Clustering versus Projection Debiasing for Contextualized Embeddings on WinoBias and CrowS-Pairs Gender Benchmarks

Assignee Research

June 11, 2026

Abstract

Contextualized word embeddings have been replacing standard embeddings as the representational knowledge source of choice in NLP systems. Since a variety of biases have previously been found in standard word embeddings, it is crucial to assess biases encoded in their replacements as well. Focusing on BERT (Devlin et al., 2018), we measure gender bias by studying associations between gender-denoting target words and names of professions in English and German, comparing the findings with real-world workforce statistics. We mitigate bias by fine-tuning BERT on the GAP corpus (Webster et al., 2018

1 Introduction

This paper examines: Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. Research question: How do clustering-based debiasing techniques for contextualized embeddings compare to projection-based methods in terms of accuracy on the WinoBias and CrowS-Pairs gender bias benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.7/10.

3 Results

13 papers retrieved. 14 claims extracted; 13 independently verified. Quality review score: 8.7/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
A template-based corpus in English and German was created to measure gender bias with respect to different profession gr	✓	0.21
The dataset and code for all experiments are publicly available at https://github.com/marionbartl/gender-bias-BERT.	✓	0.26
The method of querying BERT's underlying MLM (Masked Language Model), proposed by Kurita et al. (2019), can be used for	✓	0.27
The BERT language model does not only encode biases that reflect real-world data, but also those that are based on stere	✓	0.22
A technique on BERT, previously applied on ELMo (Peters et al., 2018; Zhao et al., 2019), was shown to be successful for	✓	0.19
The cross-lingual transfer of a bias measuring method proposed for English is impaired by the morphological marking of g	✓	0.26
Gender bias is the systematic unequal treatment on the basis of gender (Moss-Racusin et al., 2012; Sun et al., 2019).	✓	0.27
Gender bias occurs when one gender is more closely associated with a profession than another in language use, resulting	✓	0.25
Gender bias, irrespective of whether it is representative of real-world data, may lead to allocational harm, because mal	✓	0.35
The Huggingface transformers library (Wolf et al., 2019) for PyTorch with a default random seed of 42 was used for all e	✓	0.27
A pre-trained BERTBASE model (Devlin et al., 2018) with a language modelling head on top was used for bias evaluation an	✓	0.26
For English, the tokenizer and model are loaded with the standard pre-trained uncased BERT-BASE model.	✓	0.26
For German, the cased model provided by DBMDZ was used.	×	0.12
The method for measuring bias used in this work is based on the prediction of masked tokens and relies on masking tokens	✓	0.29

References

- http://arxiv.org/abs/2101.09523v1
- http://arxiv.org/abs/1901.03116v2
- http://arxiv.org/abs/2010.14534v1