male identifying human group, *woman* as a female-identifying group, and *non-binary* for the pending groups who can't belong to the above two.

4.2 LingDA Pipeline

The paper further explains the methodology for automatic discrimination detection in text by using linguistic discourse analysis (LingDA). LingDA method uses two qualitative features which entail the semantics surrounding the text: *Nomination* (identifying the person-named-entity and how they are named Knobloch (1996)) and *Predication* (detecting the traits, characteristics, qualities, and

features Kamlah Wilhelm (1996)). A pipeline as shown in Figure 2 is created that combines nomination and predication using Information Extraction (IE) techniques which stores the predication in a knowledge base and this is further used to analyze the whole text for discrimination.

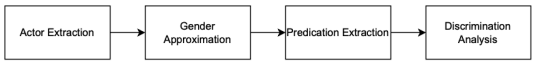

Figure 2: Visualization of the flexible and language agnostic pipeline

4.3 Discrimination Detection and Validation

The knowledge base from the above section 4.2 is extended by the sentiment of each predicated sentence and gender-coded words used in the predication. A value between -1 (very negative sentiment) and 1 (very positive sentiment) is assigned to each sentence. A Naive Bayes classifier trained on movie reviews is used to determine if the gender-coded words occur in the predication. A report is then generated to understand the spread of these words across the three genders. For testing our pipeline, we generate three texts with ChatGPT that contain several actors, with at least one respectively using feminine, masculine, or gender-neutral/non-binary pronouns.

The paper indicated that the report was able to distinguish sentences with very negative sentiments which can then be excluded from the model training. The strength of this pipeline is that it focuses on actors and not the text as a whole and results in determining more subtle discrimination. The pipeline is adaptable to different languages and also can be scaled from single texts to whole corpora level to fully report on the discourse. The pipeline was able to generate sentences for non-binary genders which is inclusive and eliminates any potential as many other methods would consider only binary genders.

5 Paper 3: Gender Bias between PLMs and Annotators

This paper Zhu et al. (2024) explores gender bias from the context of Pretrained Language Models (PLMs) and Human Annotators. There is no denying the fact that Pretrained Language Models (PLMs) have achieved success in various NLP tasks, but they do present safety problems like offensive language, social bias, and toxic behaviors. Of these issues, social bias, more specifically gender bias for two reasons: one, it is implicit and subtle. Second, there are different perspectives on bias across gender groups and within groups. Human annotators can unintentionally introduce gender bias and PLMs can carry forward the gender bias if the input data it is trained on, carries the bias. The exploration is made to answer the question **do PLMs and annotators share the same gender bias?**

5.1 Contextualized Gender Bias

This paper defines a metric called Contextualized Gender Bias (CGB) which serves to measure the implicit gender bias in both PLMs and annotators. The sentences are modified to replace the gender-specific words with MASK tags and PLMs and annotators are asked to fill in the MASK tag with either male or female words. If the MASK tag is filled with gender-specific gender words equally (e.g., like MOTHER and FATHER for childcare), then it is termed as unbiased context. Generally,

humans learn to replicate behavior from the social and cultural context around us based on what we see in news, media, or personal experiences. Based on this, PLMs and annotators replace the MASK tag with MOTHER for the above example and this over-association is consistent in both PLMs and annotators.

The paper introduces CGBDataset which contains 20K Chinese words from news articles and is further divided into Measuring Sentences where no clues can be inferred for gender words and Objective Sentences which provide definitive clues.

5.2 Measurement Design

To measure CGB from annotators, three humans were chosen (two female and one male) who were expected to have low gender bias and were provided with 200 sentences with Measuring Sentences and Controllable Questions. As shown in 3 the fine-grained correlation score is validated to measure CGB. In the end, annotators show no bias for measuring questions and gender bias (male specifically for this example sentence) for the unbiased contextual texts.

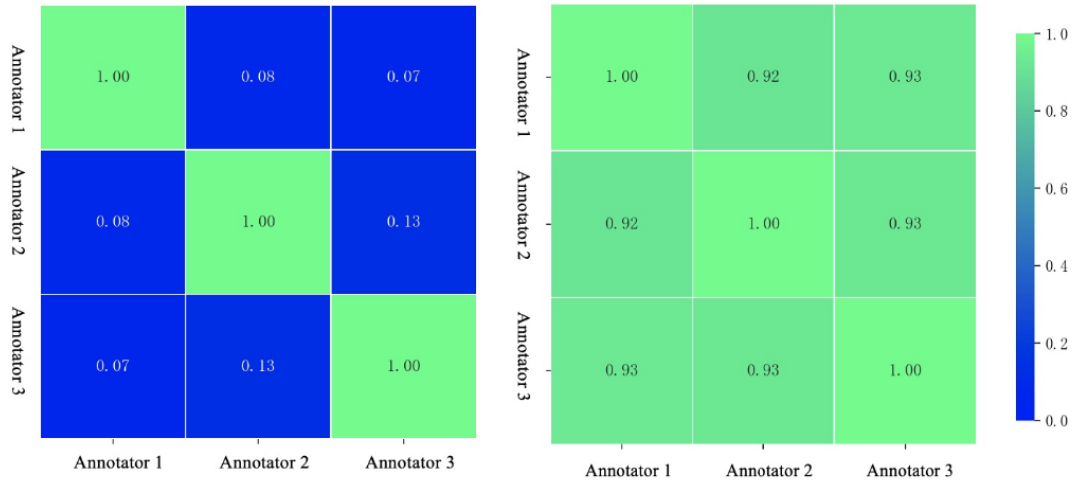

Figure 3: The left diagram shows the correlation of Measuring Questions among three annotators. The right diagram shows the correlation of Accuracy Controllable Questions among three annotators. Pearson's r is calculated as a correlation.

For measuring CGB from PLMs, below three widely used models were chosen:

- BERT (Devlin et al., 2019)
- RoBERTa (Liu et al., 2019)
- ELECTRA (Clark et al., 2020)

For each sentence S in CGBDataset, each PLM will give a female word probability $p_f(S)$ and a male word probability $p_m(S)$. Then, the CGB score of a sentence $PB(S)$ measured by a PLM can be calculated as follows:

$$PB(S) = log\frac{p_m(S)}{p_f(S)}$$

A positive value indicates the PLM indexes the sentence towards males, while a negative value indicates the PLM indexes the sentence towards females. Table 1 below shows the results of CGB measured by different PLMs where all PLMs exhibit different CGB. The results indicate we get correct answers in Objective Sentences while remain low CGB in Measuring Sentences.

	BERT-base	BERT-wwm	BERT-wwm-ext	RoBERTa	ELECTRA
Accuracy of OS	0.819	0.809	0.823	0.842	0.502
Bias Score of MS	0.540	0.589	0.627	0.570	0.779
SD of PB(MS)	0.697	0.750	0.800	0.750	0.940

Table 1: Results of CGB measured by different PLMs. We show the accuracy of Objective Sentences (OS) and the bias score of Measuring Sentences (MS) measured by PLMs. We also show the standard deviation (SD) of PB(MS).

After comparing the CGB score from PLMs and Annotators, it seems most PLMs show the same implicit gender bias as annotators. Additionally, this current method did not include non-binary gender and their expressions which doesn't seem inclusive in considering our society population. Future work is needed to include the non-binary gender and a more diverse group of annotators can be considered.

6 Paper 4: Fairness Analysis of Human and AI-Generated Student Reflections Summaries

This paper Baghel et al. (2024) aims to review the analysis of student reflection summaries and if there are any gender biases exhibited between manual human-generated summaries versus AI-generated summaries. Student reflections are a key meta-cognitive technique where the students from various classes like STEM, Arts, Law, and so on give their views about the classes and the instructor's teaching methods and structures. Humans or AI systems will review these reflections and provide a summary that helps the instructors and students determine the topic of confusion. The reflections are usually in the language texts and due to large amounts available for certain classes in colleges with huge student populations, automated AI-extraction techniques are created. Human-generated summaries can be used for the small volume of reflections. The generated summaries are prone to have gender bias towards male or female students and this study focuses on female student reflections from STEM class reflections. A concept of Structural Topic Model (STM) is introduced to measure how closely the summaries expose the reflections from male and female genders. The following research questions are being explored by this study:

- RQ1 What differences, if any, are there between reflections from male or female students?
- RQ2 Are summaries biased towards any specific gender?
- RQ3 If so, what is the nature of the gender bias in reflection summaries?

6.1 Dataset for reflections

Bias in summaries is examined Huang et al. (2023) from the perspective of opinion diversity as each students have varying opinions on the classes. REFLECTSUMM Zhong et al. (2024) dataset is selected for this analysis as it contained student reflections, their summaries, and student demographic

information of which gender is the metadata that is on focus. Three types of summaries were annotated or generated: extractive, phrase-level extractive, and abstractive. These summaries were saved to a knowledge base for further analysis. STM technique is used to measure the variance of reflection topics versus summaries based on the gender metadata. STM is an unsupervised learning procedure where the model learns the distribution over a set of topics given the text (reflections and summaries).

Topic 22: tree, binari, travers, search
- Under-representing Female
1: how do you delete a black node vs. a red node from a red-black bst?
2: How to label and binary search tree. And the build
tree method in the binary tree code
Topic 13: point, big, runtim, collis
1: I was confused about the Big O runtime details. I would love further explanation on how we can determine the estimated runtime. I would also like to know any tricks to more easily determine Big O. Additionally, I do not understand the difference between Big O, Little O, theta, and tilde.
2: BFS - how to keep track of what is seen/unseen
Topic 3: abl, group, team, meet
- Over-representing Male
1: A04 and dividing work amongst team members
2: It was interesting to join groups and work together. It helped eliminate most confusion. And it was interesting to meet new people
Topic 42: class, today', assign, onlin
- Over-representing Male
1: I think that the part that was most confusing today was what we were supposed to do for the in class assignment in class 2b
2: Due dates for assignment 10

Table 2: Top Reflections for Discrepant Topics

6.2 Analyzing Reflection Topics

STM represents topics as 'probability distribution over words' and documents as 'probability distribution over topics'. It also provides a tool to estimate a regression model predicting the learned topic proportion from document metadata which can be used to examine the association between topics and particular genders. The findings reveal that there is only subtle differences in male and female reflections.

6.3 Analyzing Reflection Summaries

Similar to topic modeling, STM is used to learn the topic distribution from the summaries generated by human annotations and AI systems and determine their associations to the particular gender. Summaries would typically represent topics from both male and female students as the reflections are mostly similar for the topic. A summary's closeness to a gender indicates a bias towards that gender. To determine this, the average distance between summary topic distributions and their

corresponding reflection topic distributions per gender is calculated. A smaller average distance indicates closeness.

After analyzing the findings and documenting the top 4 discrepancies (as seen in Table 2) it is evident that AI abstractive summaries exhibit bias toward reflections from male students, while summaries from humans and AI extractive models do not show a consistent bias. We find that AI abstractive summaries appear to under-represent specific topics suggested by female students while over-representing pedagogical themes such as teamwork from male student reflections. There are limitations though with this methodology as the non-binary and self-identifying gender groups are not considered and they would be unrepresented from this study's findings. Future work is needed to collect data on those groups.

7 Conclusion

After reviewing the four papers chosen on this topic of Gender Bias in Natural Language Processing (NLP), it seems educative and opens the door for future work and further research questions. Gender bias is an ethical consideration that many computer systems and software attempt to address to detect, mitigate and potentially avoid, if possible. The challenge though lies with the gender bias having varying perspectives and that it being subtle or implicit or not obvious. All four papers emphasized this concept in detail. My original question was: **Are LLMs capable of detecting gender bias and can they tailor the word generations without compromising on the context?** The answer to this question has been answered partially in that, the LLMs are capable of detecting the gender-specific bias words and eliminate them in the training dataset so it doesn't make it to the model nor in the output of generated sentence texts. Gender itself is contradictory in some of the research papers here because only binary genders of male and female were considered without consideration for non-binary and self-identifying groups. Only Paper 2 4 above considered the non-binary gender and all other papers didn't consider them in scope either due to very limited or unavailable data about those non-binary groups. The papers also exposed many different bias types like allocative bias, representational bias, AI-extractive bias, AI-abstractive bias, implicit bias, explicit bias, and human bias which further insisted that gender bias detection and mitigation needs more in-depth consideration rather than explicit bias and binary gender. Paper 1 3 and Paper 2 4 propose a pipeline that explains how to curate the dataset, choose the available models, define metrics and analyze the results. Paper 3 5 and Paper 4 6 make a different attempt at comparing the gender bias exhibited by human annotators and automated AI agents (PLMs, AI-extractor systems). To conclude, the LLMs can detect gender bias and find ways to mitigate it though there are limitations that needs future work and contributions from future research.

References

Bhiman Baghel, Arun Balajiee Lekshmi Narayanan, and Michael Miller Yoder. 2024. A fairness analysis of human and AI-generated student reflection summaries. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 60–77, Bangkok, Thailand. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics, pages 5454–5476, Online. Association for Computational Linguistics.

Nannan Huang, Lin Tian, Haytham Fayek, and Xiuzhen Zhang. 2023. Examining bias in opinion summarisation through the perspective of opinion diversity. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 149–161, Toronto, Canada. Association for Computational Linguistics.

Lorenzen Paul Kamlah Wilhelm. 1996. *Die Elementare Prädikation*, pages 23–44. J.B. Metzler, Stuttgart.

Clemens Knobloch. 1996. *Nomination: Anatomie eines Begriffes*, pages 21–53. VS Verlag für Sozialwissenschaften, Wiesbaden.

Marco Stranisci, Pere-Lluís Huguet Cabot, Elisa Bassignana, and Roberto Navigli. 2024. Dissecting biases in relation extraction: A cross-dataset analysis on people's gender and origin. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 190–202, Bangkok, Thailand. Association for Computational Linguistics.

Stefanie Urchs, Veronika Thurner, Matthias Aßenmacher, Christian Heumann, and Stephanie Thiemichen. 2024. Detecting gender discrimination on actor level using linguistic discourse analysis. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 140–149, Bangkok, Thailand. Association for Computational Linguistics.

Yang Zhong, Mohamed Elaraby, Diane Litman, Ahmed Ashraf Butt, and Muhsin Menekse. 2024. ReflectSumm: A benchmark for course reflection summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13819–13846, Torino, Italia. ELRA and ICCL.

Shucheng Zhu, Bingjie Du, Jishun Zhao, Ying Liu, and Pengyuan Liu. 2024. Do PLMs and annotators share the same gender bias? definition, dataset, and framework of contextualized gender bias. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 20–32, Bangkok, Thailand. Association for Computational Linguistics.