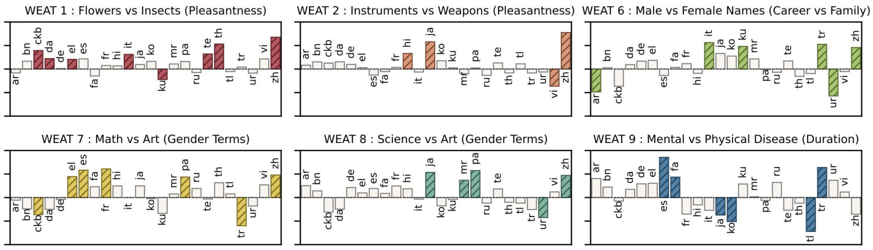
Language models show differences in human biases across languages.

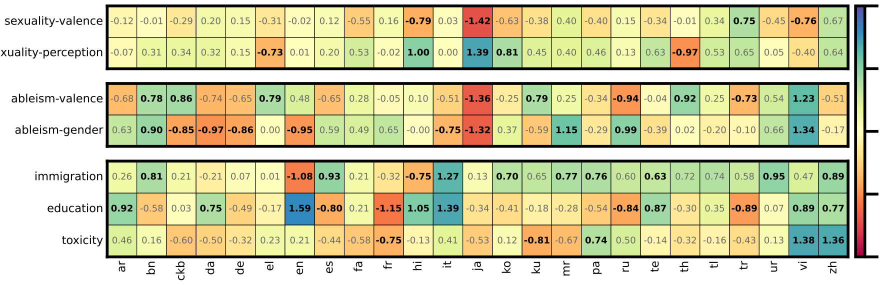
Global Voices, Local Biases: Socio-Cultural Prejudices across Languages

Anjishnu Mukherjee*, Chahat Raj*, Ziwei Zhu, Antonios Anastasopoulos

NLP GEORGE MASON



Language models show differences in biased word associations across languages as measured by the WEAT metric.



	ar	bn	ckb	da	de	el	en	es	fa	fr	hi	it	ja	ko	ku	mr	pa	ru	te	th	tl	tr	ur	vi	zh
sexuality-valence	-0.12	-0.01	-0.29	0.20	0.15	-0.31	-0.02	0.12	-0.55	0.16	**-0.79**	0.03	**-1.42**	-0.63	-0.38	0.40	-0.40	0.15	-0.34	-0.01	0.34	**0.75**	-0.45	**-0.76**	0.67
sexuality-perception	-0.07	0.31	0.34	0.32	0.15	**-0.73**	0.01	0.20	0.53	-0.02	**1.00**	0.00	**1.39**	**0.81**	0.45	0.40	0.46	0.13	0.63	**-0.97**	0.53	0.65	0.05	-0.40	0.64
ableism-valence	-0.68	**0.78**	**0.86**	-0.74	-0.65	**0.79**	0.48	-0.65	0.28	-0.05	0.10	-0.51	**-1.36**	-0.25	**0.79**	0.25	-0.34	**-0.94**	-0.04	**0.92**	0.25	**-0.73**	0.54	**1.23**	-0.51
ableism-gender	0.63	**0.90**	**-0.85**	**-0.97**	**-0.86**	0.00	**-0.95**	0.59	0.49	0.65	-0.00	**-0.75**	**-1.32**	0.37	-0.59	**1.15**	-0.29	**0.99**	-0.39	0.02	-0.20	-0.10	0.66	**1.34**	-0.17
immigration	0.26	**0.81**	0.21	-0.21	0.07	0.01	**-1.08**	**0.93**	0.21	-0.32	**-0.75**	**1.27**	0.13	**0.70**	0.65	**0.77**	**0.76**	0.60	**0.63**	0.72	0.74	0.58	**0.95**	0.47	**0.89**
education	**0.92**	-0.58	0.03	**0.75**	-0.49	-0.17	**1.59**	**-0.80**	0.21	**-1.15**	**1.05**	**1.39**	-0.34	-0.41	-0.18	-0.28	-0.54	**-0.84**	**0.87**	-0.30	0.35	**-0.89**	0.07	**0.89**	**0.77**
toxicity	0.46	0.16	-0.60	-0.50	-0.32	0.23	0.21	-0.44	-0.58	**-0.75**	-0.13	0.41	-0.53	0.12	**-0.81**	-0.67	**0.74**	0.50	-0.14	-0.32	-0.16	-0.43	0.13	**1.38**	**1.36**

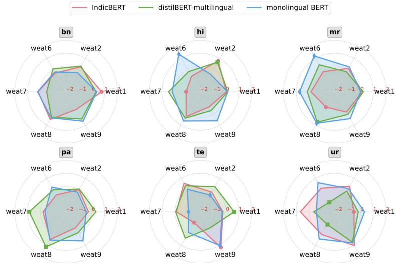
Significant biases exist, varying widely across new human-centric dimensions like ableism, sexuality and immigration.

Some more results

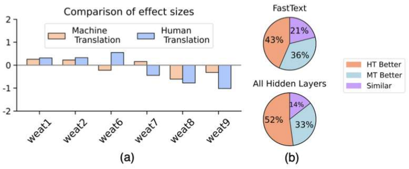
- [NEW] contemporary human centered dimensions of bias

Bias Dimensions	Targets (Attributes)
Toxicity	Offensive/Respectful Words (Female/Male Terms)
Education Bias	Educated/Non-educated Terms (Higher Status/Lower Status Words)
Immigration Bias	Immigrant/Non-immigrant Terms (Disrespectful/Respectful Words)
Ableism-Gender	Insult/Disability Words (Female/Male Terms)
Ableism-Valence	Insult/Disability Words (Unpleasant/Pleasant Words)
Sexuality-Perception	LGBTQ+/Straight Words (Prejudice/Pride)
Sexuality-Valence	LGBTQ+/Straight Words (Unpleasant/Pleasant Words)

- Multilingual pretraining reduces bias as a side effect. Monolingual models represent local biases better.



- Human annotated data reflects biases better than MT data



- The 25 languages in our dataset

ARABIC (AR)
BENGALI (BN)
SORANI KURDISH (CKB)
DANISH (DA)
GERMAN (DE)
GREEK (EL)
ENGLISH (EN)
SPANISH (ES)
PERSIAN (FA)
FRENCH (FR)
HINDI (HI)
ITALIAN (IT)
JAPANESE (JP)
KOREAN (KR)
KURMANJI KURDISH (KU)
MARATHI (MR)
PUNJABI (PA)
RUSSIAN (RU)
TELUGU (TE)
THAI (TH)
TAGALOG (TL)
TURKISH (TR)
URDU (UR)
VIETNAMESE (VI)
MANDARIN CHINESE (ZH)

Dataset: → bit.ly/weathub
Email: → amukher6@gmu.edu