# Gender Classification Accuracy and Fairness in Contextualized Models Under Static Embedding Debiasing on BiasBios

Assignee Research

June 11, 2026

**Abstract**

Neural machine translation has significantly pushed forward the quality of the field. However, there are remaining big issues with the output translations and one of them is fairness. Neural models are trained on large text corpora which contain biases and stereotypes. As a consequence, models inherit these social biases. Recent methods have shown results in reducing gender bias in other natural language processing tools such as word embeddings. We take advantage of the fact that word embeddings are used in neural machine translation to propose a method to equalize gender biases in neural mach

## 1 Introduction

This paper examines: Equalizing Gender Biases in Neural Machine Translation with Word Embeddings Techniques. Research question: How do static embedding debiasing techniques affect the gender classification accuracy and fairness metrics of contextualized models on the BiasBios dataset?.

## 2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.7/10.

## 3 Results

9 papers retrieved. 11 claims extracted; 10 independently verified. Quality review score: 7.7/10.

## 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 |
|---|---|---|
| Improvements in combination of approaches such as attention and translation systems algorithms like the Transformer have | ✓ | 0.16 |
| Models trained with human-generated corpora learn social biases and stereotypes from the data. | ✓ | 0.16 |
| Word embeddings are a vector representation of words. | × | 0.12 |
| The presence of biases in the data can directly impact downstream applications and are at risk of being amplified. | ✓ | 0.15 |
| The word 'friend' in the English sentence 'She works in a hospital, my friend is a nurse' would be correctly translated | ✓ | 0.32 |
| The word 'friend' in the English sentence 'She works in a hospital, my friend is a doctor' would be incorrectly translat | ✓ | 0.33 |
| This study provides progress on the recent detected problem of gender bias in machine translation (MT). | ✓ | 0.20 |
| The progress towards reducing gender bias in MT is made in two directions: defining a framework to experiment, detect an | ✓ | 0.26 |
| This is the first study in proposing debiasing techniques for MT. | ✓ | 0.18 |
| The language pair used for the experiments is English-Spanish. | ✓ | 0.17 |
| The training set consists of 16,554,790 sentences from a variety of sources including United Nations, Europarl, CommonCr | ✓ | 0.22 |

## References


- http://arxiv.org/abs/1908.02810v1
- http://arxiv.org/abs/1901.03116v2
- http://arxiv.org/abs/2606.07964v1