Beyond Gender: Addressing Age and Disability Biases with Context-Debias

Even Li **Jessica Wang** **Joanna Wang**
MIT 6.8610
{even1314, jessyw22, joannawa}@mit.edu

Abstract

Bias in pre-trained contextualized word embeddings, like BERT, can propagate to NLP applications, leading to unfair outcomes. While existing methods primarily address gender and racial biases, we propose extending Context-Debias to mitigate biases related to age and disability. This work identifies attribute-target associations, fine-tunes embeddings via orthogonal projections, and evaluates bias reduction and NLP task performance using SEAT and GLEU benchmarks. Our result reduces biases while preserving semantic integrity and downstream performance.

1 Introduction

Pre-trained language models (PLMs) like BERT have achieved state-of-the-art performance across numerous natural language processing (NLP) tasks. However, they also encode social biases prevalent in their training corpora, leading to the propagation of stereotypes in downstream applications (Ladhak et al., 2023a). Bias mitigation in these models is an ongoing challenge, with existing efforts primarily focused on gender and racial biases (Meade et al., 2022).

Contextualized embeddings present unique challenges for bias mitigation due to their dynamic nature, which reflects word meanings in specific contexts (Kaneko and Bollegala, 2021). While methods like Context-Debias (Kaneko and Bollegala, 2021) have effectively addressed gender biases, they have not been extensively applied to other dimensions, such as age and disability. This gap limits the inclusivity of NLP systems and underscores the need for extended methodologies.

In this work, we build upon the Context-Debias framework to mitigate biases related to age and disability in contextualized embeddings. Our contributions are as follows:

1. We extend the Context-Debias method to address age and disability biases, focusing on orthogonal projection techniques for bias mitigation.

2. We evaluate the effectiveness of our approach using the Sentence Encoder Association Test (SEAT) (May et al., 2019) and General Language Understanding Evaluation (GLUE) benchmarks (Wang et al., 2018), demonstrating significant bias reduction without sacrificing downstream performance.

3. We provide insights into the interplay between bias dimensions, emphasizing the need for comprehensive approaches that address multiple biases simultaneously.

Our findings indicate that mitigating age and disability biases is not only feasible but also critical for developing more equitable NLP systems. By addressing underexplored dimensions, this work contributes to the broader goal of creating fair and accountable AI applications.

2 Related Works

Bias mitigation in pre-trained language models (PLMs) has garnered significant attention as these models often encode stereotypes embedded in their training data, which can propagate into downstream applications. Existing approaches to debiasing can be broadly categorized into three levels of intervention: (1) data-level strategies, (2) embedding-level adjustments, and (3) model-level fine-tuning.

Data-Level Strategies. Counterfactual Data Augmentation (CDA) (Zmigrod et al., 2019; Webster et al., 2020) aims to mitigate bias at the data source by generating balanced counterfactual examples. For instance, demographic terms are swapped to create datasets that encourage models

to generalize beyond stereotypical patterns. Although effective in reducing biases at the pretraining stage, CDA requires full retraining of large models on modified corpora, making it computationally prohibitive for contextualized embeddings like BERT. Additionally, retraining introduces the risk of losing task-specific knowledge that PLMs acquire during their initial pretraining phase, making this approach less practical for addressing biases in widely used contextualized embeddings (Li et al., 2023).

Embedding-Level Adjustments. Post-processing techniques focus on mitigating biases within pretrained embeddings by altering their representations. For example, Sent-Debias (Liang et al., 2020) and Auto-Debias (Guo et al., 2022) project word embeddings away from bias subspaces, while FairFil (Cheng et al., 2021) employs adversarial training to suppress biased dimensions. These methods are computationally efficient as they do not require model retraining. However, many of these methods primarily mask rather than eliminate biases, leaving residual associations in the embeddings that may influence downstream tasks.

Model-Level Fine-Tuning. Fine-tuning-based approaches integrate fairness objectives during model adaptation to downstream tasks. For instance, MABLE (He et al., 2022) introduces fairness-aware loss functions to mitigate bias while preserving task-specific performance. These methods are particularly advantageous for contextualized embeddings, as they address biases without requiring full retraining. However, challenges such as the reintroduction of biases during fine-tuning and potential trade-offs between fairness and task accuracy remain prevalent (Gonen and Goldberg, 2019; Zhou et al., 2023).

Despite the progress in debiasing techniques, most existing methods predominantly focus on addressing gender and racial biases, leaving other dimensions, such as age and disability, largely underexplored (Meade et al., 2022; Yu et al., 2023; Ladhak et al., 2023b; Kaneko et al., 2024). Moreover, many techniques are designed for static word embeddings or specific downstream tasks, limiting their generalizability to modern contextualized embeddings (Dev et al., 2020; Nangia et al., 2020; Nadeem et al., 2020). Contextualized embeddings introduce unique challenges due to their dynamic, context-dependent nature, where biases emerge from both the target word and its co-occurring context.

For contextualized embeddings, Context-Debias(Kaneko and Bollegala, 2021) stands out by leveraging orthogonal projections to effectively reduce biased associations, demonstrating success in mitigating gender bias. However, its focus on gender bias and reliance on predefined word lists indicate the need for further extensions to address a broader range of bias dimensions comprehensively. This work seeks to address these gaps by extending the Context-Debias framework to mitigate biases related to age and disability in contextualized embeddings. By doing so, we aim to contribute to the development of more inclusive debiasing techniques, advancing the equity and accountability of NLP systems.

3 Methodology

Our approach extends the Context-Debias method (Kaneko and Bollegala, 2021) to mitigate biases related to age and disability in contextualized word embeddings. This is achieved through a two-part loss function: a debiasing loss and a regularization loss, which together reduce biased associations while preserving semantic information.

3.1 Debiasing Loss

The debiasing loss minimizes the inner product between the non-contextualized embedding of an attribute word (e.g., "elders," "teenagers") and the contextualized embedding of a target word (e.g., "burdens," "immature"). As illustrated in Figure 2, attribute words are selected (e.g., "elders," "teenagers," and "kids") and paired with target words that may have biased associations (e.g., "burdens," "immature," "naive").

To compute the non-contextualized embedding $v_i(a)$, we average the contextualized embeddings of the attribute word a across all sentences where it appears:

$$v_i(a) = \frac{1}{|\Omega(a)|} \sum_{x \in \Omega(a)} E_i(a, x; \theta_e),$$

where $\Omega(a)$ is the set of sentences containing a. The debiasing loss, which penalizes the inner product between the contextualized target embedding $E_i(t, x; \theta_e)$ and the non-contextualized attribute embedding $v_i(a)$, is defined as:

$$L_{\text{debias}} = \sum_{t \in V_t} \sum_{x \in \Omega(t)} \sum_{a \in V_a} \left(v_i(a)^\top E_i(t, x; \theta_e) \right)^2,$$

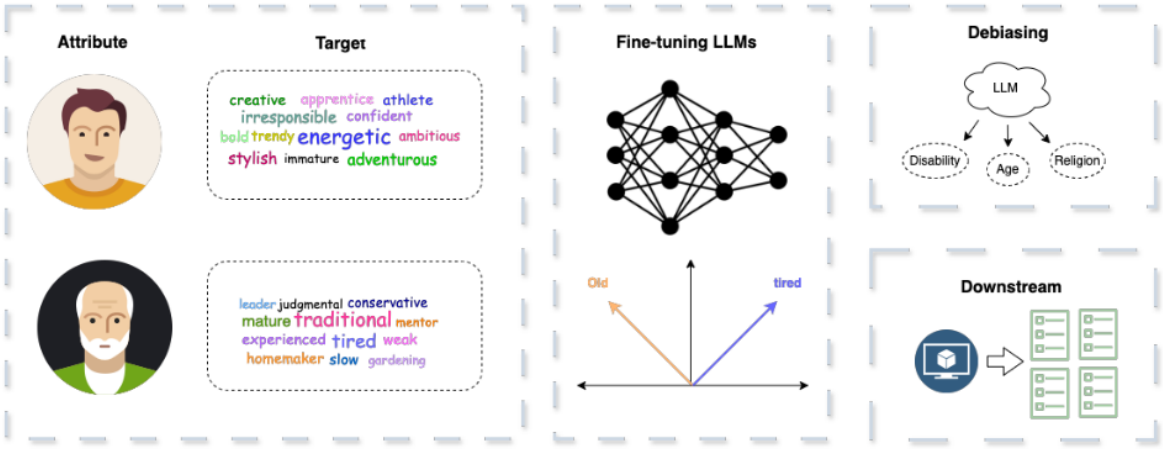

Figure 1: Overview of the proposed approach for mitigating age and disability biases. The process includes identifying attribute and target words, fine-tuning language models, and debiasing embeddings while maintaining downstream task performance.

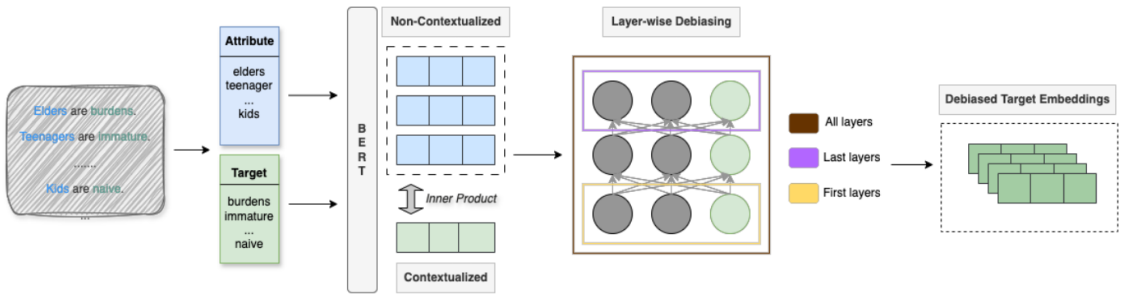

Figure 2: Layer-wise debiasing mechanism, showing the inner product minimization between non-contextualized attribute embeddings and contextualized target embeddings. Debiasing can be applied across specific layers (first, last, or all layers).

where V_t and V_a are the sets of target and attribute words, respectively. This process, highlighted in Figure 2, ensures that target word embeddings become orthogonal to attribute embeddings, effectively reducing biased associations layer by layer.

3.2 Regularization Loss

The regularization loss complements the debiasing loss by preserving the semantic information in the original embeddings. It minimizes the squared ℓ_2-distance between the contextualized embeddings of the original (pretrained) and debiased models for each word w in a sentence x:

$$L_{\text{reg}} = \sum_{x \in \mathcal{A}} \sum_{w \in x} \sum_{i=1}^{N} \|E_i(w, x; \theta_e) - E_i(w, x; \theta_{\text{pre}})\|^2,$$

where $\mathcal{A}$ is the set of sentences containing attribute words, and N is the total number of layers. This step ensures that the structure of contextualized embeddings remains intact while biases are reduced.

3.3 Overall Framework

Figure 1 provides an overview of the entire framework, showing how the process integrates attribute and target words, fine-tunes embeddings through layer-wise debiasing, and evaluates the resulting embeddings in downstream tasks. The final objective combines the debiasing and regularization losses as:

$$L = \alpha L_{\text{debias}} + \beta L_{\text{reg}},$$

where α and β are hyperparameters balancing bias mitigation and semantic preservation.

By applying this framework, we ensure that large language models (LLMs) are debiased across

multiple layers while retaining their downstream utility.

4 Experiments

4.1 Data

To evaluate the effectiveness of our debiasing framework, we created word lists to represent two bias dimensions: **age-related biases** and **disability-related biases**.

Disability-Related Biases: Word lists for disability-related biases were manually curated from the Disability Language Style Guide (National Center on Disability and Journalism, 2024). The attribute words were categorized into *mental* (e.g., "depression," "anxiety") versus *physical* (e.g., "deaf," "blind") disabilities, while the target words consisted of stereotypical associations with these categories (e.g., "burden," "crazy"). This categorization allowed us to systematically assess the specific biases embedded in contextualized embeddings.

Age-Related Biases: Similarly, the word lists for age-related biases were constructed from (Gendron et al., 2015). Attribute words were categorized as *young* (e.g., "teenager," "child") versus *old* (e.g., "elderly," "retiree"), and target words reflected stereotypes associated with these categories (e.g., "immature," "stubborn").

These word lists were used to identify sentences from the News-Commentary v15 corpus (Statistical Machine Translation, 2024), which provided a diverse dataset for finetuning the debiased embedding. Each sentence contained either an attribute word or a target word in our defined lists.

4.2 Bias Evaluation Settings

To systematically analyze the performance of our debiasing framework, we conducted experiments under 12 distinct configurations that varied across three dimensions.

- **Debiasing Granularity:** We applied debiasing at both the *sentence level* (optimizing the loss L on all tokens in a sentence containing a target word) and *token level* (computing L only on the target word in the sentence).
- **Layer Selection:** Debiasing was applied to the *first*, *last*, or *all* layers of the contextualized embedding model.
- **Bias Dimensions:** These experiments were conducted for both age-related and disability-related biases, resulting in 6 settings per bias dimension and a total of 12 settings.

Evaluation Metrics: We assessed debiasing performance using the Sentence Encoder Association Test (SEAT) (May et al., 2019), a well-established benchmark for measuring bias in word embeddings. SEAT tests compare the relative strength of associations between pairs of attribute and target word sets, akin to the Implicit Association Test (IAT) in psychology(May et al., 2019).

- **SEAT9:** This test evaluates biases in embeddings by measuring the association between *mental disability* words and *physical disability* words with the attributes *Temporary* and *Permanent*.
- **SEAT10:** This test examines the association between *young people's names* and *old people's names* with the attributes *Pleasant* and *Unpleasant*.

The SEAT tests provide effect sizes and p-values to quantify the strength and statistical significance of the observed biases. Lower effect sizes and non-significant p-values reflect successful bias mitigation. The results of SEAT9 and SEAT10 are presented in Table 1, illustrating how effectively our method reduces biased associations in contextualized embeddings while maintaining semantic utility.

4.3 Downstream Task Evaluation

To ensure that debiasing does not negatively impact model performance on practical NLP tasks, we evaluated the debiased embeddings on the General Language Understanding Evaluation (GLEU) benchmark (Wang et al., 2018). GLEU includes five tasks: **Stanford Sentiment Treebank (SST-2)**, **Microsoft Research Paraphrase Corpus (MRPC)**, **Semantic Textual Similarity Benchmark (STS-B)**, **Recognizing Textual Entailment (RTE)**, and **Winograd Schema Challenge (WNLI)**.

GLEU is a widely used benchmark for testing the generalizability of embeddings across multiple linguistic tasks. It provides a comprehensive assessment of whether the debiasing process preserves semantic information critical for downstream performance. By evaluating on tasks rang-

Model	Layer	Unit	SEAT-9	SEAT-10	†#
Disability	all	token	**0.04**	0.50†	1
		sent	0.48	0.45†	1
	last	token	0.28	**0.43**†	1
		sent	0.25	0.63†	1
	first	token	0.25	0.57†	1
		sent	0.21	0.60†	1
	original		0.51†	0.51†	2
Age	all	token	0.51	**0.13**	0
		sent	-0.47	0.43†	1
	last	token	-0.28	0.73†	1
		sent	-0.85	0.57†	1
	first	token	**-0.23**	0.64†	1
		sent	0.94	0.97†	1
	original		0.51†	0.51†	2

Table 1: Disability and age bias of contextualized embeddings on SEAT9 and SEAT10. † denotes significant bias effects at $\alpha < 0.05$.

ing from sentiment classification to textual entailment, GLEU ensures that the debiased embeddings remain versatile and effective for real-world applications.

The results of the GLEU benchmark are summarized in Table 2. These results provide insights into the trade-offs between reducing biases and preserving downstream task performance, which will be discussed in detail in the conclusion and discussion section.

5 Results

5.1 Debiasing Performance

Table 1 summarizes the results of our debiasing experiments using SEAT9 and SEAT10, comparing pre-trained embeddings with various debiasing configurations. We found that the best results for both disability and age debiasing were achieved with token-level debiasing applied across all layers of the contextualized embedding model. For disability debiasing, the effect size for SEAT9 decreased from 0.51 to 0.04, with a non-significant p-value. Similarly, for age debiasing, the effect size for SEAT10 was reduced from 0.51 to 0.13. These results demonstrate the effectiveness of our method in mitigating age and disability biases across contextualized embeddings.

Interestingly, debiasing one dimension influenced bias metrics in another. For example, debiasing for age impacted SEAT9 results, highlighting the complex interplay between different bias dimensions. This cross-dimensional influence underscores the importance of developing methods that simultaneously address multiple biases, as addressing them in isolation may yield unintended interactions.

Certain configurations, such as debiasing only the first or last layer, or applying debiasing at the sentence level, were less effective. In some cases, these configurations led to the unintended emphasis of age bias (as seen in SEAT10 scores). This could be due to incomplete neutralization of attribute-target associations in intermediate layers, reinforcing the importance of a comprehensive, all-layer approach like token-level debiasing applied across all layers.

5.2 Downstream Task Performance

Table 2 presents the downstream task performance evaluated on the GLEU benchmark. Debiased embeddings consistently preserved performance levels comparable to their original pretrained counterparts across a range of tasks. For example, the all-token model debiased for disability achieved an average GLEU score of 74.2%, identical to the performance of the original embeddings. This result confirms that the proposed debiasing method preserves the semantic integrity required for accurate downstream predictions.

Additionally, the method showed robustness across a variety of tasks, including sentiment classification (SST-2), paraphrase detection (MRPC),

Model	Layer	Unit	SST-2	MRPC	STS-B	RTE	WNLI	Avg
Disability	all	token	91.2	84.2	81.3	60.0	54.3	74.2
		sent	91.2	82.3	80.8	58.4	55.0	73.5
	last	token	91.2	83.8	81.6	57.8	55.3	73.9
		sent	91.6	84.5	82.0	55.6	52.6	73.3
	first	token	90.0	85.8	82.1	59.9	53.1	74.2
		sent	90.5	86.8	81.9	57.6	53.5	74.1
	original	-	90.0	85.6	82.1	59.8	53.5	**74.2**
Age	all	token	90.0	85.8	82.1	59.0	52.8	73.9
		sent	90.1	84.7	82.2	61.4	52.6	74.2
	last	token	91.4	83.4	81.6	61.2	52.2	74.0
		sent	91.9	85.1	82.0	60.3	51.5	74.2
	first	token	90.7	85.5	82.0	60.3	51.9	74.1
		sent	90.8	83.8	82.3	61.5	51.1	73.9
	original	-	90.0	85.6	82.1	59.8	53.5	**74.2**

Table 2: GLEU evaluation of debiased embeddings across five downstream NLP tasks. The average performance (Avg) summarizes task utility after debiasing.

and textual entailment (RTE), demonstrating the versatility and real-world applicability of the debiased embeddings. The average GLEU scores for debiased and original embeddings across tasks were nearly identical, indicating that the debiasing process did not significantly degrade task performance.

6 Discussion

Our proposed debiasing method demonstrates promising results in mitigating biases related to age and disability. However, it also highlights several limitations and potential areas for future work. A notable limitation is the focus on only two dimensions of bias, leaving other critical dimensions—such as socioeconomic status, religion, and intersectional biases—unaddressed. Addressing these biases often requires tailored approaches that extend beyond the scope of this study. Furthermore, our findings reveal cross-dimensional interactions, where debiasing one dimension (e.g., age) inadvertently impacts another (e.g., disability), as observed in the SEAT results. These interactions underscore the complexity of bias mitigation and emphasize the need for methods that can disentangle and simultaneously address multiple dimensions of bias.

Another challenge lies in the reliance on manually curated word lists for attribute-target pairs. While effective, these lists introduce subjectivity and are limited in their scalability and representativeness. They may fail to capture the full range of nuanced biases in diverse datasets, particularly in multilingual or domain-specific contexts. Automating the identification of these pairs using techniques such as unsupervised learning or knowledge graphs could mitigate this limitation and improve the generalizability of debiasing methods.

Future work could build on these findings by exploring several promising directions. First, developing a unified framework for multidimensional debiasing that explicitly accounts for cross-dimensional interactions is essential. Such a framework would enable a more comprehensive and inclusive approach to bias mitigation. Second, advancing automated bias detection methods—such as leveraging pretrained language models or knowledge-based resources to identify attribute-target pairs—could reduce the dependency on manual curation and improve scalability.

Finally, extending the proposed method to other contextualized embedding architectures, such as GPT, ALBERT, and multilingual embeddings, is critical for broadening its applicability. These extensions could address biases across a wider range of tasks, including cross-lingual applications. Addressing these challenges will enhance the inclusivity, robustness, and scalability of debiasing methods, ultimately contributing to the development of fairer and more equitable NLP systems.

7 Conclusion

This work extends the Context-Debias framework to mitigate biases related to age and disability in contextualized word embeddings. By employing token-level debiasing across all layers, our approach achieves significant bias reduction, as demonstrated by non-significant SEAT effect sizes, while preserving downstream task performance with GLEU scores comparable to the original embeddings. These results highlight the effectiveness of our method in addressing biases without sacrificing utility.

Building on prior efforts primarily focused on gender and racial biases, this study broadens the scope of bias mitigation to underexplored dimensions, demonstrating that contextualized embeddings can be effectively debiased for age and disability. By addressing these critical dimensions, the proposed approach advances fairness in NLP systems while ensuring robustness across diverse tasks. This work contributes to the growing body of literature on practical and effective debiasing techniques, representing a step towards more equitable and accountable applications of language technologies.

References

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *Proceedings of the International Conference on Learning Representations*.

Sunipa Dev, Tao Li, Jeff Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 7659–7666.

Tracey Gendron, E. Welleford, Jenny Inker, and Jay White. 2015. The language of ageism: Why we need to use words carefully. *The Gerontologist*, 56.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 60–63, Florence, Italy. Association for Computational Linguistics.

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 (ACL)*, pages 1012–1023. Association for Computational Linguistics.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. 2022. Mabel: Attenuating gender bias using textual entailment data. *arXiv preprint arXiv:2210.14975*.

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.

Masahiro Kaneko, Danushka Bollegala, and Timothy Baldwin. 2024. The gaps between pre-train and downstream settings in bias evaluation and debiasing. *arXiv preprint arXiv:2401.08511*.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto. 2023a. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto. 2023b. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Conference of the European Chapter of the Association for Computational Linguistics*.

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)*, pages 14254–14267, Toronto, Canada. 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.

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

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)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 617–622.

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*, Online. Association for Computational Linguistics.

National Center on Disability and Journalism. 2024. Disability language style guide. Accessed: 2024-12-07.

Statistical Machine Translation. 2024. News commentary v16 dataset. Accessed: 2024-12-07.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. 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*.

Liu Yu, Yuzhou Mao, Jin Wu, and Fan Zhou. 2023. Mixup-based unified framework to overcome gender bias resurgence. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1755–1759.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. pages 4227–4241.

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661.

8 Impact Statement

Our project tackles critical ethical and societal issues by addressing biases in pre-trained language models (PLMs), specifically focusing on age and disability biases in contextualized word embeddings. Language models are increasingly integral to applications that impact everyday life, such as hiring algorithms, healthcare triage systems, and customer service chatbots. Biases within these systems can propagate and exacerbate societal inequalities, particularly for vulnerable populations such as older adults and individuals with disabilities. By mitigating these biases, our work aims to create more equitable and inclusive AI systems, ensuring that such groups are not unfairly disadvantaged in automated decision-making processes.

From a societal perspective, our research contributes to broader efforts to develop fair and equitable AI systems. By providing tools to systematically address biases, our work allows developers, organizations, and policymakers to critically evaluate the ethical implications of deploying NLP models in diverse and sensitive contexts. However, the reliance on curated word lists and current evaluation metrics poses limitations in scalability and applicability across different cultural or linguistic domains. Future iterations must consider how these tools can adapt to new contexts and evolving definitions of fairness. Furthermore, mitigating biases is only one component of responsible AI development; ensuring transparency, accountability, and stakeholder inclusion in these processes remains an ongoing challenge. This research marks our attempt toward creating more socially responsible AI, but it must be integrated with broader systemic efforts to address inequality and ensure justice in AI-powered applications.