Do clustering-based debiasing techniques maintain higher semantic textual similarity scores than projection-based methods across

Assignee Research

June 11, 2026

Abstract

In comparison to the numerous debiasing methods proposed for the static noncontextualised word embeddings, the discriminative biases in contextualised embeddings have received relatively little attention. We propose a fine-tuning method that can be applied at token-or sentence-levels to debias pre-trained contextualised embeddings. Our proposed method can be applied to any pretrained contextualised embedding model, without requiring to retrain those models. Using gender bias as an illustrative example, we then conduct a systematic study using several state-of-the-art (SoTA) contextualised repr

1 Introduction

This paper examines: Debiasing Pre-trained Contextualised Embeddings. Research question: Do clustering-based debiasing techniques maintain higher semantic textual similarity scores than projection-based methods across diverse domain benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.5/10.

3 Results

10 papers retrieved. 9 claims extracted; 8 independently verified. Quality review score: 8.5/10.

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
There are numerous debiasing methods proposed for static non-contextualised word embeddings.	✓	0.22
The discriminative biases in contextualised embeddings have received relatively little attention.	✓	0.31
A fine-tuning method is proposed that can be applied at token-or sentence-levels to debias pre-trained contextualised em	✓	0.38
The proposed debiasing method can be applied to any pre-trained contextualised embedding model without requiring to retr	✓	0.37
Gender bias is used as an illustrative example in the study.	×	0.13
The study evaluates the level of biases encoded in different contextualised embeddings before and after debiasing using	✓	0.36
Applying token-level debiasing for all tokens and across all layers of a contextualised embedding model produces the bes	✓	0.39
There is a trade-off between creating an accurate vs. unbiased contextualised embedding model.	✓	0.33
Different contextualised embedding models respond differently to the trade-off between accuracy and unbiasedness.	✓	0.25

References

- https://doi.org/10.18653/v1/2021.eacl-main.107
- https://doi.org/10.1145/3543507.3583206
- https://doi.org/10.1145/3631326