Beyond Single-Score Evaluation: A Bidirectional Analysis of Gender Bias in Language Models

Author: Säde Amanda Pesti, Computer Science Bachelor's
Supervisors: Faiz Ali Shah, PhD; Ahmed Abdulmajeed A Sabir, PhD
Institute of Computer Science

UNIVERSITY OF TARTU
Institute of Computer Science

Introduction

Language models have become widely adopted across natural language processing (NLP) applications, demonstrating their capacity to perform a broad range of tasks. However, despite these advancements, previous research has consistently shown that these models exhibit biases. These biases raise concerns about fairness and ethical deployment, as they can reinforce stereotypes and lead to discriminatory behaviour by models.

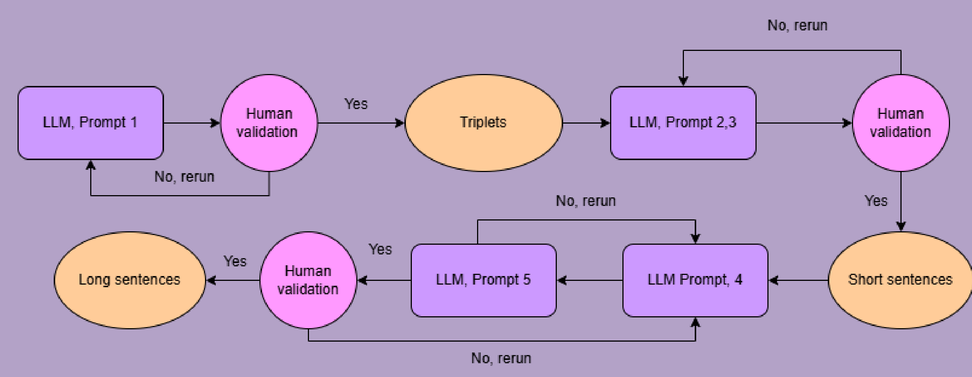
Transformer-based decoder models, such as large language models that power chatbots like ChatGPT, have gained widespread popularity. Meanwhile, BERT-family language models, which utilise an encoder-transformer architecture, remain relevant, especially for delivering high-quality results on classification tasks. Understanding the gender bias present in these models is crucial to preventing biased responses from the language models.

Previous work has studied learned social biases in language models, particularly in BERT-family models. However, most existing research relies on a single metric for bias evaluation using established benchmarks. In addition, limited research has examined bias bidirectionally, for example, by considering both the relationship from occupational context to pronouns and from pronouns to occupational context.

This thesis addresses this gap by investigating bias in both directions and measuring the degree of agreement between the resulting bias scores. This thesis makes three main contributions. Firstly, two new datasets are created for evaluating language models. Then, existing benchmark datasets, including WinoBias, Winogender, and GenderLex, are adapted for bidirectional agreement evaluation. Third, a bidirectional evaluation framework is applied to measure the consistency between context-to-pronoun and pronoun-to-context bias scores.

The results reveal that agreement varies across datasets and BERT-family models, indicating that unidirectional bias evaluation may be insufficient to fully capture the range of gendered associations encoded in bidirectional masked language models.

Dataset creation



Based on previous work with the gender bias benchmark dataset WinoBias, we focused on the most common 40 occupational title biases (e.g., *CEO* with males and *cleaner* with females), supported by the US labour force statistics. Therefore, both the long and short datasets followed a sentence structure in which the occupation was presented first. Optionally, there was some context, and the third part of the sentence was the action verb followed by optional context; the fourth part was the object noun followed by optional context; and last in the sentence was the gendered pronoun, which was either a male or female pronoun. To generate short and later long sentences for each occupation that would follow the defined structure, triplets that included the **occupation**, an **action verb**, and an **object noun** were created. To do that, LLM was used for keyword generation by giving 10 occupations at a time, in alphabetical order, along with a specific prompt. An example of a triplet generated would be: (laborer; shovel; dug), or (receptionist; phone; answered). The resulting datasets can be seen in the table.

Dataset	Size	Avg. Length	Median	Std. Dev.	Obj. Nouns	Act. Verbs
Winogender [8]	178	14.81	15	2.91	×	55
WinoBias [7]	785	12.55	12	1.98	×	187
GenderLex [9]	837	11.95	12	0.90	223	111
Thesis's short	400	7.91	7	1.57	359	283
Thesis's long	400	12.30	12	1.05	359	283

Results

For the pronoun-level bias the thesis's long dataset was the only dataset to have a slightly female-leaning pronoun bias, average of 54%. Thesis's short dataset showed a varied female preference, with the highest at 71.5% and the lowest at 25% indicating that the shorter sentences in the dataset did not provide enough context for the models to give biased responses Both the GenderLex and the WinoBias datasets showed a strong male pronoun preference, averaging 70% and 80%, respectively.

Bidirectional agreement scores showed the following results. The thesis's short dataset's object noun category bidirectional agreement was close to 50%, for the action verb category the average agreement was again close to random, averaging 49%, and the occupation category showed the strongest agreement on bidirectional bias, with and average of 58%.

For the thesis's long dataset for the object nouns the overall agreement was similar to that in the shorter dataset, averaging 48%. Similarly, action verb overall agreement for models was similar to the previous results, averaging 50%, and occupation had the overall bias agreement higher at 56%.

For the GenderLex dataset the object noun category showed bidirectional agreement close to 50%. The action verb category had high distilled model agreement, 80% and 74% for Distilled BERT and RoBERTa, respectively, and the occupation category had relatively high bias agreement scores, averaging 61%

WinoBias dataset showed the action verb category with overall agreement scores on average 56% and the occupation category showed a relatively high overall bias agreement score with 74%.

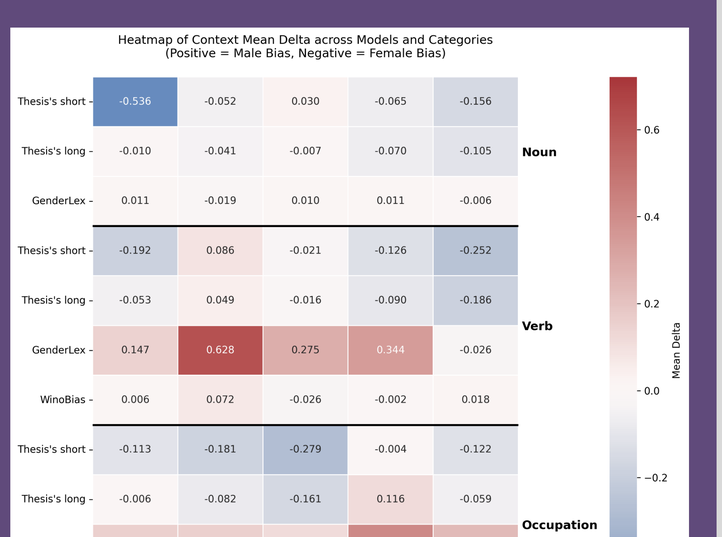
Out of 15 tests 8 were statistically significant for the object noun category, and 14 and 17 out of 20 tests for the action verb and occupation categories, respectively.

Table 7. Example of match (√) or mismatch (×) with both bias direction for occupations context Bias_c and pronoun Bias_p for all dataset for ModernBERT.

WinoBias						GenderLex					
Context	Δ_c	Bias_c	Δ_p	Bias_p	Match	Context	Δ_c	Bias_c	Δ_p	Bias_p	Match
developer	1.817	M	1.665	M	√	psychologist	-0.031	F	-1.002	F	√
supervisor	-0.042	F	0.796	M	×	conductor	1.274	M	-1.277	F	×
mover	0.531	M	1.709	M	√	mathematician	0.065	M	0.899	M	√
writer	0.374	M	0.155	M	√	doctor	0.551	M	0.423	M	√
attendant	-0.254	F	0.267	M	×	leader	0.568	M	-1.948	F	×
Thesis's short						**Thesis's long**					
Context	Δ_c	Bias_c	Δ_p	Bias_p	Match	Context	Δ_c	Bias_c	Δ_p	Bias_p	Match
salesperson	-0.739	F	-0.463	F	√	receptionist	-0.019	F	-2.229	F	√
sewer	0.224	M	-0.702	F	×	auditor	0.314	M	-0.175	F	×
clerk	0.103	M	0.915	M	√	nurse	-0.331	F	-1.401	F	√
cook	-0.104	F	-0.963	F	√	carpenter	0.002	M	1.023	M	√
supervisor	0.080	M	-0.324	F	×	accountant	0.485	M	-0.833	F	×

Table 8. Pronoun preference scores on bidirectional agreement on occupation context. The bidirectional agreement score is based on the WinoBias occupation category. Non-significant values (mean delta not different from zero at $p < 0.05$) are shown in gray.

	WinoBias (%)			GenderLex (%)			Thesis's long (%)			Thesis's short (%)		
Model	M	F	Agr	M	F	Agr	M	F	Agr	M	F	Agr
BERT	79.1	20.9	79.2	67.9	32.1	55.7	33.5	66.5	48.0	28.5	71.5	58.5
RoBERTa	73.9	26.1	77.3	79.0	21.0	69.3	53.2	46.8	57.0	75.0	25.0	52.5
ModernBERT	76.8	23.2	76.3	27.0	73.0	42.9	28.8	71.2	52.0	48.8	51.2	63.7
DistilBERT	91.5	8.5	70.6	89.2	10.8	71.1	73.8	26.2	58.5	68.8	31.2	48.0
DistilRoBERTa	80.9	19.1	64.5	87.8	12.2	66.8	47.8	52.2	66.5	54.2	45.8	65.2



Discussion – unidirectionality is not enough, women as passive agents

The main finding of the analysis is that, especially in the object noun and action verb categories, bidirectional agreement is low in both datasets. This indicates that the framework of using bidirectional agreement categories compared to the previous one-directional bias assessment could be crucial for understanding the biases embedded in BERT-family language models and in mask language models in general. The highest bidirectional agreement across datasets for the models was in the occupation category, suggesting that occupation is more similarly biased bidirectionally. Still, the bidirectional agreement scores were not very high, ranging from 50 – 80%, indicating that, for the occupation-level bias assessment, a bidirectional approach would still be beneficial for uncovering bias patterns that unidirectional assessments do not reveal. A clear example is attendant in the WinoBias dataset, the context-based signal leans female while the pronoun-based signal leans male, producing a directional mismatch that a single-direction could not have detected.

The heatmap shows model-level comparison per context. Across the board, the object noun category is the only female-leaning category, while the action verb category is moderately male-leaning, and the occupation category is the most male-leaning. Although the object noun and action verb categories in the thesis's datasets were similar in bias strength, the GenderLex object noun category was much more female-leaning than the GenderLex action verb category. In addition, the WinoBias dataset had slightly more male-leaning action-verb bias than the other model's object-noun bias.

If it is not the case that object nouns and action verbs are differently biased due to data or evaluation characteristics, and if, across the board, action verbs are still more male-leaning than object nouns due to biases in the models, then that could be explained by men being seen as more active agents by the models. In contrast, the models portray women as more passive agents, which could explain differences in model preference across contexts: women are more closely associated with object nouns, while men are more closely associated with action verbs. This finding aligns with previous work in philosophy, where the famous feminist philosopher Simone de Beauvoir writes that: "Humanity is male, and man defines woman, not in herself, but in relation to himself; she is not considered an autonomous being. Woman, the relative being..." .In addition, this reading is supported by Charlotte Knowles, who interprets Beauvoir's *The Second Sex* as follows: "Similarly, for Beauvoir, women's complicity in their own subordination is expressed in the way women cling to limited and limiting self-conceptions that reduce their agency by casting them into the role of the 'Other': man's passive and dependent counterpart". Furthermore, this is explained by the passivity-activity dichotomy within the femininity masculinity binary, in which identifying as a passive object in relation to an active agent is seen as constitutive of a woman's identity. In contrast, the active agent whose identity is defined in and by itself is the man's, as Beauvoir seems to argue.

Acknowledgements

Hereby, I acknowledge my Bachelor's thesis supervisor, Ahmed Sabir, for providing me with weekly thorough feedback for almost the entire academic year, for providing the reading materials and explanations needed to write the thesis, and for offering help with the practical part of the thesis.