cases

doc_1		doc_2		decision	id
authors	• Jacqueline He • Mengzhou Xia • Christiane Fellbaum • Danqi Chen	**authors**	• Chen, Danqi • Fellbaum, Christiane • He, Jacqueline • Xia, Mengzhou	DUPLICATES	146
title	MABEL: Attenuating Gender Bias using Textual Entailment Data	**title**	MABEL: Attenuating Gender Bias using Textual Entailment Data		
publication_date	2022-10-26 00:00:00	**publication_date**	2022-10-26 00:00:00		
source	SupportedSources.INTERNET_ARCHIVE	**source**	SupportedSources.CORE		
journal		**journal**			
volume		**volume**			
doi		**doi**	None		
urls	• https://web.archive.org/web/20221028065707/https://arxiv.org/pdf/2210.14975v1.pdf	**urls**	• http://arxiv.org/abs/2210.14975		
id	id-3272820383237452972	**id**	id8560142824019134244		
abstract	Pre-trained language models encode undesirable social biases, which are further exacerbated in downstream use. To this end, we propose MABEL (a Method for Attenuating Gender Bias using Entailment Labels), an intermediate pre-training approach for mitigating gender bias in contextualized representations. Key to our approach is the use of a contrastive learning objective on counterfactually augmented, gender-balanced entailment pairs from natural language inference (NLI) datasets. We also introduce an alignment regularizer that pulls identical entailment pairs along opposite gender directions closer. We extensively evaluate our approach on intrinsic and extrinsic metrics, and show that MABEL outperforms previous task-agnostic debiasing approaches in terms of fairness. It also preserves task performance after fine-tuning on downstream tasks. Together, these findings demonstrate the suitability of NLI data as an effective means of bias mitigation, as opposed to only using unlabeled sentences in the literature. Finally, we identify that existing approaches often use evaluation settings that are insufficient or inconsistent. We make an effort to reproduce and compare previous methods, and call for unifying the evaluation settings across gender debiasing methods for better future comparison.	**abstract**	Pre-trained language models encode undesirable social biases, which are further exacerbated in downstream use. To this end, we propose MABEL (a Method for Attenuating Gender Bias using Entailment Labels), an intermediate pre-training approach for mitigating gender bias in contextualized representations. Key to our approach is the use of a contrastive learning objective on counterfactually augmented, gender-balanced entailment pairs from natural language inference (NLI) datasets. We also introduce an alignment regularizer that pulls identical entailment pairs along opposite gender directions closer. We extensively evaluate our approach on intrinsic and extrinsic metrics, and show that MABEL outperforms previous task-agnostic debiasing approaches in terms of fairness. It also preserves task performance after fine-tuning on downstream tasks. Together, these findings demonstrate the suitability of NLI data as an effective means of bias mitigation, as opposed to only using unlabeled sentences in the literature. Finally, we identify that existing approaches often use evaluation settings that are insufficient or inconsistent. We make an effort to reproduce and compare previous methods, and call for unifying the evaluation settings across gender debiasing methods for better future comparison.Comment: Accepted to EMNLP 2022. Code and models are publicly available at https://github.com/princeton-nlp/mabe		
versions		**versions**			