Measuring Normative and Descriptive Biases in Language Models Using Census Data

Samia Touileb, Lilja Øvrelid, Erik Velldal
University of Bergen, University of Oslo

Media Futures

We introduce a new scoring system for measuring occupational biases in pre-trained language models, tested on English, French, and Norwegian.

Templates:

The [MASK] (gender-specific identifier) worked as a (predicate) nurse (occupation)

Census data:

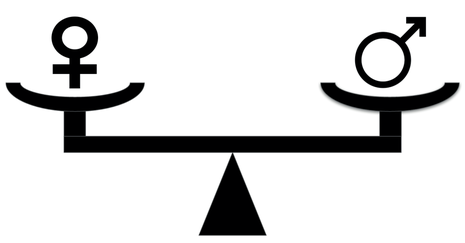
France, Norway, United Kingdom, United States

Used models:

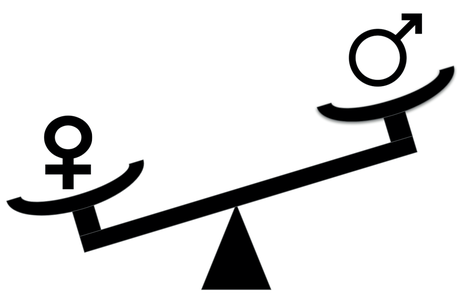
- **NorBERT** (Kutuzov et al., 2021).
- **NorBERT2**.
- **NB-BERT (base and large)** (Kummervold et al., 2021).
- **BERT (base and large)** (Devlin et al, 2019).
- **RoBERTa (base and large)** (Liu et al., 2019).
- **CamemBERT** (Martin et al, 2020).
- **BARThez** (Eddine et al., 2020).

The woman worked as a nurse VS the man worked as a nurse

Normatively



Descriptively (in Norway)



Model	Normative	Descriptive
NorBERT	16.23	39.31
NorBERT2	3.17	34.67
NB-BERT	18.55	36.50
NB-BERT_Large	11.35	40.90
BERT_UK	18.05	35.33
BERT_large_UK	13.73	40.43
RoBERTa_base_UK	0.15	34.56
RoBERTa_large_UK	0.00	34.56
BERT_US	17.25	43.29
BERT_Large_US	12.46	48.88
RoBERTa_base_US	0.15	42.81
RoBERTa_Large_US	0.31	42.81
CamemBERT	10.46	34.10
BARThez	6.45	37.08

Table 1: Normative and descriptive occupational bias scores.

Model	Neutral	Female	Male
NorBERT	1.46	22.34	15.50
NorBERT2	0.24	33.57	0.85
NB-BERT	1.46	23.68	11.35
NB-BERT_Large	0.12	33.82	6.95
BERT_UK	1.54	33.02	0.77
BERT_Large_UK	1.23	31.63	7.56
RoBERTa_base_UK	0.00	34.56	0.00
RoBERTa_Large_UK	0.00	34.56	0.00
BERT_US	2.39	39.93	0.95
BERT_Large_US.	1.75	40.09	7.02
RoBERTa_base_US	0.00	42.81	0.00
RoBERTa_Large_US	0.00	42.81	0.00
CamemBERT	0.00	0.00	34.10
BARThez	0.00	0.00	37.08

Table 2: Descriptive bias scores of gender-imbalanced and gender-neutral occupations.