cases	doc_1		doc_2		decision	id
	authors	• Hongyin Luo • James Glass	**authors**	• Hongyin Luo • James Glass	DUPLICATES	9
	title	Logic Against Bias: Textual Entailment Mitigates Stereotypical Sentence Reasoning	**title**	Logic Against Bias: Textual Entailment Mitigates Stereotypical Sentence Reasoning		
	publication_date	2023-03-10 02:52:13+00:00	**publication_date**	2023-03-10 00:00:00		
	source	SupportedSources.ARXIV	**source**	SupportedSources.INTERNET_ARCHIVE		
	journal	None	**journal**			
	volume		**volume**			
	doi		**doi**			
	urls	• http://arxiv.org/pdf/2303.05670v1 • http://arxiv.org/abs/2303.05670v1 • http://arxiv.org/pdf/2303.05670v1	**urls**	• https://web.archive.org/web/20230313043854/https://arxiv.org/pdf/2303.05670v1.pdf		
	id	id4502012054496866620	**id**	id5525618078148985153		
	abstract	Due to their similarity-based learning objectives, pretrained sentence encoders often internalize stereotypical assumptions that reflect the social biases that exist within their training corpora. In this paper, we describe several kinds of stereotypes concerning different communities that are present in popular sentence representation models, including pretrained next sentence prediction and contrastive sentence representation models. We compare such models to textual entailment models that learn language logic for a variety of downstream language understanding tasks. By comparing strong pretrained models based on text similarity with textual entailment learning, we conclude that the explicit logic learning with textual entailment can significantly reduce bias and improve the recognition of social communities, without an explicit de-biasing process	**abstract**	Due to their similarity-based learning objectives, pretrained sentence encoders often internalize stereotypical assumptions that reflect the social biases that exist within their training corpora. In this paper, we describe several kinds of stereotypes concerning different communities that are present in popular sentence representation models, including pretrained next sentence prediction and contrastive sentence representation models. We compare such models to textual entailment models that learn language logic for a variety of downstream language understanding tasks. By comparing strong pretrained models based on text similarity with textual entailment learning, we conclude that the explicit logic learning with textual entailment can significantly reduce bias and improve the recognition of social communities, without an explicit de-biasing process		
	versions		**versions**			