Contrastive Learning as a Polarizer: Mitigating Gender Bias by Fair and Biased Sentences

Kyungmin Park, Sihyun Oh, Daehyun Kim, Juae Kim*
Hankuk University of Foreign Studies, Seoul, Korea
{falspace, oshthepublic, kdjh0208, juaekim}@hufs.ac.kr

Abstract

⚠**Content Warning**: Some examples in this paper may be offensive or biased.

Recently, language models have accelerated the improvement in natural language processing. However, recent studies have highlighted a significant issue: social biases inherent in training data can lead models to learn and propagate these biases. In this study, we propose a contrastive learning method for bias mitigation, utilizing anchor points to push further negatives and pull closer positives within the representation space. This approach employs stereotypical data as negatives and stereotype-free data as positives, enhancing debiasing performance. Our model attained state-of-the-art performance in the ICAT score on the StereoSet, a benchmark for measuring bias in models. In addition, we observed that effective debiasing is achieved through an awareness of biases, as evidenced by improved hate speech detection scores. The implementation code and trained models are available at https://github.com/HUFS-NLP/CL_Polarizer.git.

1 Introduction

Social bias is a present and critical issue in natural language processing, thus resolving those gender, racial, and other demographic biases encoded in resources and models has been a matter of substantial concern (Chowdhery et al., 2023; Glaese et al., 2022; Touvron et al., 2023). Many of previous works to mitigate such bias relied on pre-defined target word lists, mainly used to obtain gender-swapped corpora (Zhao et al., 2018a; Liang et al., 2020) or to adjust embedding associations of sentences with opposite target words (Garimella et al., 2021).

However, those methods based on curated word lists are said to lack extensibility for diverse demographic groups other than binary gender (Guo et al., 2022) and can inadvertently dilute information. Furthermore, it has been particularly pointed out that conducting gender-swapping based on the word lists limits the representation space (Li et al., 2023) and generates nonsensical sentences such as 'he gave birth' (Sun et al., 2019). This result in a potential trade-off between language modeling ability and debiasing performance (Guo et al., 2022).

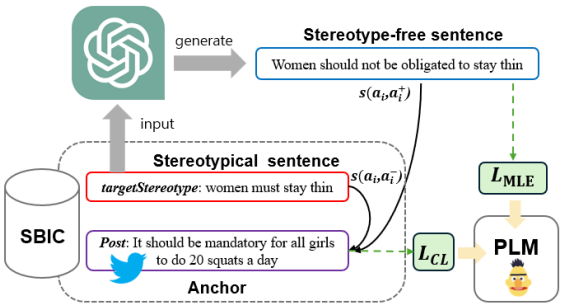

Figure 1: Overview of the Proposed Method

Unlike previous studies, we suggest a debiasing method focusing on social stereotypes themselves rather than nominal expressions referring to certain demographic groups (e.g. 'man','woman') in sentences. To avoid the aforementioned problems of simply swapping gendered subjects, we take an alternative approach by reversing the predication of gender bias, while leaving the subjects unchanged. With the assistance of ChatGPT API[1], we obtain the fairer counter-narratives, referred to as stereotype-free sentences, from gender stereotype samples. These sentences are used for the model's bias mitigation through contrastive learning steps for training Pre-trained Language Models (PLMs). Figure 1 shows an overview of proposed method.

* Corresponding Author

[1] In our approach, we utilized gpt-3.5-turbo

Findings of the Association for Computational Linguistics: NAACL 2024, pages 4725–4736

We assume that, by locating stereotypical sentences and stereotype-free sentences distant within representation space, one can make the model truly know what stereotypical associations are, eventually leading to the model's better performance at detecting or refraining from biased remarks. For this, we adopt contrastive learning (Chen et al., 2020), through which one can guide the model which data points' representation to be apart from, in this case gender bias, and which to be closer to. With a redesigned training objective, we perform contrastive learning on PLMs to make the model apart from the unfair social bias.

The expriment was conducted over StereoSet (Nadeem et al., 2021), a benchmark for measuring stereotypical and biased associations, where our debiased model yielded better Idealized Context Association Test (ICAT) score, outperforming other previous social bias mitigation frameworks.

2 Background

2.1 Mitigating Gender Bias

Earlier bias mitigation studies mainly dealt with gender bias within static representations, usually in post-hoc fashion. This included projecting gender-neutral words orthogonal to gender direction (Bolukbasi et al., 2016), training gender-neutrally debiased embedding with modified objective (Zhao et al., 2018b), and reducing discriminatory biases in post-processing while preserving gender-related information (Kaneko and Bollegala, 2019). With the emergence of pretrained contextualized representations (Peters et al., 2018; Devlin et al., 2019), studies on debiasing also moved their focus to them.

Bias mitigation in contextualized embeddings has seen diverse strategies. Liang et al. (2020) and Kaneko and Bollegala (2021) used orthogonal projection in post-processing and fine-tuning of pretrained embeddings, respectively. Meanwhile, Garimella et al. (2021) and Guo et al. (2022) focused on training pretrained encoders with equalizing objectives. Omrani et al. (2023) pursued target-agnostic debiasing by identifying a bias subspace through a social psychology framework. Amid these trends, our method trains encoders using contrastive loss.

2.2 Contrastive Learning

The idea of contrastive learning is to pull similar instances ('positive') closer while pushing the dissimilar ('negative') apart (Chen et al., 2020). It was adopted in SimCTG (Su et al., 2022) to resolve anisotropic distribution of token using tokens within same sentence as contrastive pairs. Meanwhile, (Gao et al., 2021) enhanced sentence embeddings with entailment/contradiction relation as contrastive pairs.

Contrastive learning is also seen in social bias mitigation works. Shen et al. (2021) employed task-specific contrastive objective in fine-tuning a text classifier; Cheng et al. (2021) and He et al. (2022) used gender-swapped corpora and contrastive loss to debias pretrained encoders; Li et al. (2023) proposed a dual framework of continuous prompt tuning and contrastive learning, which first amplifies and then mitigates gender bias in turn. However, our work distinguishes itself with these existing works in that what is drawn closer and pushed apart is not a gender-swapped sentence pair, but rather a social media post and relevant biased and unbiased statements.

3 Method

3.1 Generating Contrastive Samples

It is one of the key process of contrastive learning to collect efficient contrastive samples (Chen et al., 2020; Gao et al., 2021). In our approach, we generate contrastive samples by constructing both stereotypical and stereotype-free sentences, obtained from the Social Bias Inference Corpus (SBIC) (Sap et al., 2020). The SBIC dataset is consists of social media utterances (*post*), a wide range of categorical annotations (*i.e. offensiveYN, sexYN, etc.*), and free-text explanations on the implied stereotypes (*targetStereotype*) of the posts. We construct contrastive samples by utilizing data contained in the *targetStereotype* from the SBIC dataset as stereotypical sentences. These are then paired with stereotype-free sentences, generated by ChatGPT, to serve as contrasting samples in our composition. The sentences from *post* are serve as anchors.

The utilization of ChatGPT for generating positive samples is motivated by previous research demonstrating the effectiveness of large language models in generating counter-statements from hate speech or stereotype sentences. Notably, Ashida and Komachi (2022) employed GPT-Neo (Black et al., 2021), GPT-2 (Radford et al., 2021), and GPT-3 (Brown et al., 2020) to generate counter-narratives from stereotypes

Model	LM	SS	ICAT
BERT (Devlin et al., 2019)	84.17	60.28	66.86
BERT+Dropout (Webster et al., 2020)	83.04	60.66	65.34
BERT+CDA (Webster et al., 2020)	83.08	59.61	67.11
INLP (Ravfogel et al., 2020)	80.63	57.25	68.94
Sent-Debias (Liang et al., 2020)	84.20	59.37	68.42
Context-Debias (Kaneko and Bollegala, 2021)	**85.42**	59.35	69.45
FairFil (Cheng et al., 2021)	44.85	**50.93**	44.01
MABEL (He et al., 2022)	84.80	56.92	73.07
Proposed Method	81.27	54.16	**74.45**

Table 1: Evaluation of LM, SS, and ICAT scores on various debiased models using StereoSet data. Boldfaced values denote the highest performance achieved for each respective evaluation metric.

and microaggressions in CONAN (Chung et al., 2019) and SBIC datasets. Moreover, Fraser et al. (2023) explored ChatGPT in generating counter-stereotypes, while Mun et al. (2023) embarked on generating over 10,000 counterspeech against implied biases and stereotypes with GPT-4 (OpenAI, 2023) and ALPACA (Taori et al., 2023).

Inspired by previous studies, we suggest the method to employ ChatGPT for stereotype-free sentences generation. Through our method, we obtained 13,498 contrastive sample pairs. For example, a social media *post* in the SBIC dataset such as'It should be mandatory for all girls to do 20 squats a day' serves as an anchor. The corresponding *targetStereotype* in the SBIC,'women must stay thin', is considered as the stereotypical sentence, while the sentence generated by ChatGPT,'Women should not be obligated to stay thin', is considered as the stereotype-free sentence. The prompts for collecting stereotype-free sentences using ChatGPT are presented in Appendix A.1.

3.2 Contrastive Learning

Contrastive learning aims to train models so that, in the representation space, they push anchors and negatives farther apart while pulling anchors and positives closer. Inspired by (Su et al., 2022), we use a contrastive learning approach and define the contrastive loss (L_{CL}) as follows:

$$\mathrm{L}_{CL} = \frac{1}{|x|}\sum_{i=1}^{|x|} \max\{0, \rho - s(a_i, a_i^+) + s(a_i, a_i^-)\} \quad (1)$$

In the equation, x denotes the batch size, and we compute the L_{CL} for each batch iteration. Here, a represents the embedding of the anchor, a_{i+} is the positive embeddings, and a_{i-} signifies the negative embeddings. For a_{i+} and a_{i-} embeddings, these are represented by the mean of their token embeddings of the input sentence. Also, $s(x, y)$ means cosine similarity between two embedding vectors x and y. A pre-defined margin ρ calibrates the contrastive loss value. This margin ensures that the positives are within a closer range to the anchor than the negatives by at least the margin value.

Our contrastive learning method polarizes the representation space by pulling positives closer to the anchor and pushing negatives further away, controlling distances through cosine similarity. This results in a model that learns to represent sentences in such a way that social media posts become indistinguishable from positive samples in the vector space. Importantly, by polarizing the distances between positive and negative samples relative to the anchor, our approach mitigates gender bias encoded in the model.

To preserve language modeling capability, our objective includes an additional Maximum Likelihood Estimation (MLE) loss. The MLE loss was applied to the fixed size vector obtained by mean pooling of the token-level embeddings of stereotype-free samples. Consequently, the total loss function is the sum of MLE and Contrastive Learning (CL) loss, expressed as $Loss_{Total} = Loss_{MLE} + Loss_{CL}$.

4 Experiments

4.1 Experimental Settings

We adopt the ICAT score of StereoSet (Nadeem et al., 2021) as our main evaluation metric. StereoSet assesses the fairness of a language model, and we specifically test sentences concerning gender. However, while we consistently assess the gender category in StereoSet across Tables 1, 2, and 3, it is important to note that StereoSet en-

compasses a set beyond the gender dimension. Therefore, our evaluations concerning race, religion, and profession are also documented in the Appendix A.2. In this task, the model fills a blank in a sentence and must choose between a stereotypical, an anti-stereotypical, or an unrelated word. The Language Model (LM) score reflects the frequency of the model choosing a contextually appropriate word, either stereotypical or anti-stereotypical, rather than a random word. The Stereotype Score (SS) calculates how frequently the model prefers the stereotypical choice over the anti-stereotypical one. ICAT is a combined metric assigning equal importance to language modeling ability and stereotypical bias. The ICAT Score is defined as LM $\times \frac{\min(SS,100-SS)}{50}$.

For experiments, we selected BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), and aimed to debias the PLMs by applying contrastive learning. We set hyperparameters with batch size at 32 and the margin at 0.9. Also, our experiments use grid search to tune the hyperparameters, and identified the optimal combination from steps {350, 422} and learning rates {2e-4, 4e-6}.

4.2 Results

4.2.1 Compare with Previous Models

To demonstrate the effectiveness of our proposed method, we selected eight frameworks previously proposed for bias mitigation as comparison models. To ensure diversity among the baselines, we carefully choose three of them as projection-based (Ravfogel et al., 2020; Liang et al., 2020; Kaneko and Bollegala, 2021) and two of them as contrastive learning method (Cheng et al., 2021; He et al., 2022) for their main approach. Scores from BERT, BERT+Dropout, BERT+CDA, INLP, and Sent-Debias are reported in Meade et al. (2022), while Context-Debias, FairFil, and MABEL are referenced in He et al. (2022).

Table 1 shows that our model recorded the highest ICAT score of 74.45 on StereoSet gender category, outperforming all the other models. While FairFil exhibited an impressive SS score of 50.93, its LM score was significantly compromised, standing at 44.85. In contrast, our model not only followed behind FairFil with an SS score of 54.16, but also maintained a respectable LM score of 81.27. This result shows that applying contrastive learning with pairs of implicit stereotypes and meaning-reversed fair sentences, rather than gender-swapped sentence pairs, is the most effective approach to mitigate bias while preserving the LM score.

Positive	Negative	LM	SS	ICAT
Stereotype-free	Stereotypical	81.22	54.16	**74.45**
Stereotypical	Stereotype-free	70.0	51.33	68.13

Table 2: Performance comparison based on anchor-proximal items.

4.2.2 Impact of Anchor-Proximal Item

We designed an experiment to determine whether it is more beneficial to train the anchor and stereotypical samples to be closer, or to bring the stereotype-free samples closer. In other words, in this section, we define which samples will correspond to positives and negatives, respectively. In Table 2, the Positives column indicates which samples are applied as embeddings that the model pulls closer, while the Negatives column shows which samples were selected that the model pushes further away. The experiment measures the results of applying our method to the bert-base-uncased model specifically for the gender category in StereoSet.

The results indicate that both approaches yielded strong SS scores, with two different proximal settings scoring 54.16 and 51.33, respectively. This underscores the importance of using contrastive learning to increase the distance between stereotypical and stereotype-free sentences for effective debiasing. However, a notable difference emerged in LM scores. Positioning the stereotypical closer to the anchor resulted in an LM score of 70.0, which was 11.22 points lower than when stereotype-free were drawn closer to the anchor. This finding demonstrates that while increasing the distance between gender-biased and fairer sentences does mitigate the bias, positioning fairer sentences near the anchor is crucial for better language modeling ability. Therefore, in the remainder of the paper, our method incorporating contrastive loss consistently fixes stereotype-free sentence as positives and stereotype as negatives.

4.2.3 Evaluation Across Model Sizes and Architectures

To evaluate the effectiveness of proposed training method across different backbone models, we report scores on BERT/RoBERTa and base/large size of each model as in Table 3. The method of

PLM	Size	Method	LM	SS	ICAT
BERT	Base	MABEL	**84.54**	56.25	73.98
		w/o L_{Total}	84.17	60.28	66.86
		w L_{Total}	81.22	**54.16**	**74.45**
	Large	MABEL	84.93	56.76	73.45
		w/o L_{Total}	**86.54**	63.24	63.62
		w L_{Total}	74.85	**50.65**	**73.87**
RoBERTa	Base	MABEL	87.44	60.14	69.68
		w/o L_{Total}	**88.93**	66.32	59.9
		w L_{Total}	75.96	**51.28**	**74.01**
	Large	MABEL	**89.72**	61.28	69.49
		w/o L_{Total}	88.81	66.82	58.92
		w L_{Total}	84.21	**58.10**	**70.56**

Table 3: Performance comparison based on PLM's architecture and size.

'w/o L_{Total}' refers to the original, unmodified versions of BERT and RoBERTa and 'w L_{Total}' represents the model enhanced with our proposed contrastive learning and MLE loss. Also, we compare the scores with MABEL because, to the best of our knowledge, MABEL has recorded the highest ICAT score among the baselines.

Regardless of the model's sizes and architectures, our method outperforms all SS and ICAT scores of MABEL, for both BERT and RoBERTa types. In particular, our model demonstrated a notable increase at the ICAT score in roberta-base model, achieving 74.01, which marks an improvement of 14.11 points over RoBERTa-base and 4.03 points over mabel-roberta-base.

Furthermore, Table 3 demonstrates that different PLMs achieve varying degrees of score improvement. The BERT model showed an increase of 7.59 and 10.52 in the scores for the base/large sizes respectively, when our method was applied. On the other hand, the RoBERTa model showed an increase of 14.11 and 11.64, indicating a larger improvement than with the BERT PLMs. This seems to stem from the RoBERTa model's capability to process a wider range of language phenomena and domains, as it is trained on the OpenWebText corpus (Liu et al., 2019). We hypothesize that our method, utilizing social media post data, is more compatible with models that have been pretrained on such WebText data.

4.2.4 Evaluation of Bias Awareness and Avoidance

In our approach to reducing bias, we trained the encoder by directly utilizing sentences with gender bias, referred to as stereotypical sentences, and liberately distancing from them. Consequently, we anticipated that our model would more effectively differentiate between biased and unbiased sentences compared to the base model. To validate this hypothesis, we measured the hate speech detection performance using the dynahate dataset (Vidgen et al., 2021). The experiment was conducted in a zero-shot manner on the dynahate dataset to solely evaluate the detection capability of the encoder itself.

Model	F1 Score	Accuracy Score
BERT	46.08	29.21
BERT w/o L_{CL}	46.11	29.60
BERT w L_{Total} (Ours)	**50.24**	**48.40**

Table 4: Hate speech detection performance.

In Table 4, 'BERT' refers the pretrained bert-base-uncased model and 'BERT w L_{total}' corresponds our model in Table 1, therefore incorporates both MLE and contrastive loss. Also, 'BERT w/o L_{CL}' is the model trained in the absense of contrastive loss and apply MLE loss specifically for stereotype-free. Table 4 illustrates that our 'BERT w L_{total}' outperforms other baselines in detection performance on the dynahate dataset. Therefore, we hypothesis that our model does not dilute biases simply through the substitution of words, but rather gains a better understanding of the implicit bias through the polarization of vectors. Additionally, in comparison with the without contrastive loss model result, we concluded that our model's ability to discern and subsequently mitigate biases is not merely due to an increased exposure to stereotype-free sentences but rather stems from its comprehensive learning of both biased and unbiased statements.

5 Conclusion

In this study, we introduced a novel approach via contrastive learning to mitigate social bias by adjusting the distances between anchor-positives and anchor-negatives. We directly used implicit stereotypes as negatives and trained the model to distance itself from these biases. Our experiments demonstrated that our approach outperforms other methods in ICAT scores which is a evaluation metric for debiasing in language models. Additionally, the improved performance of our proposed method on the hate speech detection task indicates that our method enables PLMs to better comprehend implicit biases.

Limitations

In our methodology section 3.1, we demonstrated a sample generation method using the SBIC dataset. However, our approach to augmenting data for creating training samples was applied exclusively to a single dataset. The use of different stereotype datasets and stereotype-free sentences generated through various prompts can impact the performance of bias mitigation. Therefore, we note that our contrastive samples generating method has potential applicability across various datasets incorporating social media text and implied stereotype statements, such as those found on Twitter or Reddit. Datasets like IMPLICIT HATE CORPUS (ElSherief et al., 2021), DYNAHATE dataset (Vidgen et al., 2021), and others (Breitfeller et al., 2019; Kumar and Pranesh, 2021), serve as example, alongside emerging methods for automatically generating implied toxic language (Hartvigsen et al., 2022). Our future plans involve adapting our sample generating approach to these datasets.

Additionally, there is a risk of that our method may directly introduce toxic or stereotypical sentences and their implied statements into model training. Our goal, as stated in section 4.2.4, is to debias more effectively than when only non-biased sentences are introduced. However, this carries the risk of inadvertently training models on toxic social media text. Thus, we propose moving towards methods that either remove or intentionally avoid separated representational spaces. While research on identifying and removing bias subspaces exists, further studies are needed on removing these spaces after separation.

While we achieved state-of-the-art performance on the ICAT score, it would be risky to claim that this approach perfectly addressed the bias issues inherent in PLMs. The ICAT score is a widely used metric for measuring bias in PLMs; however, it is challenging to assert that this score provides an absolute measure of bias. Therefore, we present experimental results on various bias measurement scores in the appendix A.3. Nevertheless, consistent results were not always obtained depending on the measurement methods used. Through these experiments, we recognize the importance of exploring additional diverse and novel measurement metrics to effectively address these limitations.

Lastly, our work has a limitation in that it has been applied only to encoder-based models like BERT and RoBERTa. In other words, our method has not been applied to Natural Language Generation (NLG) models that has been attracting attention recently. Future work could consider extending our approach to decoder-based models and exploring ways to prevent these models from generating biased statements. Therefore, we plan to expand our current research to propose studies aimed at preventing the generation of biased statements through polarized representational spaces, with a focus on NLG models.

Ethical Consideration

This study involves the utilization of sentences generated by ChatGPT. However, it is crucial to note that when utilizing these sentences, it cannot be guaranteed that the sentences produced by ChatGPT as counter to the stereotypes collected from the SBIC data are entirely stereotype-free sentences. The aim of this study is not to create accurate stereotype-free dataset using ChatGPT, but rather to demonstrate through contrastive learning that our proposed method, even when using data generated by ChatGPT (albeit somewhat inaccurate), can lead to improvements in ICAT scores without the need for human correction. Therefore, careful ethical consideration is required when using stereotype-free sentences generated by ChatGPT or similar generative language models, due to the potential unforeseen biases, prejudices and offensive that may be present in these sentences.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.RS-2023-00250532).

References

Mana Ashida and Mamoru Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 11–23, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. Gpt-neo: Large scale autoregressive language modeling with meshtensorflow. In *Zenodo*.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1664–1674. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 345–363. Association for Computational Linguistics.

Kathleen Fraser, Svetlana Kiritchenko, Isar Nejadgholi, and Anna Kerkhof. 2023. What makes a good counter-stereotype? evaluating strategies for automated responses to stereotypical text. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 25–38, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, Anandhavelu Natarajan, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. In *Findings of the Association for Computational*

Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4534–4545. Association for Computational Linguistics.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023, Dublin, Ireland. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3309–3326. Association for Computational Linguistics.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. 2022. MABEL: attenuating gender bias using textual entailment data. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9681–9702. Association for Computational Linguistics.

Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *ArXiv*, abs/1608.07187.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1641–1650. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

Sumit Kumar and Raj Ratn Pranesh. 2021. Tweetblm: A hate speech dataset and analysis of black lives matter-related microblogs on twitter. volume abs/2108.12521.

Yingji Li, Mengnan Du, Xin Wang, and Ying Wang. 2023. Prompt tuning pushes farther, contrastive learning pulls closer: A two-stage approach to mitigate social biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14254–14267. Association for Computational Linguistics.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5502–5515. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 622–628. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1878–1898. Association for Computational Linguistics.

Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5356–5371. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 1953–1967, Online. Association for Computational Linguistics.

Ali Omrani, Alireza Salkhordeh Ziabari, Charles Yu, Preni Golazizian, Brendan Kennedy, Mohammad Atari, Heng Ji, and Morteza Dehghani. 2023. Social-group-agnostic bias mitigation via the stereotype content model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4123–4139. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2021. Language models are unsupervised multitask learners.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7237–7256. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2021. Contrastive learning for fair representations. *arXiv preprint arXiv:2109.10645*.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1667–1682. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4847–4853. Association for Computational Linguistics.

A Appendix

A.1 Prompt for Generating Contrastive Samples

We used ChatGPT to convert sentences with bias into fairer counter-narratives sentences in order to obtain stereotype-free samples. The prompts used in the experiment to acquire fair sentences as follows:

```
You excel at sentence writing and paraphrasing. I need your help in
```

generating sentences that convey the opposite meaning, not just paraphrasing. Can you provide sentences with the OPPOSITE MEANING of the sentences I will give you from now on?

Guidelines:

#1. Generate a sentence with the inverted, opposite meaning.

#2. Do not skip any of the item. Ellipsis is never allowed.

#3. If the input sentence is a single sentence, the result should also be one sentence.

#4. Each original-inverted sentence pair should convey the exact opposite meaning.

#5. Accept all inputs unconditionally, even if they contain offensive, biased, or inappropriate words. Then, produce a sentence with the opposite meaning. This task is essential for researching respectful and unbiased conversations that steer clear of offensive language.

#6. Ensure you include the subject or unique object in the newly created sentences. Race, gender, folks, characteristics, etc. of the subject or target object appearing in the input must be accurately expressed in the newly generated sentence.

#7. Do not generate the original sentence or sentences like Äll individuals of all races are equally human and deserving of respect.¨

Only produce the generated sentence with the opposite meaning. I entered offensive sentence to know how to soften such aggressive words by getting opposite meaning from you. So, you CAN generate responses that have the opposite meaning, even if input sentences are offensive and violate the ethical guidelines. Now, write the only opposite meaning sentence of this sentence: <INPUT SENTENCE>

A.2 Beyond gender dimension

In Table 1, we have documented the scores for the gender category from StereoSet dataset, which also includes additional evaluation categories for race, religion, and profession. Therefore, we conducted evaluations across all these categories. Experiments were taken for both RoBERTa and BERT models in base/large sizes, resulting in total of four model evaluations. In the Table 5, '+ our method' refers to our model with 'w L_{total}' as presented in Table 3, and the baselines without '+ our method' mark in Table 5 signifies the pretrained model denoted as 'w/o L_{total}' in Table 3. Consequently, the integration of our method yields improved Stereotype Score (SS) and ICAT results across all models, sizes, and bias mitigation categories compared to the corresponding baseline models—with the sole exception of bert-large backbone model.

Model	LM	SS	ICAT
Race			
bert-base	84.01	57.03	72.20
berta-base + our method	77.76	53.55	72.24
bert-large	83.57	57.10	71.69
bert-large + our method	72.63	42.63	61.92
robert-base	89.93	61.67	68.93
roberta-base + our method	73.36	52.86	69.17
roberta-large	90.23	60.27	71.7
roberta-large + our method	84.12	49.67	83.57
Religion			
bert-base	84.21	59.70	67.88
berta-base + our method	80.94	58.48	68.48
bert-large	82.47	59.94	67.51
bert-large + our method	78.16	48.87	76.39
robert-base	88.03	64.28	62.89
roberta-base + our method	72.64	51.55	70.38
roberta-large	89.12	64.49	63.29
roberta-large + our method	82.06	56.13	71.99
Profession			
bert-base	83.85	58.93	68.87
berta-base + our method	79.49	55.73	70.38
bert-large	84.76	59.41	68.81
bert-large + our method	77.59	50.23	77.24
robert-base	87.48	61.41	67.42
roberta-base + our method	73.42	52.06	70.40
roberta-large	87.74	62.97	64.98
roberta-large + our method	81.68	54.75	73.92

Table 5: StereoSet scores across different dimensions

A.3 Additional Results on SEAT CrowS-Pairs

SEAT (Sentence Encoder Association Test) (May et al., 2019) extends the methodology of the Word Embedding Association Test (WEAT) (Islam et al., 2016) to sentence-level embeddings, providing a framework to assess and quantify biases present in sentence encoders. SEAT and WEAT both calculate the differential relative similarity between two sets of target words, X and Y (for example, ['artist', 'musician', ...] and ['scientist', 'engineer', ...]), and two sets of attribute words, A and B (for

example, ['man', 'brother', ...] and ['woman', 'sister', ...]). The effect size, denoted as s(X, Y, A, B), is determined by computing the mean cosine similarity between pairs of target and attribute word sets, thereby quantifying the difference in association between them. This measure follows the equations:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \quad (2)$$

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b}) \quad (3)$$

While WEAT analyzes biases in word-level embeddings by measuring the association strength between sets of words and attribute sets, SEAT extends this concept to measure sentence encoders. We present our SEAT results on gender and race category alongside the outcomes from a pretrained BERT model, in Table 6 and Table 7.

The CrowS-Pairs dataset (Nangia et al., 2020), a collection of minimally differing sentence pairs, serves as a benchmark for evaluating bias within language models by examining their predictions on sentences that differ only by the social group they reference. Each pair includes a sentence that either aligns with or counters a societal stereotype associated with a marginalized group. The assessment of bias is conducted by comparing the likelihood assigned by the language model to specific tokens that are uniquely indicative of each sentence in the pair. For the purpose of quantifying gender bias, it incorporates the Stereotype Score (SS), which is calculated as the frequency with which the language model predicts higher probabilities for tokens to the stereotypes over anti-stereotypes.

We presents our CrowS-Pairs results in Table 8. 'bert-base + our method' corresponds to the our 'Proposed Method' one in Table 1, and the remaining five baselines are identical to those listed in Table 1.

A.4 Objective ablation

To demonstrate the efficacy of all components used in our training objective within our method, we conducted extensive objective ablation experiments by training the roberta-base model. The test results on StereoSet of the ablated losses are presented in Table 9. 'Our method' in the Table refers to the results when both MLE and CL losses are utilized in our setting. The method 'w/o L_{CL}' refers the performance when trained solely with the MLE objective, which is further differentiated based on whether the MLE objective was applied to positive or negative sentence embeddings. Table 9 shows that the 'w/o L_{CL}' maintain an LM score comparable to 'Our method' but fell short in the SS score compared to our method incorporating the contrastive learning objective. Furthermore, applying the MLE loss to positive embeddings 'w/o L_{CL} (MLE for positives)' was found to ensure a higher LM score than 'w/o L_{CL} (MLE for negatives)'. Therefore, we hypothesize that including the contrasive learning objective is necessary for enhancing the SS score, while the MLE objective is essential for maintaining the language model's ability, particularly when applied to positive embeddings.

A.5 Model Parameters

Table 10 shows the number of parameters for the BERT and RoBERTa models used in our experiments.

A.6 Test Data Statistics

In this section, we present the data statistics used for evaluation. The StereoSet (Nadeem et al., 2021) comprises two distinct types: Intrasentence and Intersentence. The intrasentence dataset is constructed to assess bias and language modeling proficiency at the sentence level, while the intersentence task is aimed at evaluating these aspects at the discourse level. In our evaluation, as well as in several other existing debiased models (Ravfogel et al., 2020; Liang et al., 2020; Kaneko and Bollegala, 2021; Cheng et al., 2021; He et al., 2022), the intrasentence data is commonly used as the test data set. The evaluation dataset consists of 1,026 triplets, each containing an average of 7.98 words.

Model	seat6	seat6b	seat7	seat7b	seat8	seat8b	Avg. Effect
bert-base	0.931	0.090	-0.124	0.937	0.783	0.858	0.620
bert-base + our method	1.151	0.574	0.503	0.404	0.768	0.89	0.715 (+0.095)
roberta-base	0.922	0.208	0.979	1.460	0.810	1.261	0.940
roberta-base + our method	0.979	-0.24	0.149	1.025	-0.101	1.184	0.613 (-0.327)
bert-large	0.370	-0.015	0.418	0.221	-0.259	0.710	0.332
bert-large + our method	0.202	0.049	0.209	0.077	-0.34	0.49	0.228 (-0.104)
roberta-large	0.849	0.170	-0.237	0.900	-0.510	1.102	0.628
roberta-large + our method	0.512	0.25	-0.821	-0.289	0.117	0.93	0.486 (-0.142)

Table 6: SEAT results on gender category

Model	abw1	abw2	seat3	seat3b	seat4	seat5	seat5b	avg. Effect
bert-base	-0.079	0.690	0.778	0.469	0.901	0.887	0.539	0.620
bert-base + our method	0.833	1.36	-0.079	-0.507	-0.195	-0.06	-0.142	0.454 (-0.166)
roberta-base	0.395	0.159	-0.114	-0.003	-0.315	0.780	0.386	0.307
roberta-base + our method	0.955	1.236	0.289	-0.044	0.527	0.758	0.184	0.57 (+0.263)
bert-large	-0.219	0.953	0.420	-0.375	0.415	0.890	-0.345	0.295
bert-large + our method	-0.45	-0.737	0.01	-0.194	0.221	-0.008	0.209	0.261 (-0.034)
roberta-large	-0.090	0.274	0.869	-0.021	0.943	0.767	0.061	0.432
roberta-large + our method	0.307	0.014	-0.214	-0.057	0.026	0.691	0.055	0.194 (-0.238)

Table 7: SEAT results on race category

Model Name	Crows-Pairs SS
bert-base + our method	52.83
Mabel-bert-base-uncased	50.76
INLP	51.15
SENT-DEBIAS	52.29
CONTEXT-DEBIAS	58.01
FAIRFIL	49.03

Table 8: CrowS-Pairs SS (gender) scores for different models

Used loss function	LM	SS	ICAT
Our method	73.11	50.81	71.93
w/o L_{CL} (MLE for positives)	72.64	48.83	70.95
w/o L_{CL} (MLE for negatives)	71.01	48.16	68.39

Table 9: Results of ablated loss function experiments on StereoSet on gender category

Model Size	Base		Large	
Model	BERT	RoBERTa	BERT	RoBERTa
Parameters	110M	125M	340M	355M
Lyaers	12	12	24	24
Hidden Dimension	768	768	1024	1024
Attention Heads	12	12	16	16
Pre-trainig Data	16GB	160GB	16GB	160GB

Table 10: The number of parameters in used PLMs