Content Conditional Debiasing for Fair Text Embedding

Anonymous ACL submission

Abstract

Mitigating biases in machine learning models has gained increasing attention in Natural Language Processing (NLP). Yet, only a few studies focus on fair text embeddings, which are crucial yet challenging for real-world applications. In this paper, we propose a novel method for learning fair text embeddings. We achieve fairness while maintaining utility trade-off by ensuring conditional independence between sensitive attributes and text embeddings conditioned on the content. Specifically, we enforce that embeddings of texts with different sensitive attributes but identical content maintain the same distance toward the embedding of their corresponding neutral text. Furthermore, we address the issue of lacking proper training data by using Large Language Models (LLMs) to augment texts into different sensitive groups. Our extensive evaluations demonstrate that our approach effectively improves fairness while preserving the utility of embeddings, representing a pioneering effort in achieving conditional independence for fair text embeddings.

1 Introduction

Embedding text into dense representations is popular in modern Natural Language Processing (NLP), enables applications like sentiment analysis (Dang et al., 2020), recommendation systems (Zhang et al., 2016), and search engines (Palangi et al., 2016). However, as these embeddings become widespread, their inherent biases could affect various applications (Packer et al., 2018; Baeza-Yates, 2018; Zerveas et al., 2022; Rabelo et al., 2022), potentially harming marginalized groups by perpetuating gender, race, and other biases. Unfortunately, various forms of biases, including gender, racial, and religious biases, have been identified in text embeddings generated by pre-trained language models (PLMs), as reported in several studies (Bolukbasi et al., 2016; Nissim et al., 2020; Liang et al., 2020; May et al., 2019). Consequently, attaining fairness in text embedding models is crucial.

Recent debiasing techniques (Liang et al., 2020; Kaneko and Bollegala, 2021) for text embeddings use post-training to address biases, avoiding the inefficiency of retraining sentence encoders for each new bias. When removing bias, projection-based methods (Liang et al., 2020; Kaneko and Bollegala, 2021) reduce an embedding's projection onto each bias subspace. The distance-based method (Yang et al., 2023) constructs embeddings for sensitive groups and equalizes distances to text embeddings across these groups. Nevertheless, these methods persist in pursuing independence between sensitive attributes and text embeddings completely remove the sensitive information. As a result, these approaches do not effectively find the *sweet spot* between fairness and utility trade-off (Zhao and Gordon, 2022; Deng et al., 2023; Zliobaite, 2015).

Recent studies (Mary et al., 2019; Deng et al., 2023; Pogodin et al., 2022) suggest that using datasets labeled with sensitive information to achieve conditional independence — specifically, conditioning on the content class to preserve semantic information within the text — provides a more effective approach to achieving fairness while preserving utility. Yet, the scarcity of text datasets with sensitive labels (Gallegos et al., 2023) limits the practical application of these findings. Although Counterfactual Data Augmentation (CDA) (Zhao et al., 2018) and LLMs (Schick and Schütze, 2021; Shao et al., 2023) provide ways for generating data, they face challenges with grammatically incorrect artificial instances (Dinan et al., 2019) or inherent systematic biases (Yu et al., 2023), respectively.

In this paper, we propose a new method to improve text embedding fairness via Content Conditional sensitive attributes Debiasing (dubbed `CCD`). Specifically, we first use LLMs to augment a dataset into texts across various sensitive groups linked to the same content by substituting sensi-

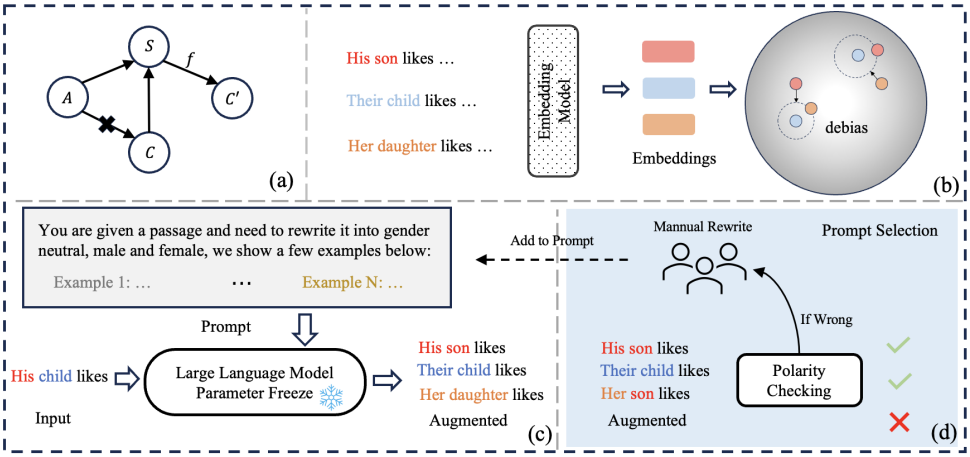

Figure 1: Pipleline of our method with **gender** as the sensitive attributes. (a) Graphical demonstration of the fairness issue. (b) The debiasing procedure achieves a content-conditioned equal distance to improve the fairness. (c) Overview of the data augmentation strategy, including the prompt template used to replace sensitive words with their equivalents from all sensitive groups. (d) Prompt search module: Augmented texts are sent to the demographic polarity checking block. Incorrectly augmented samples are then manually labeled and added to the prompts.

tive words with equivalents from different groups (*i.e.*, "son", "daughter" and "child" for gender bias). Then, we achieve conditional independence between sensitive attributes and text embeddings by ensuring that texts with varied sensitive attributes but identical content have embeddings equidistant from their neutral counterparts. Extensive evaluations on debiasing benchmarks and downstream tasks demonstrate our method's effectiveness in promoting fairness while preserving utility.

2 Method

2.1 Problem Setting

This section outlines the problem of fairness in text embeddings. We define key variables: S for input text, C for the content, and f for the text embedding model that generates text embedding $C' = f(S)$. We also denote sensitive attributes as $A = [a_1, ..., a_{|A|}]$ (e.g., gender, race, age), words related to attribute a_i as X^{a_i}, and neutral words as X^n. A text is neutral (S^n) if it lacks sensitive words, and is associated with attribute a_i (S^{a_i}) based on its sensitive polarity (Wang et al., 2023). We list all notations in Appendix Table 3.

Fairness Issue: Fig. 1 (a) shows there exists an association between attributes A and content variable C, if model f superficially considers A as a proxy for C, leading to issues of fairness.

2.2 Content Conditional Fairness Loss

We propose to achieve conditional independence between sensitive attributes and content $A \perp C' \mid C$. The conditional independence allows prediction C' to depend on A but only through the content variable C, prohibiting abusing A as a *proxy* for C thus mitigating the fairness issue while preserving the utility. Echoing the methodologies in previous research (Hinton and Roweis, 2002; Yang et al., 2023), we can define the probability $P(i|j)$ as the likelihood of variable i selecting j as its neighbor in the embedding space. Then conditional independence $A \perp C' \mid C$ can be represented as a content conditioned equal distance on the embedding space, as shown in Fig. 1 (b) and definition 2.1 (detailed proof in Appendix B.5) :

Definition 2.1. *(Content Conditional Equal Distance Fairness.) Let S_C^n be a neutral text with the content C, assume $S_C^A = [S_C^{a_1}, S_C^{a_2}, ..., S_C^{a_{|A|}}]$ being a set of texts from all sensitive groups with the same content C. Then embedding model f is content conditioned equal distance fair with respect to attribute A, for any $a_i, a_j \in A$:*

$$||f(S_C^{a_i}) - f(S_C^n)|| = ||f(S_C^{a_j}) - f(S_C^n)||. \quad (1)$$

To align with this definition of fairness, we design a loss function L_{bias} that aims to mitigate biases while preserving the representation ability of PLMs. For a sample pair $[S_C^{a_1}, ..., S_C^{a_{|A|}}, S_C^n]$:

$$L_{bias} = \sum_{i \in [A]} \sum_{j \neq i} |dist(f(S_C^{a_i}), f(S_C^n)) - dist(f(S_C^{a_j}), f(S_C^n))|, \quad (2)$$

where $dist(A, B) = \exp\left(-\frac{\|A-B\|^2}{2\rho^2}\right)$ measures the distance on the manifold (Yang et al., 2023; Hinton and Roweis, 2002) (details in Appendix B.5),

and ρ is selected as the variance of the distance over the training dataset for normalization. To further preserve the valuable information encoded in the model for downstream tasks, we design L_{rep} to enforce similarity between the text embeddings processed by the fine-tuned model f and those processed by the original model f^{ori}:

$$L_{rep} = ||f(S^n) - f^{ori}(S^n)||. \quad (3)$$

Thus, the overall training objective is:

$$L_{all} = L_{bias} + \beta * L_{rep}. \quad (4)$$

2.3 Content Conditional Data Augmentation

To meet our goal in definition 2.1, we augment a dataset into texts across various sensitive attributes A linked to the same content variable C.

Augment Text into Different Sensitive Groups: We facilitate text augmentation by using a pre-trained LLM to replace sensitive words with equivalents from different sensitive groups. As shown in Fig. 1 (a), for sensitive words $X^A = [X^{a_i}, ..., X^{a_j}], a_i, a_j \in A$ in text S, the LLM h substitutes X^A with words from different sensitive groups $h(X^A) = [[X^{a_i}_{a_m}, ..., X^{a_j}_{a_m}]]_{a_m \in A}$ respectively, creating augmented texts S^A from all sensitive groups. We also generate a neutral version S^n using neutral replacements, resulting in a dataset of texts from different groups $\mathcal{S}^A$ and a neutral text S^n. Examples are provided in Appendix Table 4.

Rule Guided Prompt Searching: Considering the impact of few-shot prompt tuning on LLM effectiveness, choosing the right prompts is essential for successful text augmentation. As shown in Fig. 1 (d), we first find inaccurately augmented texts whose augmented text's sensitive group mismatches its sensitive polarity (Wang et al., 2023) computed by:

$$g(S) = \arg\max_{a_i \in A} occ(S, V^{a_i}), \quad (5)$$

where occ counts occurrences of words from the sensitive words lists $V^{a_i}, a_i \in A$ in the sentence S, we assign polarity based on the most frequently mentioned group in A. A text S^{a_i} is considered wrongly augmented if $g(S^{a_i}) \neq i$. We then manually augment such texts to create example prompts for better alignment and repeat the process. This process is iterated to select K (ours 10) samples as few-shot prompts.

3 Experiments

In this paper, we take gender bias as an example due to its broad impact on society.

Datasets We generate training data with the News-commentary-v15 corpus (Tiedemann, 2012). For gender bias evaluation, we follow (Yang et al., 2023) to use SEAT (May et al., 2019), CrowS-Pairs (Nangia et al., 2020) and StereoSet-Intrasentence data (Nadeem et al., 2020). We additionally assess fairness on longer texts via the Bias-IR dataset (Krieg et al., 2023). To evaluate whether the biased models' representation ability is maintained, we follow (Kaneko and Bollegala, 2021; Yang et al., 2023) to select four small-scale subsequent tasks from the GLEU benchmark: Stanford Sentiment Treebank (SST-2 (Socher et al., 2013)), Microsoft Research Paraphrase Corpus (MRPC (Dolan and Brockett, 2005)), Recognizing Textual Entailment (RTE (Bentivogli et al., 2009)) and Winograd Schema Challenge (WNLI (Levesque et al., 2012)).

Baseline Methods To assess debiasing performance, we compare our algorithm with finetuning-based methods DPCE (Kaneko and Bollegala, 2021) and ADEPT-F (Yang et al., 2023). To assess the effectiveness of our data augmentation strategy, we compare our approach with CDA (Zhao et al., 2018). For the use of LLMs in data augmentation, we use ChatGPT and Gemini (Team et al., 2023).

Hyperparameters We follow (Yang et al., 2023) to conduct experiments on the *BERT-LARGE-UNCASED* pre-trained model. We use Adam to optimize the objective function. During the debiasing process, our learning rate is 5e-5, batch size is 32, and β is 1. Results for DPCE and ADEPT-F are with originally reported hyperparameters (Kaneko and Bollegala, 2021; Yang et al., 2023). Experiments are conducted on a NVIDIA A100 GPU.

3.1 Results and Analysis

We evaluate four models on all benchmarks, namely the original model (pre-trained with no explicit debiasing), the DPCE model, the ADEPT-F model, and our `CCD`.

Reducing Gender Biases In Table 1, our experiments reveal that our `CCD` with all data strategies (CDA, GPT, and Gemini) excels in debiasing, consistently outperforming baselines in the StereoSet and the CrowS-Pairs dataset and ranking second best in SEAT. It is worth noting that our method achieves a good ICAT score in the Stere-

Table 1: Evaluation results on debiasing performance. We test the debiased models on SEAT, CrowS-Pairs, and filtered StereoSet-Intrasentence, with the best result in bold.

Datasets	SEAT (0.00) the best							StereoSet:gender			StereoSet:all			CrowS-Pairs
Method	6	6-b	7	7-b	8	8-b	AVG (abs)↓	LMS↑	SS (50.00)	ICAT↑	LMS↑	SS(50.00)	ICAT↑	SS(50.00)
BERT	0.37	0.20	0.42	0.22	−0.26	0.71	0.36	86.34	59.66	69.66	84.16	58.24	70.29	55.73
DPCE	−0.21	0.27	0.44	0.07	0.25	0.21	**0.24**	81.19	56.72	65.41	64.06	**52.96**	60.26	52.29
ADEPT-F	0.83	−0.14	0.63	1.24	0.43	1.28	0.76	86.45	61.70	66.21	85.09	57.52	72.26	51.91
CCD-CDA	0.16	0.03	0.43	0.38	0.47	0.22	0.29	80.34	**53.53**	74.69	79.10	53.46	73.62	46.95
CCD-GPT	0.47	−0.06	−0.33	0.55	−0.17	0.48	0.34	83.76	55.40	74.71	82.70	54.80	**74.75**	**49.62**
CCD-Gemini	0.47	−0.00	−0.02	−0.72	−0.30	0.07	0.26	82.91	54.93	**74.72**	82.97	55.00	74.67	48.85

Table 2: Evaluation results on the GLUE dataset and the Bias-IR dataset, we additionally calculate the average deviation to 0.5 for Bias-IR as AVG-DEV. The **bold** and underline represent the best and second-best respectively.

Datasets	GLUE ↑					Bias-IR (Male Ratio, 0.50 the best)								
Method	SST-2↑	MRPC↑	RTE↑	WNLI↑	AVG↑	Appearance	Child	Cognitive	Domestic	Career	Physical	Relationship	AVG	AVG-DEV↓
BERT	92.9	84.6	72.5	38.0	72.0	0.71	0.50	0.75	0.46	0.75	0.68	0.61	0.64	0.16
DPCE	92.8	69.6	53.4	49.3	66.3	0.86	0.79	1.00	0.47	0.70	0.84	0.61	0.75	0.24
ADEPT-F	93.2	85.5	69.9	56.3	76.2	0.50	0.50	0.75	0.53	0.80	0.68	0.65	0.63	0.13
CCD-CDA	92.8	**86.3**	65.3	50.7	73.8	0.79	0.79	0.83	0.80	0.70	0.79	0.83	0.79	0.29
CCD-GPT	**93.6**	85.1	70.4	56.3	76.4	0.50	0.43	0.58	0.60	0.35	0.53	0.61	**0.51**	**0.07**
CCD-Gemini	93.5	83.6	**72.9**	**56.3**	**76.6**	0.57	0.64	0.58	0.60	0.70	0.42	0.65	0.59	0.11

oSet dataset, indicating an excellent balance between performance and fairness. However, DPCE, while preserving great fairness, adversely affects its representation capability, as evidenced by a notably lower LMS score in the StereoSet dataset.

Preserving Representation Ability In Table 2, the GLUE results demonstrate that CCD-Gemini achieves the highest average performance, suggesting that it not only maintains but also enhances the model's representational capabilities. Conversely, DPCE, which is intended to separate gender attributes from neutral text embeddings, adversely affects the model's utility.

Bias in Information Retrival We conduct an evaluation using the Bias-IR dataset to assess the fairness of text embedding for long text sequences. Table 2 illustrates that our approach, CCD-Gemini and CCD-GPT outperforms others in fairness across numerous fields, as well as in average fairness.

Comparision of Data Strategy To demonstrate the effectiveness of our proposed data strategy, we conduct comparisons with CDA as shown in Table 1 and Table 2. Integrating our debiasing loss with both data strategies resulted in improved fairness. However, CDA experienced a drawback in maintaining model utility, due to potential grammatical inconsistencies introduced by its rule-based augmentation. This highlights the superiority of our LLM-based augmentation method. For different LLMs, we found that both ChatGPT and Gemini achieve good performance on fairness and utility.

Embedding Visualization Fig. 2 reveals that original BERT clusters male and female embeddings, indicating fairness issues. DPCE and ADEPT-F debias by separating sensitive attributes from neutral embeddings, harming the representation quality. Our CCD maintains the original embedding proximity and mixes male and female embeddings, achieving fairness effectively.

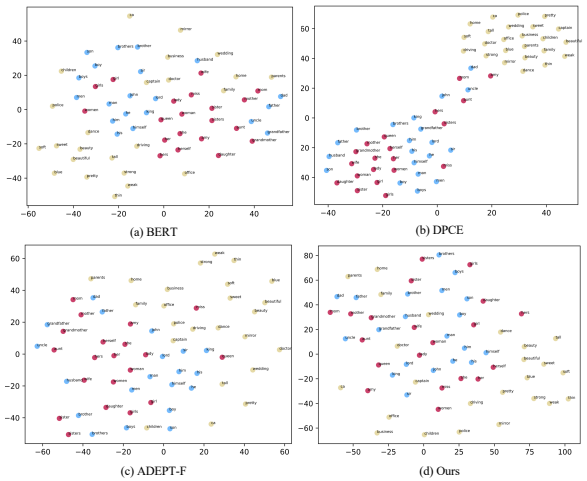

Figure 2: T-SNE plots of embeddings that are processed by different methods. Our approach maintains embedding positions similar to BERT while mixing male and female embeddings thus achieving fairness.

4 Conclusion

This paper presents a new method for generating fair text embeddings by ensuring conditional independence between sensitive attributes and text content. We show that conditional independence can be represented as content-conditioned equal distance in embedding spaces. By leveraging LLMs to augment datasets, we address the data scarcity issue and enable model debiasing. Extensive experiments in diverse datasets show that our approach enhances fairness while preserving utility.

5 Limitaions

In this study, we utilize gender bias to demonstrate the efficacy of our method. As our approach constitutes a general pipeline, we plan to extend our methodology to address other types of biases (e.g., race, age) in the future. Moreover, we discuss the application of our method in a binary gender setting, which generally does not reflect the real world where gender (and other biases) may not be strictly binary. Fortunately, our method is readily extensible to any number of dimensions. We consider this extension as part of our future work.

6 Ethical Consideration

Our work pioneers in mitigating biases in text embeddings, crucial for fairness and inclusivity in NLP applications. We introduce a method that ensures fair representation by achieving conditional independence between sensitive attributes and text embeddings, aiming to reduce societal biases. Employing LLMs for data augmentation represents ethical advancement in tackling inherent biases, moving towards equitable technology and inspiring future bias-aware research. Our contribution significantly advances AI fairness by validating a method that minimizes bias in text embeddings, promoting inclusivity in machine learning.

References

Ricardo Baeza-Yates. 2018. Bias on the web. *Communications of the ACM*, 61(6):54–61.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. *TAC*, 7(8):1.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2022. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*.

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. 2022. Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. *arXiv preprint arXiv:2203.09101*.

Nhan Cach Dang, María N Moreno-García, and Fernando De la Prieta. 2020. Sentiment analysis based on deep learning: A comparative study. *Electronics*, 9(3):483.

Wenlong Deng, Yuan Zhong, Qi Dou, and Xiaoxiao Li. 2023. On fairness of medical image classification with multiple sensitive attributes via learning orthogonal representations. In *International Conference on Information Processing in Medical Imaging*, pages 158–169. Springer.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2019. Queens are powerful too: Mitigating gender bias in dialogue generation. *arXiv preprint arXiv:1911.03842*.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Geoffrey E Hinton and Sam Roweis. 2002. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. *arXiv preprint arXiv:2101.09523*.

Klara Krieg, Emilia Parada-Cabaleiro, Gertraud Medicus, Oleg Lesota, Markus Schedl, and Navid Rekabsaz. 2023. Grep-biasir: A dataset for investigating gender representation bias in information retrieval results. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pages 444–448.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*.

Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui. 2019. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pages 4382–4391. PMLR.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

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

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Ben Packer, Yoni Halpern, Mario Guajardo-Cspedes, and Margaret Mitchell. 2018. Text embedding models contain bias. here's why that matters. *Google Developers*.

Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):694–707.

Roman Pogodin, Namrata Deka, Yazhe Li, Danica J Sutherland, Victor Veitch, and Arthur Gretton. 2022. Efficient conditionally invariant representation learning. *arXiv preprint arXiv:2212.08645*.

Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022. Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021. *The Review of Socionetwork Strategies*, 16(1):111–133.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. *arXiv preprint arXiv:2104.07540*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Synthetic prompting: Generating chain-of-thought demonstrations for large language models. *arXiv preprint arXiv:2302.00618*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Rui Wang, Pengyu Cheng, and Ricardo Henao. 2023. Toward fairness in text generation via mutual information minimization based on importance sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 4473–4485. PMLR.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 322–330.

Ke Yang, Charles Yu, Yi R Fung, Manling Li, and Heng Ji. 2023. Adept: A debiasing prompt framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10780–10788.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Zerogen: Efficient zero-shot learning via dataset generation. *arXiv preprint arXiv:2202.07922*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. *arXiv preprint arXiv:2306.15895*.

George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. 2022. Mitigating bias in search results through contextual document reranking and neutrality regularization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2532–2538.

Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 353–362.

Han Zhao and Geoffrey J Gordon. 2022. Inherent tradeoffs in learning fair representations. *The Journal of Machine Learning Research*, 23(1):2527–2552.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Indre Zliobaite. 2015. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*.

A Related Work

LLMs in creating dataset: Leveraging the success of large language models (LLMs), researchers have begun using them to generate various forms of training data, such as tabular data (Borisov et al., 2022), relation triplets (Chia et al., 2022), sentence pairs (Schick and Schütze, 2021), and instruction data (Shao et al., 2023),*et al.*. As we focus on obtaining data with sensitive attribute information, data generation for text classification would be the most similar one among those applications. Recent efforts in generating data for text classification (Meng et al., 2022; Ye et al., 2022; Wang et al., 2019) primarily employ simple class-conditional prompts while focusing on mitigating issues of low quality after generation. However, these efforts encounter the challenge of inherent systematic biases present in LLMs (Yu et al., 2023). Although recent study (Yu et al., 2023) considered generated data bias, it focuses only on the diversity of topics and overlooks the inherent bias within words in a text (*i.e.*, "child" occurs more frequently with "mother").

Debias Text Embedding: Bias in text embeddings (also known as sentence embedding) is a significant issue that arises when these models reflect or amplify societal stereotypes and prejudices found in their training data. To resolve the issue, (Liang et al., 2020) contextualizes predefined sets of bias attribute words to sentences and applies a hard-debias algorithm (Bolukbasi et al., 2016). Contextualized bebiasing methods (Kaneko and Bollegala, 2021; Yang et al., 2023) apply token-level debiasing for all tokens in a sentence and can be applied at token- or sentence-levels (Kaneko and Bollegala, 2021) to debias pretrained contextualized embeddings. However, all the above methods aim to strictly achieve independence between text embedding and sensitive attributes, which may not achieve a good balance between fairness and utility.

B Algorithm Details

B.1 Notation

Table 3: Main notations used in this paper.

		Basic Variables
L	$\triangleq$	loss function
f,f^{ori}	$\triangleq$	finetuned and original text embedding model.
h	$\triangleq$	Large language model.
θ_p	$\triangleq$	Few-shot prompts that used to empower a LLM.
A, a_i	$\triangleq$	Sensitive attribute set and i-th sensitive attribute.
S^{a_i},S^n	$\triangleq$	Text that relate to sensitive attribute a_i and neutral text.
C,C'	$\triangleq$	Content variable and predicted content.
X^{a_i}, X^n	$\triangleq$	words from group a_i and neutral words in a text.
V^{a_i}	$\triangleq$	words list that contains all collected words related to attribute a_i.

B.2 Fainess Goals

Improving Fairness There are two goals in improving the fairness of a text embedding model (Yang et al., 2023; Kaneko and Bollegala, 2021). (1) Remove the superficial connection between the content C and sensitive attributes A so that the fairness is improved (2) Maintain the utility of the text embedding model so that the performance is protected.

B.3 Data Augmentation Details

We generated training data using the News-Commentary-v15 corpus (Tiedemann, 2012) focusing on gender bias. By employing Gemini and ChatGPT for data augmentation, we obtained datasets comprising 43,221 and 42,930 sample pairs, respectively. Each pair contains texts with identical content from male, female, and neutral perspectives.

Table 4: We utilize LLM to augment text into three gender categories: Male, Female, and Neutral. Below are sample examples of the generated data, where words containing gender information are highlighted in colors: red for male, blue for neutral, and orange for female.

Gender	Generated Text
Male	But because Rumsfeld wanted to prove a point about transforming strategy. After championing the continuation of his hardline policy, his current strategy of negotiation is risky. He has been very vocal in voicing discontent with the rule of Kirchner and that of his husband and predecessor, Néstor Kirchner.
Neutral	But because the individual wanted to prove a point about transforming strategy. After championing the continuation of their hardline policy, the current strategy of negotiation is risky. They have been very vocal in voicing discontent with the rule of Kirchner and that of their spouse and predecessor, Néstor Kirchner.
Female	But because Rachel wanted to prove a point about transforming strategy. After championing the continuation of her hardline policy, her current strategy of negotiation is risky. She has been very vocal in voicing discontent with the rule of Kirchner and that of her wife and predecessor, Néstor Kirchner.

B.4 Data Augmentation Results

In this section, we show data samples augmented by our algorithm, as described in Section 2.3, in Table 4. Words associated with sensitive information in the text are replaced with corresponding words from all different sensitive groups and their neutral counterparts. Ultimately, we obtained a dataset with texts of content C from all groups A and neutral.

B.5 Ommited Proofs

In this section, we give a detailed proof of definition 2.1.

Proof. Firstly, we write the conditional independence for any $a_i, a_j \in A$:

$$P(C'|A = a_i, C) = P(C'|A = a_j, C) = P(f(S^n)|C) \tag{6}$$

where C' is the content embedding. Then, assume an equal probability for different sensitive attributes $P(a_1|C) = ... = P(a_A|C)$, which is true for a fair dataset that has balanced texts from all groups within content C (can be obtained through our data augmentation strategy in Section 2.3). we can rewrite the Eq. 6 as:

$$\begin{aligned} P(C', A = a_i|C) &= P(C', A = a_j|C) \\ P(f(S^{a_i})|C) &= P(f(S^{a_j})|C) \end{aligned} \tag{7}$$

Following (Hinton and Roweis, 2002; Yang et al., 2023) the probability that C would pick a_i as its neighbor (in other words, the probability a_i conditioned on C):

$$P_{f(S^{a_i})|C} = \frac{\exp\left(-\frac{\|f(S_C^{a_i}) - f(S_C^n)\|^2}{2\rho^2}\right)}{\sum_{a_i \in A} \exp\left(-\frac{\|f(S_C^{a_i}) - f(S_C^n)\|^2}{2\rho^2}\right)} \tag{8}$$

where ρ is either set by hand or found by calculating the variance of $f(S^{a_i}) - f(S^n)$ over the whole training set (as in our experiments). Eq. 8 can be interpreted as follows: (1) Consider setting a Gaussian distribution with a covariance matrix equal to ρ times the identity matrix at the embedding of a neutral text S_C (with content C), which is denoted as $f(S_C^n)$. Then, a text with the same content but containing sensitive information a_i appears in the distribution with a probability proportional to $\exp\left(-\frac{\|f(S_C^{a_i}) - f(S_C^n)\|^2}{2\rho^2}\right)$, represented as the numerator. (2) The denominator aggregates the aforementioned probabilities across all sensitive groups $a_i \in A$ and serves as the normalization factor. Then we combine the Eq. 7 and Eq. 8 and

obtain:

$$\frac{\exp\left(-\frac{\|f(S_C^{a_i})-f(S_C^n)\|^2}{2\rho^2}\right)}{\sum_{a_i\in A}\exp\left(-\frac{\|f(S_C^{a_i})-f(S_C^n)\|^2}{2\rho^2}\right)} = \frac{\exp\left(-\frac{\|f(S_C^{a_j})-f(S_C^n)\|^2}{2\rho^2}\right)}{\sum_{a_j\in A}\exp\left(-\frac{\|f(S_C^{a_j})-f(S_C^n)\|^2}{2\rho^2}\right)}$$

$$\exp\left(-\frac{\|f(S_C^{a_i})-f(S_C^n)\|^2}{2\rho^2}\right) = \exp\left(-\frac{\|f(S_C^{a_i})-f(S_C^n)\|^2}{2\rho^2}\right)$$

$$\|f(S_C^{a_i})-f(S_C^n)\|^2 = \|f(S_C^{a_j})-f(S_C^n)\|^2 \tag{9}$$

Thus we obtain the definition 2.1. As a result, achieving conditional independence between sensitive attributes and content embeddings is equivalent to achieving content-conditioned equal distance. □