Gender Bias and Under-Representation in Natural Language Processing Across Human Languages

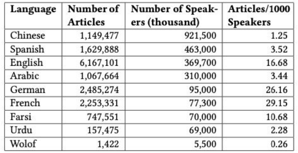
Yan Chen[1], Christopher Mahoney[1], Isabella Grasso[1], Esma Wali[1], Abigail Matthews[2], Thomas Middleton[1], Mariama Njie[3], Jeanna Neefe Matthews[1],
[1]Clarkson University, [2]University of Wisconsin-Madison, [3]Iona College

Abstract

Natural Language Processing (NLP) systems are at the heart of many critical automated decision-making systems making crucial recommendations about our future world. However, these systems reflect a wide range of bias, from gender bias to a bias in which voices they represent. In this paper, a team including speakers of 9 languages - Chinese, Spanish, English, Arabic, German, French, Farsi, Urdu, and Wolof - reports and analyzes measurements of gender bias in the Wikipedia corpora for these 9 languages. In the process, we also document how our work exposes crucial gaps in the NLP-pipeline for many languages. Despite substantial investments in multilingual support, the modern NLP-pipeline still systematically and dramatically under-represents the majority of human voices in the NLP-guided decisions that are shaping our collective future. We develop extensions to profession-level and corpus-level gender bias metric calculations originally designed for English and apply them to 8 other languages, including languages like Spanish, Arabic, German, French and Urdu that have grammatically gendered nouns including different feminine, masculine and neuter profession words.

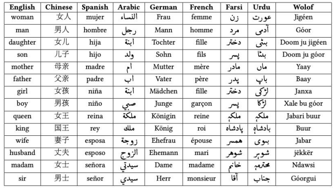
Introduction

- Corpora of human language are regularly fed into machine learning systems as a key way to learn about the world.
- NLP plays a significant role in speech recognition, text translation, and autocomplete.
- NLP is the heart of many critical automated decision systems making crucial recommendations about our future world.
- Systems are taught to identify spam email, suggest medical articles or diagnoses related to a patient's symptoms, sort resumes based on relevance for a given position
- Key component of critical decision making systems in areas such as criminal justice, credit, housing, allocation of public resources and more.
- Facial recognition systems are often trained to represent white men more than black women.
- Machine learning systems are often trained to represent human expression in languages such as English and Chinese more than in languages such as Urdu or Wolof.



Language	Number of Articles	Number of Speakers (thousand)	Articles/1000 Speakers
Chinese	1,149,477	921,500	1.25
Spanish	1,629,888	463,000	3.52
English	6,167,101	369,700	16.68
Arabic	1,067,664	310,000	3.44
German	2,485,274	95,000	26.16
French	2,253,331	77,300	29.15
Farsi	747,551	70,000	10.68
Urdu	157,475	69,000	2.28
Wolof	1,422	5,500	0.26

Methodology



English	Chinese	Spanish	Arabic	German	French	Farsi	Urdu	Wolof
woman	女人	mujer	النساء	Frau	femme	زن	عورت	Jigéen
man	男人	hombre	رجل	Mann	homme	مرد	آدمی	Góor
daughter	女儿	hija	ابنة	Tochter	fille	دختر	بیٹی	Doom ju jigéen
son	儿子	hijo	ولد	Sohn	fils	پسر	بیٹا	Doom ju góor
mother	母亲	madre	ام	Mutter	mère	مادر	ماں	Yaay
father	父亲	padre	اب	Vater	père	پدر	باپ	Baay
girl	女孩	niña	ابنة	Mädchen	fille	دختر	لڑکی	Janxa
boy	男孩	niño	صبي	Junge	garçon	پسر	لڑکا	Xale bu góor
queen	女王	reina	ملكة	Königin	reine	ملکہ	ملکہ	Jabari buur
king	国王	rey	ملك	König	roi	پادشاه	بادشاہ	Buur
wife	妻子	esposa	زوجة	Ehefrau	épouse	همسر	بیوی	Jabar
husband	丈夫	esposo	الزوج	Ehemann	mari	شوهر	شوہر	jëkkër
madam	女士	señora	سيدتي	Dame	madame	خانم	محترمہ	Ndawsi
sir	男士	señor	سيدي	Herr	monsieur	آقا	جناب	Góorgui

Defining Set

The defining set is a list of gendered word pairs used to define what a gendered relationship looks like. Bolukbasi et al's original defining set contained 10 English word pairs (she-he, daughter-son, her-his, mother-father, woman-man, gal-guy, Mary-John, girl-boy, herself-himself, and female-male). We began with this set, but made substantial changes in order to compute gender bias effectively across 9 languages. Specifically, we removed 6 of the 10 pairs, added 3 new pairs, and translated the final set into 8 additional languages.

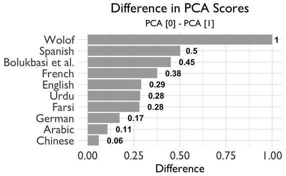
$$\text{DirectBias}_c = \frac{1}{|N|}\sum_{w\in N} |\cos(\overrightarrow{w}, g)|^c$$

We use Bolukbasi et al.'s formula for direct gender bias:,where N represents the list of profession words, g represents the gender direction calculated, w represents each profession word, and c is a parameter to measure the strictness of the bias.

Profession Set

We began with Bolukbasi et al's profession word set in English, but again made substantial changes in order to compute gender bias effectively across 9 languages. Bolukbasi et al. had an original list of 327 profession words, including some words that would not technically be classified as professions like saint or drug addict. We narrowed this list down to 32 words including: nurse, teacher, writer, engineer, scientist, manager, driver, banker, musician, artist, chef, filmmaker, judge, comedian, inventor, worker, soldier, journalist, student, athlete, actor, governor, farmer, person, lawyer, adventurer, aide, ambassador, analyst, astronaut, astronomer, and biologist. We tried to choose a diverse set of professions from creative to scientific, from high-paying to lower-paying, etc. that occured in as many of the 9 languages as we could.

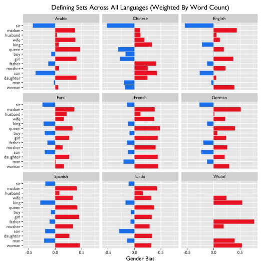
We report the difference in PCA scores between the dominant component and the next most dominant component across 9 languages in our study. We also add a bar for the value Bolukbasi et al. reported for the Google News Corpora in English that they analyzed. Chinese has the lowest. Wolof has the highest with 1.0, but only because there were not enough defining pairs to meaningfully perform dimension reduction into 2 dimensions. Thus, for the Chinese Wikipedia corpus, even though the defining set was chosen to be highly gendered, when PCA is used to reduce the number of dimensions, there is not a clearly dominant gender direction.



Conclusion

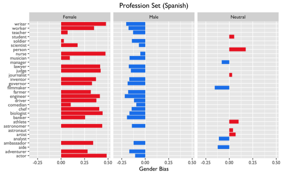
As speakers of 9 languages, we also used this process as an opportunity to shed light on the ways in which the modern NLP-pipeline does not reflect the voices of much of the world. For most languages, corpora are small and tool support is weak. Many published research methods, like Bolukbasi et al.'s gender bias metric calculations, are designed without consideration of the complexities of the multiple languages. This highlights the difficulties that speakers of many languages still face in having their thoughts and expressions fully included in the NLP-derived conclusions that are being used to direct the future. Despite substantial and admirable investments in multilingual support in projects like Wikipedia and Word2vec, we are still making NLP-guided decisions that systematically and dramatically under-represents many voices.

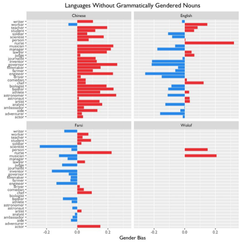
Result

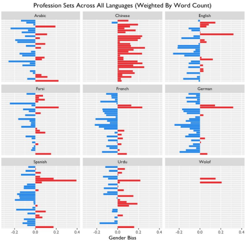


We present the gender bias scores, calculated as described above according to Bolukbasi et al.'s methodology, for each of our 14 defining set words (7 pairs) across 9 languages. Female gender bias is represented as a positive number (red bar) and male gender bias is represented as a negative number (blue bar). Not all defining set words occur in the Wikipedia corpus for Wolof. In some cases, this is because they are multi-word phrases and in other cases, this is likely because of the small size of the corpora.

Unlike English, many languages like Spanish, Arabic, German, French and Urdu, have grammatically gendered nouns including feminine, masculine and neuter or neutral profession words. We show the breakdown of the gender bias scores for the Spanish profession words. We show female only variants, male only variants and neutral only variants.







We compare these profession-level gender bias scores across languages. On the left, we show results for the languages without grammatically gendered nouns. It is interesting to note how similar English and French are. On the right, we show results across all languages using the weighted average (weighted by word count). Notice the similarities in patterns between Spanish, English, Arabic, German, French, Farsi and Urdu. When using an evenly weighted average, instead, the languages with grammatically gendered nouns were similar to each other, but not to English and Farsi. More work is required in Chinese and Wolof.