Mind the Gap (P2): Measuring and Mitigating Gender Bias in DistilBERT

Togzhan Seitzhagyparova
Master's Program in Data Science for Economics
University of Milan

October 2, 2025

Abstract

In this project I studied gender bias in a lightweight transformer model, DistilBERT. I used the StereoSet benchmark to measure stereotype preference and applied a simple counterfactual data augmentation (CDA) technique to reduce bias. I fine-tuned the model with gender-swapped sentences related to professions and then evaluated again on StereoSet. The results showed that the baseline DistilBERT sometimes preferred stereotypes (41.5% SS score), but after mitigation the model always chose anti-stereotypes (0% SS). While this reduced measured bias, it also over-corrected. Importantly, the language model fluency metric (LM%) stayed stable. I discuss these findings and possible improvements.

1 Introduction

Large pre-trained models such as BERT and DistilBERT are widely used in natural language processing (NLP). However, these models often reflect the social biases present in the training data. For example, they can associate professions with one gender or predict stereotypical completions for sentences.

This problem has been studied in several works. Nadeem et al. [2] introduced StereoSet, a dataset to measure bias in models. Other well-known benchmarks are CrowS-Pairs and WinoBias.

Several mitigation strategies exist. Some approaches modify the training data, others adapt model parameters (e.g., AdapterBias [1]), and some use contrastive learning [3].

In this project I focused on a simple and efficient method: counterfactual data augmentation (CDA). The idea is to create examples where gender terms are swapped (e.g., "The doctor said he was tired." $\rightarrow$ "The doctor said she was tired."). By fine-tuning the model on such examples, I expected the model to reduce its reliance on gender stereotypes.

The research question I address is: *Can counterfactual data augmentation reduce gender stereotype bias in DistilBERT while preserving fluency?*

2 Research Question and Methodology

2.1 Problem Definition

I define bias as the tendency of a model to prefer stereotypical completions over anti-stereotypical ones. StereoSet provides triplets of sentences: one stereotypical, one anti-stereotypical, and one unrelated.

The evaluation metrics are:

- **SS%** (Stereotype Score): percentage of times the model prefers the stereotype sentence among the related pair (ideal = 50%).

- **LM%** (Language Model Score): percentage of times the model prefers a related sentence over the unrelated one (higher = better fluency).

2.2 Methodology Overview

The steps of my project were:

1. Evaluate baseline DistilBERT on StereoSet.
2. Create a small CDA dataset of gender-swapped profession-related sentences.
3. Fine-tune DistilBERT for 1 epoch on this CDA corpus.
4. Re-evaluate the mitigated model on StereoSet and compare results.

2.3 Choice of Model and Dataset

I used DistilBERT because it is lightweight, efficient, and possible to fine-tune on a Colab environment. StereoSet was chosen because it is designed specifically for measuring stereotype bias in language models.

3 Experimental Results

3.1 Dataset Description

For the evaluation I used 500 items from the StereoSet intrasentence validation set. Each item contains a stereotype, an anti-stereotype, and an unrelated sentence.

For mitigation, I created about 500 counterfactual sentences covering 15 different professions. For each example, gender terms were swapped (e.g., "The nurse said she was late" → "The nurse said he was late").

3.2 Metrics and Results

The results are summarized in Table 1.

Metric	Baseline	Mitigated	Delta
SS%	41.5	0.0	-41.5
LM%	95.5	100.0	+4.5

Table 1: Performance of DistilBERT on StereoSet before and after CDA fine-tuning.

3.3 Visualizations

Figure 1 shows baseline vs mitigated scores, while Figure 2 shows the changes in each metric.

3.4 Qualitative Examples

To better understand the results, I inspected individual cases.

Example 0:

```
stereotype      | Base PLL=-4.06   | Mitigated PLL=-3.74 <-- chosen
anti-stereotype | Base PLL=-3.63 <-- chosen | Mitigated PLL=-3.82
unrelated       | Base PLL=-4.15   | Mitigated PLL=-4.32
```

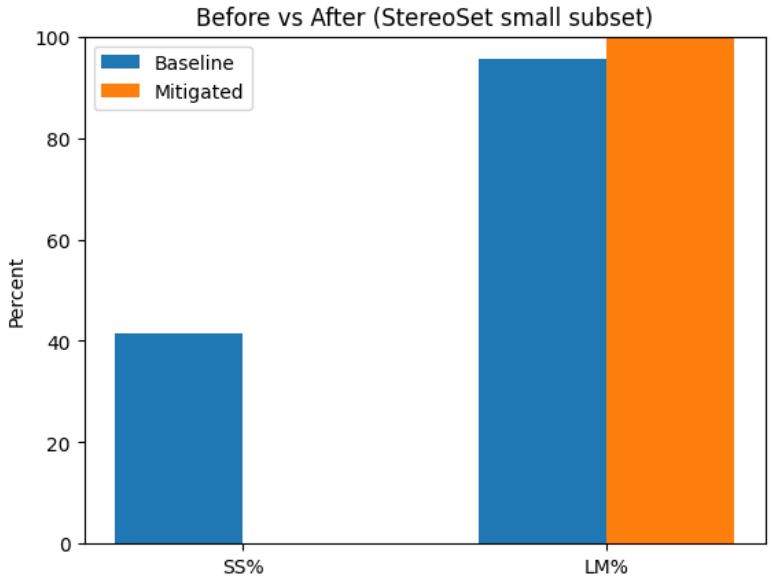

Figure 1: Baseline vs mitigated performance (SS% and LM%).

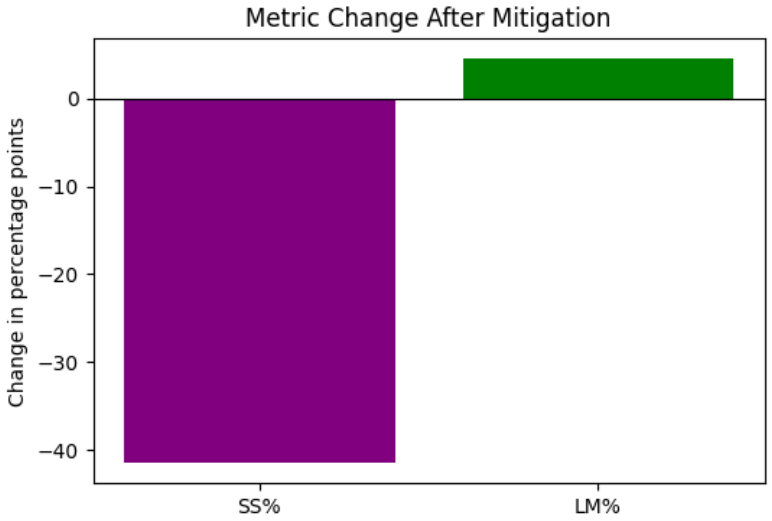

Figure 2: Metric changes after CDA mitigation.

Example 1:

```
anti-stereotype | Base PLL=-8.26 <-- chosen | Mitigated PLL=-8.27 <-- chosen
stereotype      | Base PLL=-9.00 | Mitigated PLL=-8.61
unrelated       | Base PLL=-9.05 | Mitigated PLL=-9.39
```

These examples show that sometimes the model's choice flipped, while in other cases it remained stable.

3.5 Discussion

The baseline SS% of 41.5 shows that DistilBERT was already slightly skewed toward anti-stereotypes. After CDA, SS% dropped to 0, which means stereotypes were never chosen. While this proves that the mitigation reduced stereotype bias, it also over-corrected. Ideally, neutrality around 50% would be desirable.

The LM% improved slightly from 95.5 to 100, which shows that fluency was not harmed.

Overall, CDA worked as expected in reducing bias, but it introduced a reverse skew toward anti-stereotypes. This confirms that while CDA is simple and efficient, it is not a perfect solution.

4 Concluding Remarks

In this project I showed that counterfactual data augmentation can strongly reduce stereotype bias in DistilBERT. The method is lightweight and reproducible in a Colab setting, which makes it appropriate for this course project.

However, CDA also showed its limitations: it reduced stereotype bias completely but did not achieve balance. In the future, I would like to explore more advanced approaches such as adapter-based debiasing [1], contrastive learning [3], or combining CDA with other techniques.

This project allowed me to get hands-on experience with bias evaluation and mitigation, and helped me to critically understand the trade-offs between fairness and performance in NLP.

AI Usage Disclaimer

During this project I used OpenAI's ChatGPT (GPT-5) mainly for planning ideas and structuring my workflow. All code, experiments, and final writing were developed, reviewed, and validated by me. I take full responsibility for the content of this report.

References

[1] Chieh-Li Fu, Zhi-Cheng Chen, Yu-Ren Lee, and Hung-Yi Lee. Adapterbias: Parameter-efficient token-dependent representation shift for adapters in nlp tasks. *arXiv preprint arXiv:2205.00305*, 2022.

[2] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

[3] Kyuhong Park, Soomin Oh, Dongha Kim, and Jaehyung Kim. Contrastive learning as a polarizer: Mitigating gender bias by fair and biased sentences. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4725–4736, 2024.