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

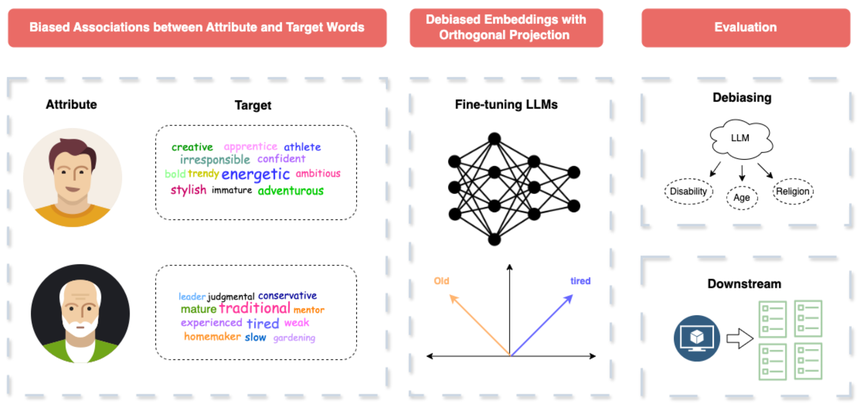
Even Li Jessica Wang Joanna Wang

Motivations

Motivating Problem
- Bias in pre-trained contextualized word embeddings, like BERT, can propagate to NLP applications, leading to unfair outcomes
- Existing debiasing methods mainly address gender and racial biases, with limited attention to biases related to age, disability, and religion

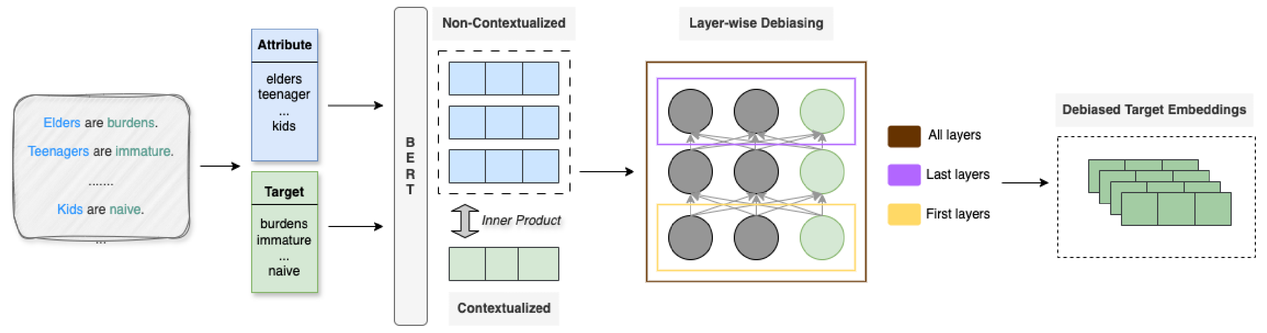
Challenges
- Contextualized embeddings have numerous parameters that interact in complex ways, making it difficult to identify and address sources of bias
- Common debiasing methods for static embeddings, such as projection-based techniques, are not directly applicable to contextualized embeddings
- For contextualized embeddings, the same word may exhibit bias in some contexts but not in others

Objectives



- Identify biased associations between attribute (e.g., age) and target words in contextualized embeddings
- Apply the Context-Debias method using orthogonal projection to fine-tune LLMs and mitigate biases
- Evaluate the debiased embeddings on under-explored dimensions for both debiasing effectiveness and downstream NLP performance

Method



Main Framework

- **Identify Attribute and Target Words**: Select attribute words representing biases (e.g., "elders," "teenagers") and target words reflecting stereotypical associations (e.g., "burdens," "immature")
- **Generate Embeddings**: Use a pre-trained BERT to create contextualized embeddings for target words; non-contextualized embeddings for attribute words are obtained by averaging the embeddings across all contexts
- **Loss Calculation**: Compute the inner product between attribute and target embeddings to quantify biased associations
- **Layer-wise Debiasing**: Apply orthogonal projections to specific layers of the model to reduce biases while preserving semantic meaning

Data Sources
We collected attribute and target words related to age from the word lists created by *Tracey L. Gendron et al,* and related to disability from the "*Disability Language Style Guide"* by the *National Center on Disability and Journalism*. Relevant biased sentences were extracted from the *News-Commentary v15 corpus*.

Loss Minimization

Debiasing loss: Squared inner product between attribute embeddings $V_i(a)$ and contextualized target embeddings $E_i(t, x; \theta_e)$

$$L_i = \sum_{t \in \mathcal{V}_t} \sum_{x \in \Omega(t)} \sum_{a \in \mathcal{V}_a} \left(\boldsymbol{v}_i(a)^\top E_i(t, x; \boldsymbol{\theta}_e) \right)^2$$

Regularization loss: Keep debiased version of the model stays close to the original version in terms of the learned representations

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

Total loss to be minimized is a weighted combination of the debiasing loss L_i and the regularization loss $L_{reg.}$ In our experiments, we set α=0.2 and β=0.8

$$L = \alpha L_i + \beta L_{reg}$$

Experiments

Our Experiments

Token-Level: Apply debiasing specifically to the target word in a sentence
Sentence-Level: Debias all words in a sentence that contains a target word
Layer-Level: Debiasing on either the first layer, the last layer, or all layers of the pretrained model

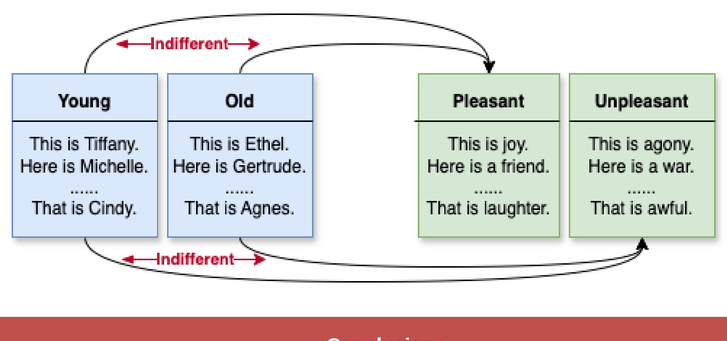
Evaluation Metrics

Bias Evaluation with Sentence Encoder Association Test (SEAT)

- We use SEAT to measure bias reduction for age and disability dimensions
- For each debiasing dimension, we collect two sets of target words (e.g., "old" vs. "young") and two sets of attribute words (e.g., "pleasant" vs. "unpleasant")
- We compute cosine similarities between sentence embeddings of target-attribute pairs to quantify bias

GLEU benchmark

- We use the GLEU benchmark to evaluate downstream performance after debiasing to ensure debiasing does not negatively impact model performance
- This benchmark comprises five tasks: Stanford Sentiment Treebank (SST-2), Microsoft Research Paraphrase Corpus (MRPC), Semantic Textual Similarity Benchmark (STS-B), Recognising Textual Entailment (RTE), Winograd Schema Challenge (WNLI)



Results and Evaluations

SEAT: Debiasing Evaluations

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 $\alpha < 0.05$.

- Best results for disability and gender debiasing are achieved with token-level debiasing across all layers, aligning with findings from the original context debias paper for gender.
- Disability Debiasing (SEAT9): Effect size reduced from 0.51 (pretrained) to 0.04 (non-significant)
- Age Debiasing (SEAT10): Effect size reduced from 0.51 (pretrained) to 0.13 (non-significant)
- Debiasing one dimension can influence performance on others. For example, age debiasing impacts SEAT9 performance. This highlights the need for methods that address multiple dimensions simultaneously.

GLEU: Downstream Performances

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**

- Debiased embeddings perform comparably to the original BERT embeddings in most settings
- Our proposed debiasing method effectively preserves the semantic information necessary for accurate downstream NLP predictions

Conclusions

Results
- Extended Context-Debias to address disability-related and age-related biases beyond gender
- Tested debiasing techniques at both token and sentence levels, as well as specific BERT layers
- Demonstrated effectiveness in reducing biases (e.g., SEAT9 and SEAT10) while maintaining NLP task performance (measured via GLEU)

Limitations
- Experimented with only one set of attribute and target word lists; need to evaluate the impact of different word lists
- Used News-Commentary v15 corpus, which lacks significant biased sentences; alternative training data with more biases should be considered

Future Steps
- Develop methods to address multi-dimensional biases simultaneously
- Investigate debiasing methods that prevent the reintroduction of biases during fine-tuning for downstream tasks

References

- Gendron, T. L., Welleford, E. A., Inker, J., & White, J. T. 2016. The Language of Ageism: Why We Need to Use Words Carefully. The Gerontologist, Volume 56, issue 6, pages 997–1006, Online. The Gerontologist.
- 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.
- National Center on Disability and Journalism. "NCDJ Style Guide." 2024. https://ncdj.org/style-guide/.
- Squiduu. "Auto-Debias Reproduction." GitHub repository, 2024. https://github.com/squiduu/auto-debias-reproduction.
- W4ngatang. "Sent-Bias: Tests for Sentence-Level WEAT." GitHub repository, 2024. https://github.com/W4ngatang/sent-bias/blob/master/tests/sent-weat10.jsonl