What is the effect of transferring English-derived debiasing methods on the alignment quality of multilingual sentence embeddings

Assignee Research

June 11, 2026

Abstract

This paper investigates the transferability of debiasing techniques across different languages within multilingual models. We examine the applicability of these techniques in English, French, German, and Dutch. Using multilingual BERT (mBERT), we demonstrate that cross-lingual transfer of debiasing techniques is not only feasible but also yields promising results. Surprisingly, our findings reveal no performance disadvantages when applying these techniques to non-English languages. Using translations of the CrowS-Pairs dataset, our analysis identifies SentenceDebias as the best technique acros

1 Introduction

This paper examines: Investigating Bias in Multilingual Language Models: Cross-Lingual Transfer of Debiasing Techniques. Research question: What is the effect of transferring English-derived debiasing methods on the alignment quality of multilingual sentence embeddings for French and German?.

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. 4 claims extracted; 4 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
Cross-lingual transfer of debiasing techniques is feasible and yields promising results in multilingual models.	✓	0.34
There are no performance disadvantages when applying debiasing techniques to non-English languages in multilingual BERT	✓	0.34
SentenceDebias is the best technique for reducing bias in mBERT across different languages, reducing bias by an average	✓	0.29
Debiasing techniques with additional pretraining exhibit enhanced cross-lingual effectiveness, particularly in lower-res	✓	0.35

References

- https://doi.org/10.18653/v1/2023.emnlp-main.175
- https://doi.org/10.1145/3560815
- https://doi.org/10.17863/cam.30462