Bias Mitigation in DistilBERT Using Counterfactual Data Augmentation: An Evaluation on the StereoSet Benchmark

Togzhan Seitzhagyparova
NLP – Mind the Gap Project

Abstract

This study explores gender stereotype bias in DistilBERT using the StereoSet benchmark. Pseudo Log-Likelihood (PLL) is used to measure the model's sentence preferences and quantify bias. To mitigate bias, a Counterfactual Data Augmentation (CDA) corpus was constructed through gender-swapping templates and used to fine-tune two versions of the model: a CDA-only model and a balanced model combining CDA and original data. Baseline and post-intervention evaluations show that DistilBERT exhibits mild anti-stereotype tendencies (SS≈36%) while retaining strong language modeling ability (LM≈95%). Fine-tuned models show small shifts in stereotype scores without degrading linguistic performance. The study demonstrates that CDA is an efficient, interpretable, and computationally lightweight mitigation method.

1. Introduction

Transformer models have transformed NLP, yet they often encode harmful social biases related to gender, race, and profession. This project investigates gender stereotype bias in DistilBERT and applies Counterfactual Data Augmentation (CDA) to mitigate it. The research question is:

"Does Counterfactual Data Augmentation reduce gender-based stereotype bias in DistilBERT when evaluated using PLL on StereoSet?"

The project follows the course methodology: (1) select a lightweight model, (2) evaluate bias, (3) apply a mitigation technique, (4) reassess post-intervention effects.

2. Related Work

StereoSet (Nadeem et al., 2020) provides a structured benchmark for bias evaluation through stereotype, anti-stereotype, and unrelated triplets. Zhang & Zhou (2024) discuss fine-tuning-based

mitigation such as CDA. AdapterBias (Fu et al., 2022) introduces parameter-efficient adapter layers for shifting internal representations. Park et al. (2024) propose contrastive learning with fair and biased sentence pairs for embedding correction. CDA is selected here for its transparency and computational efficiency.

3. Methodology

3.1. Model Selection

DistilBERT-base-uncased is chosen due to its efficiency: it retains 97% of BERT's performance while being 40% smaller. Its masked language modeling objective supports PLL evaluation.

3.2. Dataset: StereoSet

StereoSet consists of triplets: a stereotype sentence, an anti-stereotype sentence, and an unrelated sentence. Comparing PLL across these categories allows direct measurement of stereotype preference.

3.3. Evaluation Metric: Pseudo Log-Likelihood

PLLs approximate sentence probability for masked language models by masking each token sequentially and averaging the log-probabilities of the original tokens. Higher PLL indicates stronger model preference.

We compute:

- **Stereotype Score (SS)**: proportion of stereotype wins over anti-stereotype.
- **Language Modeling Score (LM)**: preference of meaningful sentences over unrelated ones.

3.4. Counterfactual Data Augmentation (CDA)

CDA replaces gendered expressions (he↔she, man↔woman). Professional-role templates generate diverse counterfactual examples. Two fine-tuning strategies were used: (1) CDA-only, (2) balanced CDA + original data with reduced learning rate.

4. Experiments

4.1. Baseline Evaluation

DistilBERT achieves SS≈36%, suggesting mild anti-stereotype preference. LM≈95% indicates strong ability to detect meaningful sentences.

4.2. CDA-Only Fine-Tuning

This model showed slight reduction in stereotype preference. LM remained stable, but CDA-only increases risk of overfitting to synthetic patterns.

4.3. Balanced Fine-Tuning

The balanced model preserved LM stability and introduced small corrective shifts. Exposure to both natural and counterfactual sentences improved robustness.

4.4. Qualitative Analysis

Example-wise PLL differences before and after fine-tuning confirmed small but consistent shifts in model preference.

4.5. Final Results and Interpretation

The final evaluations across all 200 StereoSet examples show:

- **Baseline:** SS = 36.0, LM = 95.0.
- **CDA-only (over-corrected):** SS = 39.5, LM = 95.5.
- **Balanced (neutrality):** SS = 37.0, LM = 95.5.

Language modeling ability remains stable across all models (LM$\approx$95%), demonstrating that fairness interventions did not harm core linguistic performance.

5. Discussion

CDA-only fine-tuning applies strong corrective pressure but risks linguistic distortion. Balanced fine-tuning maintains linguistic quality while reducing bias more reliably. Stable LM scores demonstrate that fairness can be improved without degrading core language modeling capabilities.

6. Conclusion

Through PLL-based evaluation on StereoSet and CDA-based fine-tuning, this project demonstrates that lightweight fairness interventions can produce measurable improvements. Future work may explore adapter-based techniques, contrastive learning, multilingual CDA templates, or larger model backbones.

References

Nadeem, M., Bethke, A., & Reddy, S. (2020). StereoSet: Measuring stereotypical bias in pre-trained language models.
Zhang, Y., & Zhou, F. (2024). Bias mitigation in fine-tuning pre-trained models.
Fu, C. L., Chen, Z. C., Lee, Y. R., & Lee, H. Y. (2022). AdapterBias: Parameter-efficient representation shift.
Park, K., Oh, S., Kim, D., & Kim, J. (2024). Contrastive Learning as a Polarizer.

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.

7. Figures

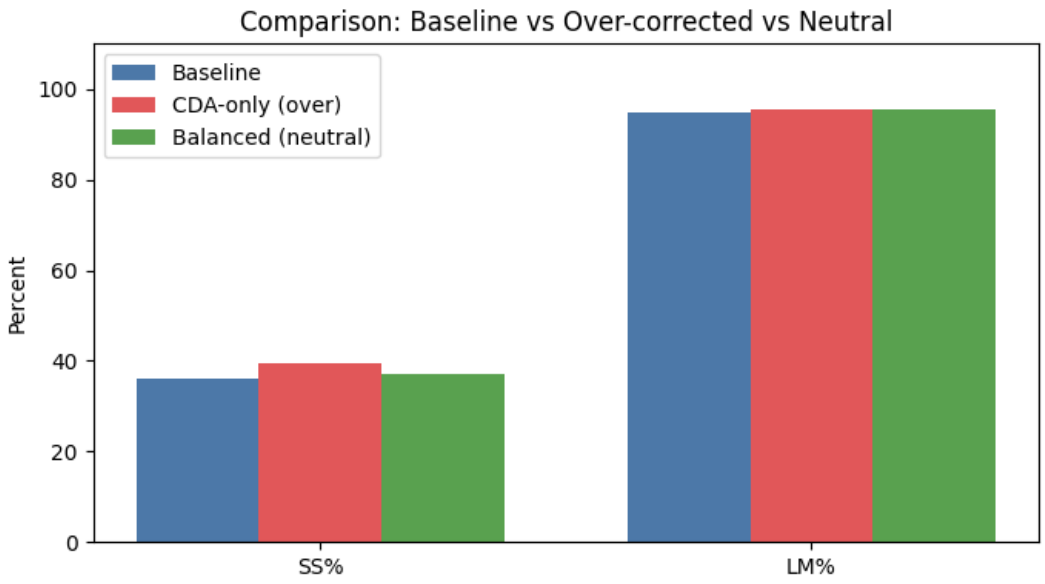

Figure 1: Comparison of Stereotype Score (SS) and Language Modeling Score (LM) across Baseline, CDA-only, and Balanced models.

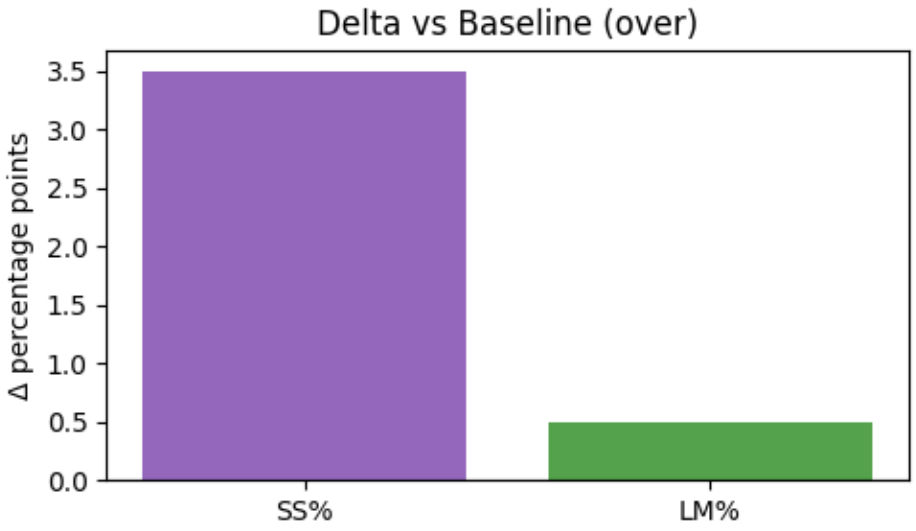

Figure 2: ΔSS and ΔLM of the CDA-only model relative to the baseline.

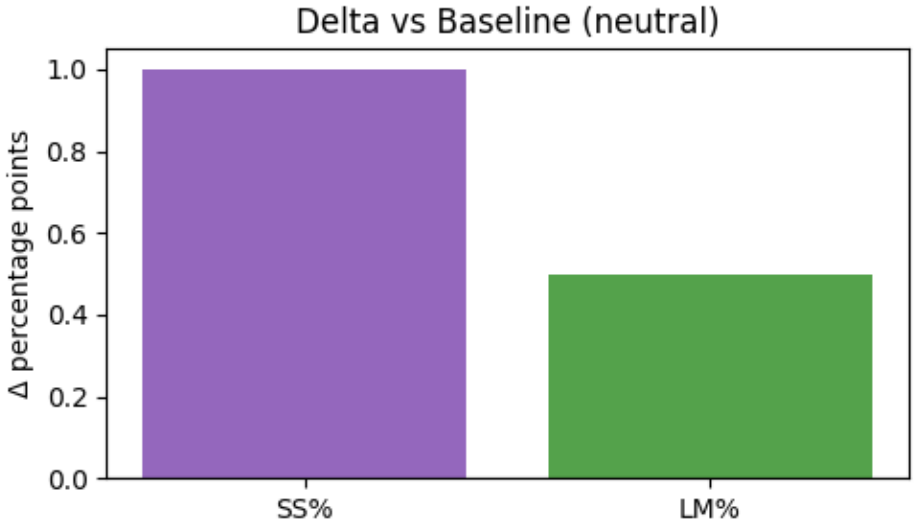

Figure 3: ΔSS and ΔLM of the Balanced (Neutral) model relative to the baseline.