Gender Bias in Natural Language Processing Across Human Languages

Abigail Matthews[1], Isabella Grasso[2], Christopher Mahoney[2], Yan Chen[2], Esma Wali[2], Thomas Middleton[2], Mariama Njie[3], Jeanna Neefe Matthews[2]
[1]University of Wisconsin-Madison, [2]Clarkson University, [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. Gender bias in NLP has been well studied in English, but has been less studied in other languages. 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. 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 that have grammatically gendered nouns including different feminine, masculine, and neuter profession words. We discuss future work that would benefit immensely from a computational linguistics perspective.

Background/Motivation

- 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.
- Expanding work on quantify gender bias done only in English to other languages

Future Work/ Opportunities for Collaboration

- Experiments with different defining sets, both specific to one languages and the best set across languages
- Experiments with changes in defining sets and profession sets over time
- Approaches to dealing with disambiguation of terms
- Identifying different corpora beyond Wikipedia. Even within one language, it would be interesting to examine collections with different emphasis such as gender of author, different time periods,different genres of text, different country of origin, etc.

Defining Set

English	Chinese	Spanish	Arabic	German	French	Farsi	Urdu	Wolof
woman	女人	mujer	المرأة	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	أم	Mutter	mère	مادر	ماں	Yaay
father	父亲	padre	أب	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	Jabaat buur
king	国王	rey	ملك	König	roi	پادشاہ	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

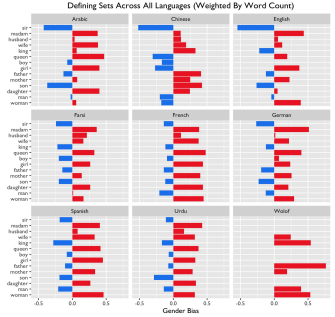
The defining set is a list of gendered word pairs used to define what a gendered relationship looks like. Bolukbasi et al's original defining set contained 10 English word pairs (she-he, daughter-son, her-his, mother-father, woman-man, gal-guy, Mary-John, girl-boy, herself- himself, and female-male). 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.

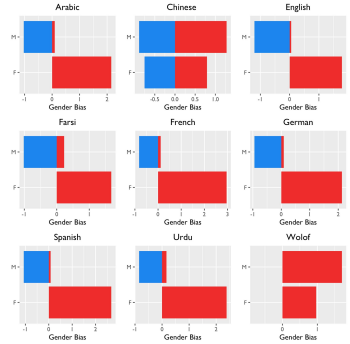
We use Bolukbasi et al.'s formula for direct gender bias:,where N represents the list of profession words, g represents the gender direction calculated, w represents each profession word, and c is a parameter to measure the strictness of the bias.

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

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.

We aggregate the gender bias for all the male words (sir, husband, king, etc.) and all the female words (madam, wife, queen, etc.) This presentation emphasizes several key aspects of the results.





Some key observations:

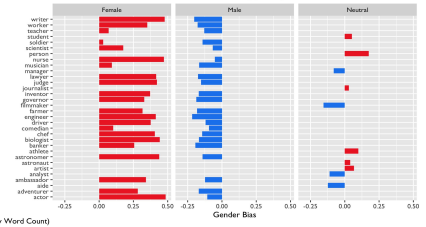
- Spanish, English, Arabic, German, French, Farsi, and Urdu, that the female words are female leaning and that most male words and male leaning as one might expect.
- Interesting exception of husband in all of these languages and also man in Farsi.
- Wolof has troubles primarily due to the small size of the corpus.
- Chinese has trouble due to difficulty finding a dominant PCA that isolates the gender direction (more info in the paper on that).

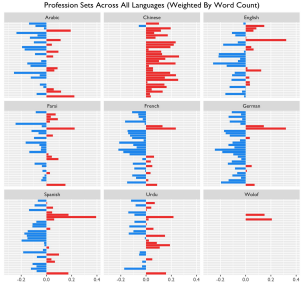
Profession Set

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.

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





We use Pearson's correlation for cluster analysis to examine 7 languages, omitting Chinese and Wolof because of problems with PCA and corpora size respectively. This exploratory analysis provokes a number of questions for future work including: How do linguistics inform the bias outputs (e.g. If English is a mixture of Germanic and Latin languages, is that why it is clustered with those languages even though it is not gendered?) and How does a language being inherently gendered affect the resulting bias of an NLP model in that language?

We compare these profession-level gender bias scores across all 9 languages using the weighted average (weighted by word count). Notice the similarities in patterns between Spanish, English, Arabic, German, French, Farsi and Urdu.For example, nurse is often highly female biased and engineer is often highly male biased.