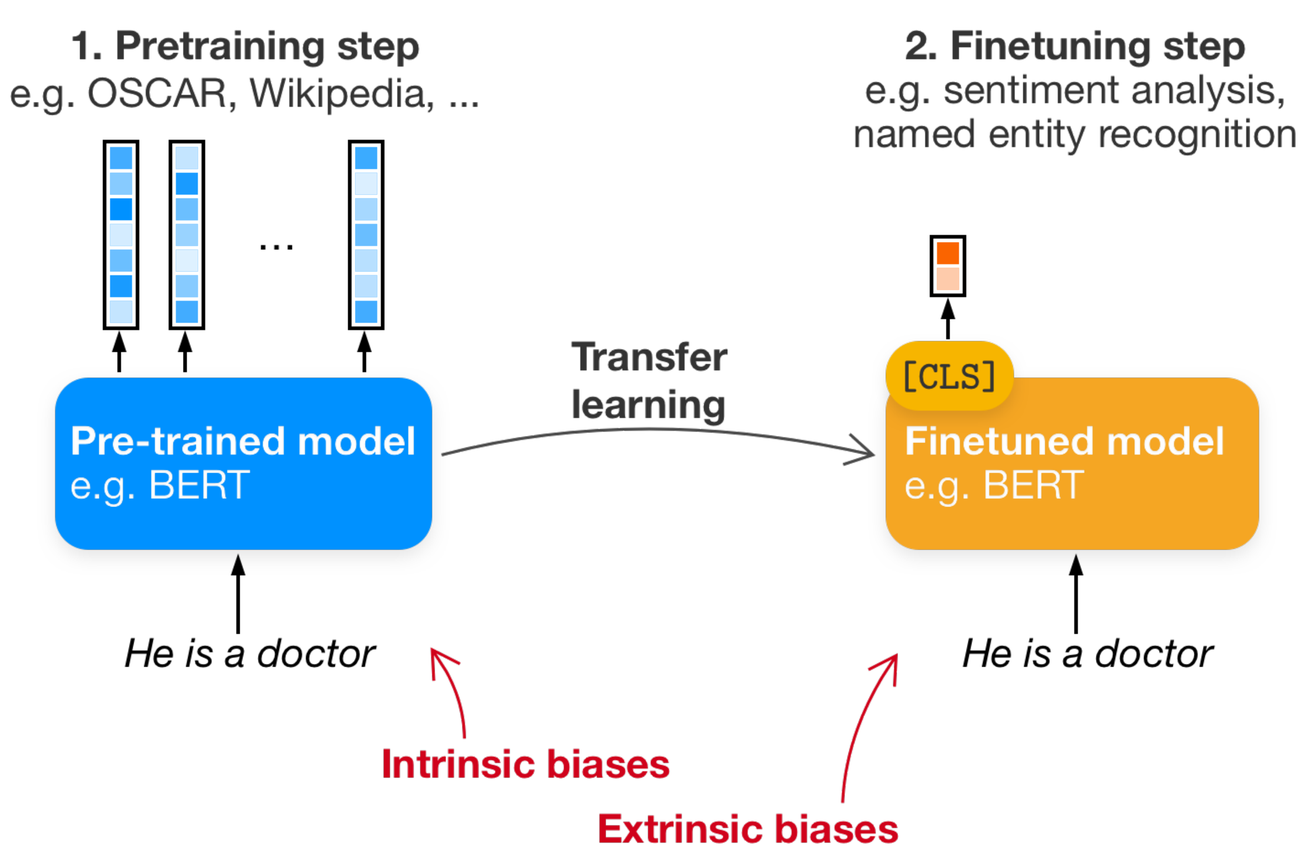
Measuring Fairness with Biased Rulers:
A Comparative Study on Bias Metrics for Pre-trained Language Models

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, Bettina Berendt
pieterdelobelle

What is *fairness* for language models?

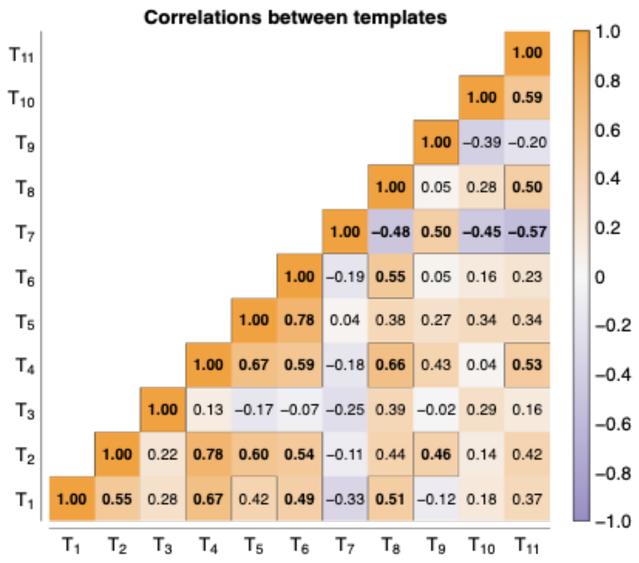


Fairness metrics

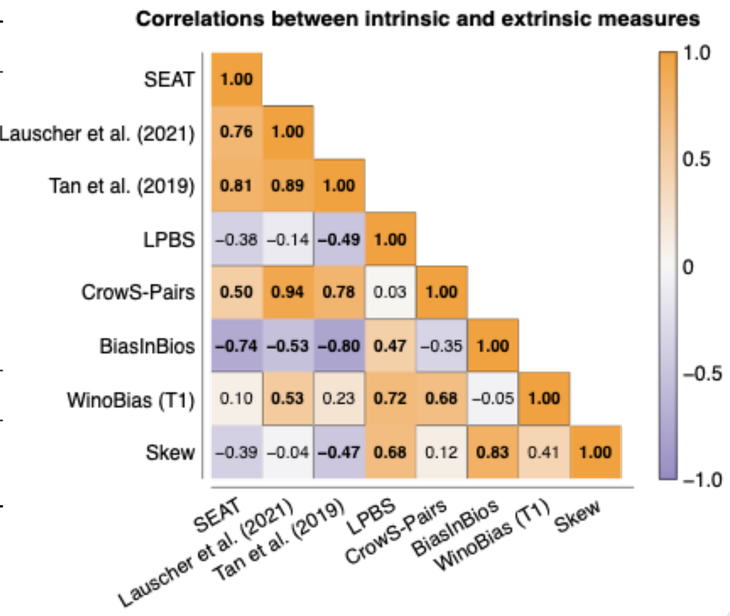
DisCo (Webster et al., 2020)
Lauscher et al. (2021)
LPBS (Kurita et al., 2019)
BEC-Pro (Bartl et al., 2020)
Based on WEAT
SEAT (May et al., 2019)
Lauscher et al. (2021)
Tan and Celis (2019)
CEAT (Guo and Caliskan, 2021)
CAT (Nadeem et al., 2021)
CrowS-Pairs (Nangia et al., 2020)
Basta et al. (2019)
Zhao et al. (2019)
Sedoc and Ungar (2019)

Fairness metrics don't correlate

We found that metrics are not compatible with each other and highly depend on (i) templates, (ii) attribute and target seeds and (iii) the choice of embeddings.



#	Type	Template sentence	D_{KL}
T_1	Bl.	"This is the _."	—
T_2	Bl.	"That is the _."	0.05
T_3	Bl.	"There is the _."	0.06
T_4	Bl.	"Here is the _."	0.13
T_5	Bl.	"The _ is here."	0.22
T_6	Bl.	"The _ is there."	0.14
T_7	Bl.	"The _ is a person."	0.17
T_8	Bl.	"It is the _."	0.05
T_9	Bl.	"The _ is a [MASK]."	0.83
T_{10}	Unbl.	"The _ is an engineer."	1.49
T_{11}	Unbl.	"The _ is a nurse with superior technical skills."	0.72



Our advice

Use a mix of some intrinsic measures of fairness that don't use embeddings directly and eliminate one source of variance, for example DisCo or LPBS. However, we also recommend to **perform extrinsic fairness evaluations on downstream tasks**.

YouTube

blog post

KU LEUVEN LEUVEN.AI University of Antwerp AI FLANDERS BUILDING OUR DIGITAL FUTURE berlin weizenbaum institut DTAI Declarative Languages & Artificial Intelligence