Is Machine Learning Speaking My Language?

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

Jeanna Matthews
Clarkson University

February 18 2021

Clarkson UNIVERSITY defy convention™

A bit about me

- PhD, UC Berkeley, 1994-1999
  - Network of Workstations project
- Professor
  - Clarkson University, 2000- 2001, 2003 - 2006
  - Cornell University, 2002- 2003
  - Clarkson University, 2006 –present
- Sabbaticals
  - VMware, Boston, 2008 – 2009
  - Data and Society New York, 2017-2019
- Industrial collaborations
  - Intel, EMC, Greenplum and others
- ACM
  - SIGOPs chair, ACM Council, Distinguished Speaker
  - US-TPC/ US-ACM

Data&Society

Association for Computing Machinery

Advancing Computing as a Science & Profession

A bit about you

- Years in school?
- Majors?
- Familiarity with natural language processing? Machine learning in general?

Natural Language Processing

- Programs that take in corpora of human text or speech
- Used for many important tasks
  - identify spam email
  - Suggest medical articles or diagnoses related to a patient's symptoms
  - Sort resumes based on relevance for a given position
  - Translation
  - And many other tasks that form key components of critical decision making systems in areas such as criminal justice, credit, housing, allocation of public resources
- Learn from humans to help make key decisions that guide our collective future

Typical NLP Pipeline

- Gathering corpora
- Processing them into text format
- Identifying key language elements
- Building trained models
- Using trained models to answer predictive questions

Types of NLP Tools and Resources

- Corpora of text (e.g. Wikipedia)
- Tools that generate text (e.g. OCR)
- Labeled corpora
- Tools that build models from text (e.g. Word2Vec, BERT)
- Pre-trained models
- Research tools and results
  - E.g. Tools that quantify gender bias in a corpora

Huge Disparities in NLP Support Across Languages

- Support varies hugely across human languages
  - E.g. Over 7000 languages spoken in the world, but only ~300 have Wikipedia corpora
- Some languages, especially English, have all the tools and support
  - Chinese also quite well supported
- Lack in the early stages contributes to lack in later stages
  - Wikipedia corpora often used to train models and tolos
  - E.g. Only 2 languages (English and Chinese) have pretrained BERT models

Wikipedia

- Over 7000 languages spoken in the world
  - ~317 have Wikipedia corpora
  - 2 have pretrained models for BERT
- Notice does not track with number of speakers

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

NLP Tools

Language	BERT	Word2Vec	NLTK	Wikipedia2Vec
Chinese	✓	✓	✓	✓
Spanish	✓	✓	✓	✓
English	✓	✓	✓	✓
Arabic	✓	✓	✓	✓
German	✓	✓	✓	✓
French	✓	✓	✓	✓
Farsi	✓	✓		
Urdu	✓		✓	
Wolof				

- But a check mark for support can hide caveats

Even when supported..

- We found:
  - Lack of testing
  - Higher error rates for tools/surprising errors in some languages

- Lack of representation in the early stages of the NLP pipeline (e.g. representation in Wikipedia) is further magnified throughout the NLP-tool chain, culminating in reliance on easy-to-use pre-trained models that effectively prevents all but the most highly resourced teams from including diverse voices.

Beyond corpora and tools, research…

- NLP research for the most part focuses on English or perhaps Chinese
- One concrete example is a paper we love from NIPS/NeurIPs 2016

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

---

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Results from Bolukbasi et al.

- Built a Word2Vec model on a corpora of Google News in English
- Each word represented as a vector
- Allows you to play word games like “man is woman as king is to what?”
- Gender bias in the results
- Developed a system for quantifying this gender bias and even “debiasing”
- But very focused on English

Defining Set and Profession Set

- Defining Set: 10 word pairs that represent a gender binary
  - She-he, daughter-son, her-his, mother-father, woman-man, gal-guy, Mary-John, girl-boy, herself-himself, and female-male
- Profession Set:
  - 327 profession words like doctor, nurse, teacher, etc
  - Including some words that would not technically be classified as professions like saint or drug addict
  - Ideally would not reflect strong gender bias but in practice sometimes do

Challenges in other languages

- Unlike English, many languages like Spanish, Arabic, German, French and Urdu, have grammatically gendered nouns including feminine, masculine and neuter or neutral profession words
- Spoken languages like Wolof might have a variety of ways of being represented in text

Troubles with defining set words

- She-he: Same words in some languages like Wolof, Farsi, and Urdu.
- Her-his: In some languages like French and Spanish, it depends on the gender of the object and not the person to which the object belongs. In German without context, “ihrer” could mean her, your, theirs or yours
- Gal-guy: Same words in some languages like Wolof. There are no translations for these words in some languages like Urdu and Arabic.
- Etc….

Defining Set: Modified and Translated

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	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

Quantifying Gender Bias: Step 1

- Calculate the center of the vectors for each definitional pair.
  - E.g. to calculate the center of the definitional pair woman/man, average the vector for "woman" with the vector for "man". Then, calculate the distance of each word in the definitional pair from the center by subtracting the center from each word in the pair (e.g. "woman" - center).

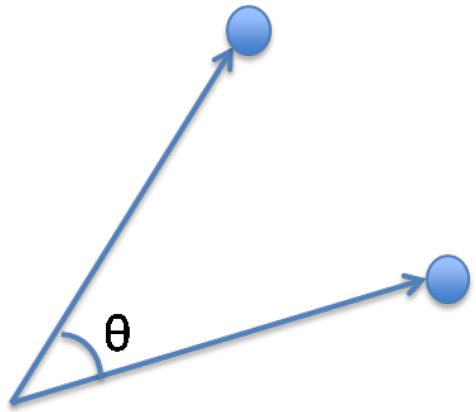
Quantifying Gender Bias: Step 2

- Apply Principal Component Analysis (PCA) to the matrix of these distances.
  - PCA is an approach that compresses multiple dimensions into fewer dimensions, ideally in a way that the information within the original data is not lost. Usually the number of reduced dimensions is 1-3 as it allows for easier visualization of a dataset.
- Because the defining set pairs were chosen to the highly gendered, we expect this dimension to be related primarily to gender and therefore call it the gender direction or the g direction.

Quantifying Gender Bias: Step 3

- Use cosine similarity as a measure of similarity between each word, w, and the g direction
- For each word, Gender Bias = The cosine of vectors w and g: cos(w,g) = (w . g) / (||w|| . ||g||).

$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$



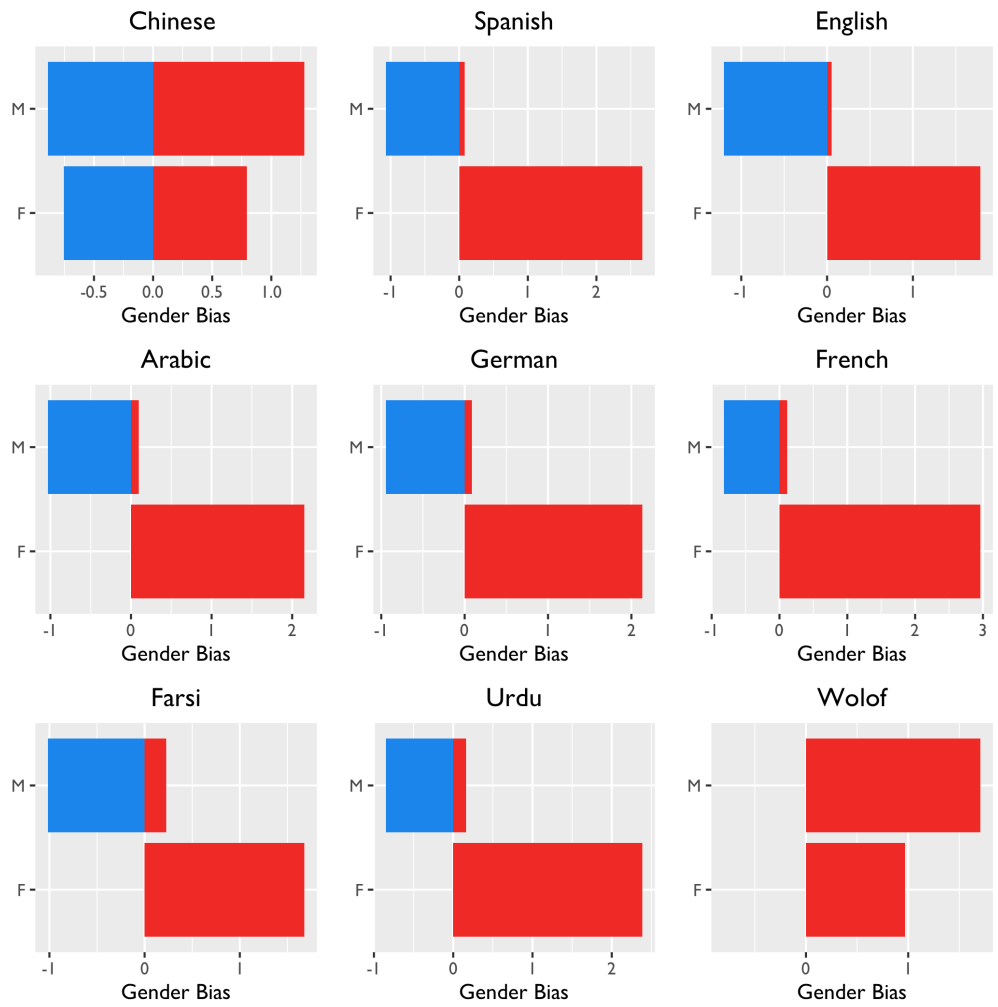
Quantifying Gender Bias: Step 4

- For a corpora or other collection of words we can average the gender bias of all the words found there.
- Bolukbasi et al used

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

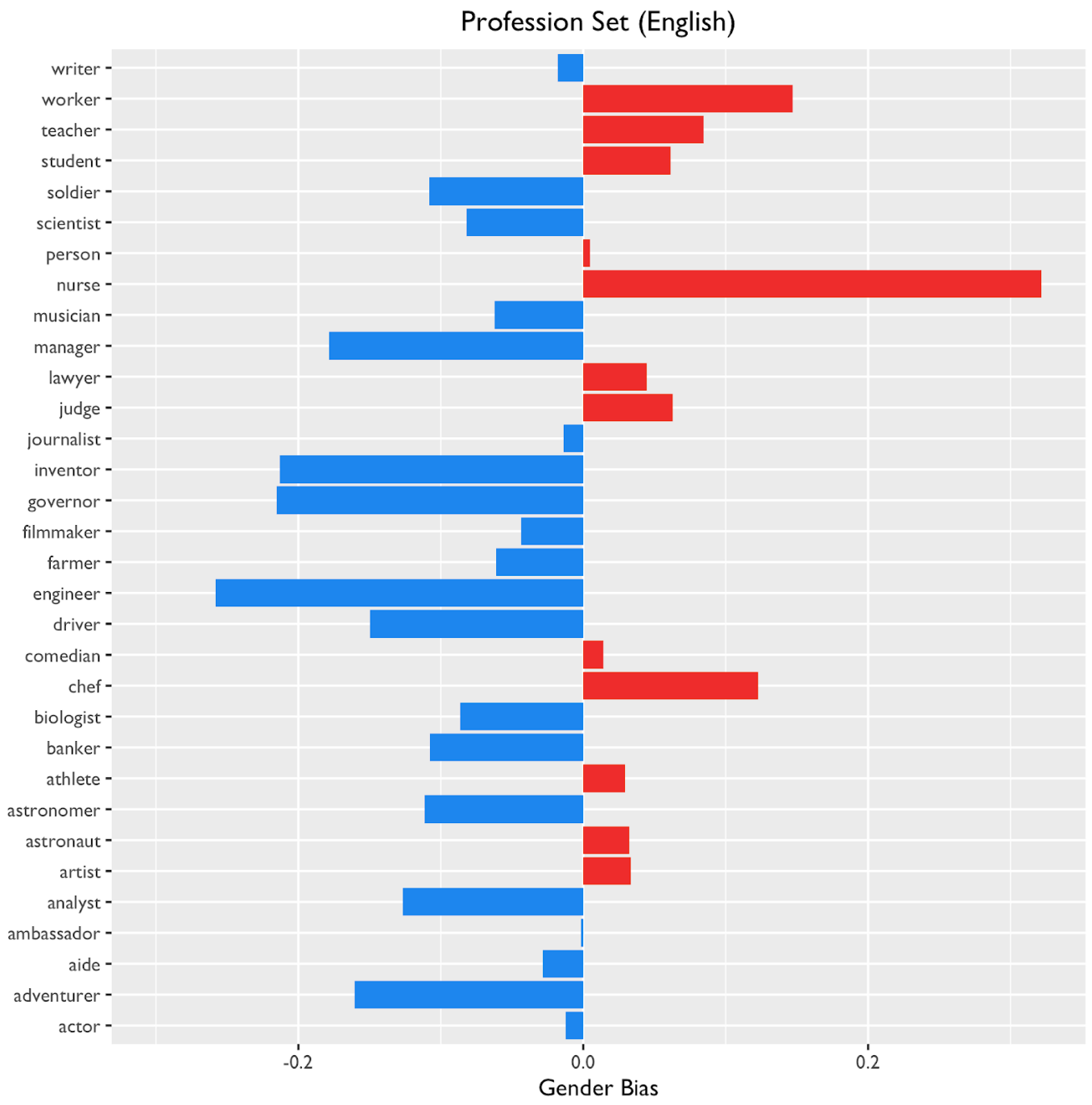
- We found weighting by the word count to be better.

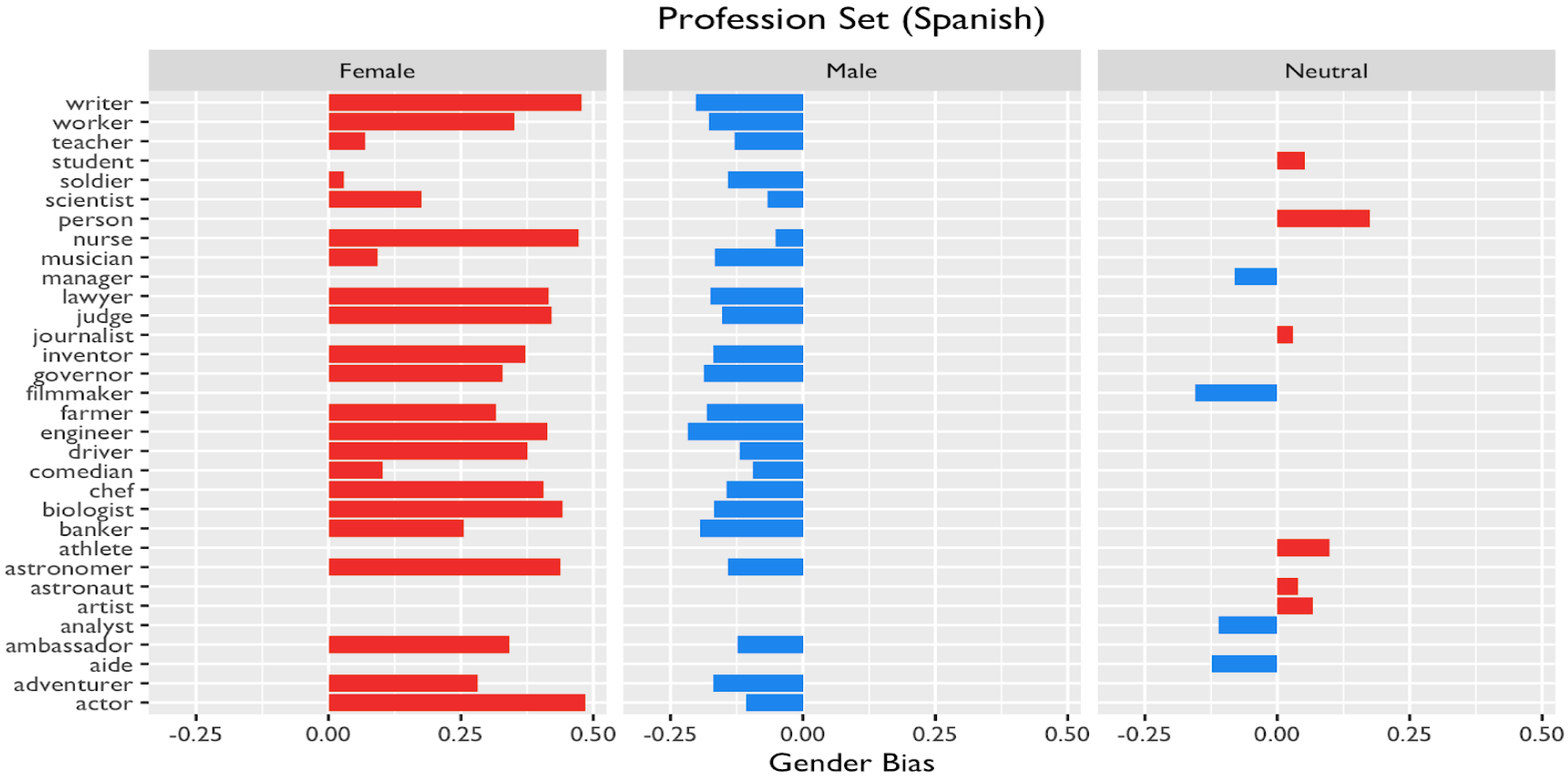

Defining Sets Across All Languages (Weighted By Word Count)
Arabic
Chinese
English
Farsi
French
German
Spanish
Urdu
Wolof
sir
madam
husband
wife
king
queen
boy
girl
father
mother
son
daughter
man
woman
-0.5
0.0
0.5
Gender Bias

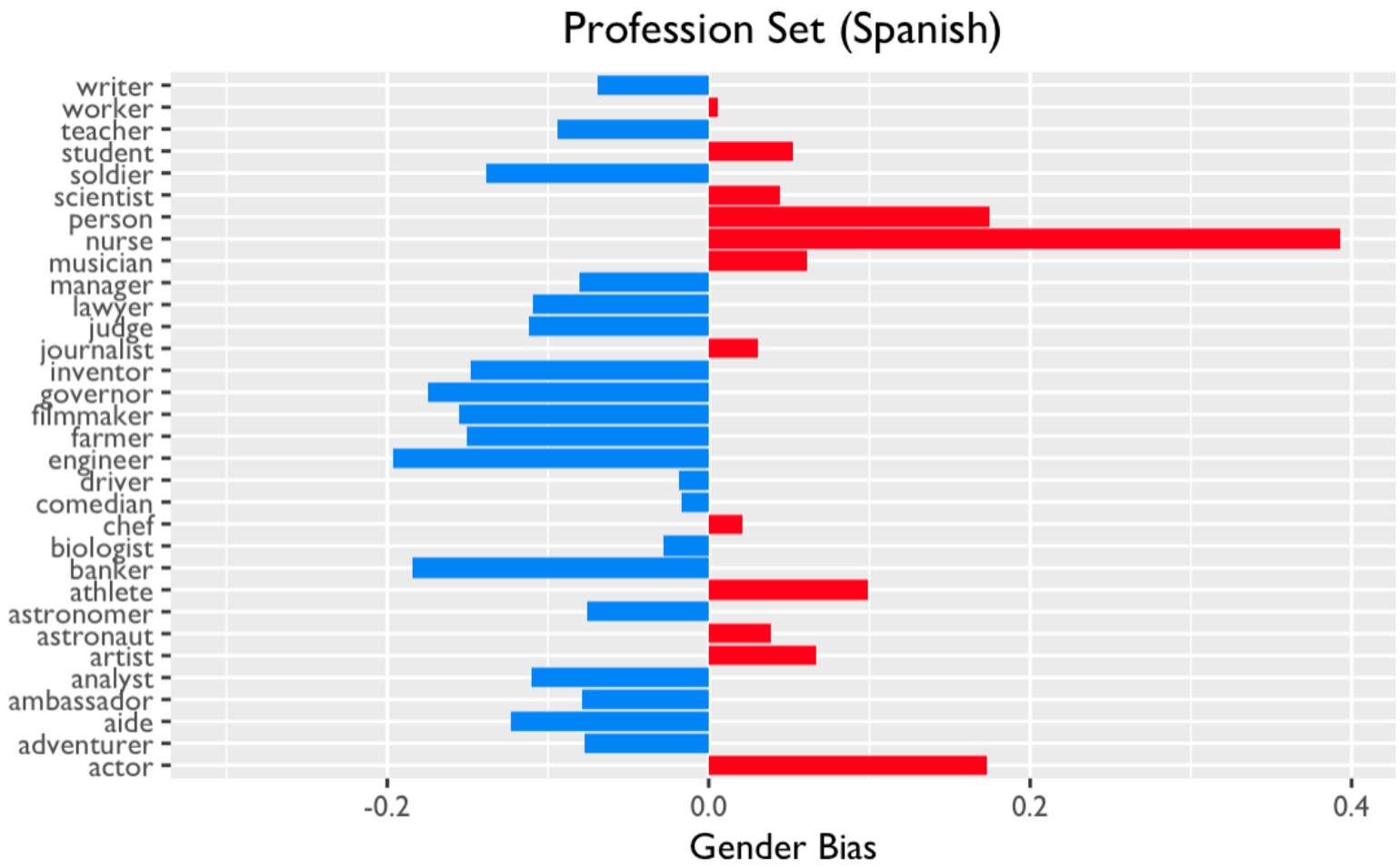

Chinese
M
F
-0.5
0.0
0.5
1.0
Gender Bias
Spanish
M
F
-1
0
1
2
Gender Bias
English
M
F
-1
0
1
Gender Bias
Arabic
M
F
-1
0
1
2
Gender Bias
German
M
F
-1
0
1
2
Gender Bias
French
M
F
-1
0
1
2
3
Gender Bias
Farsi
M
F
-1
0
1
Gender Bias
Urdu
M
F
-1
0
1
2
Gender Bias
Wolof
M
F
0
1
Gender Bias

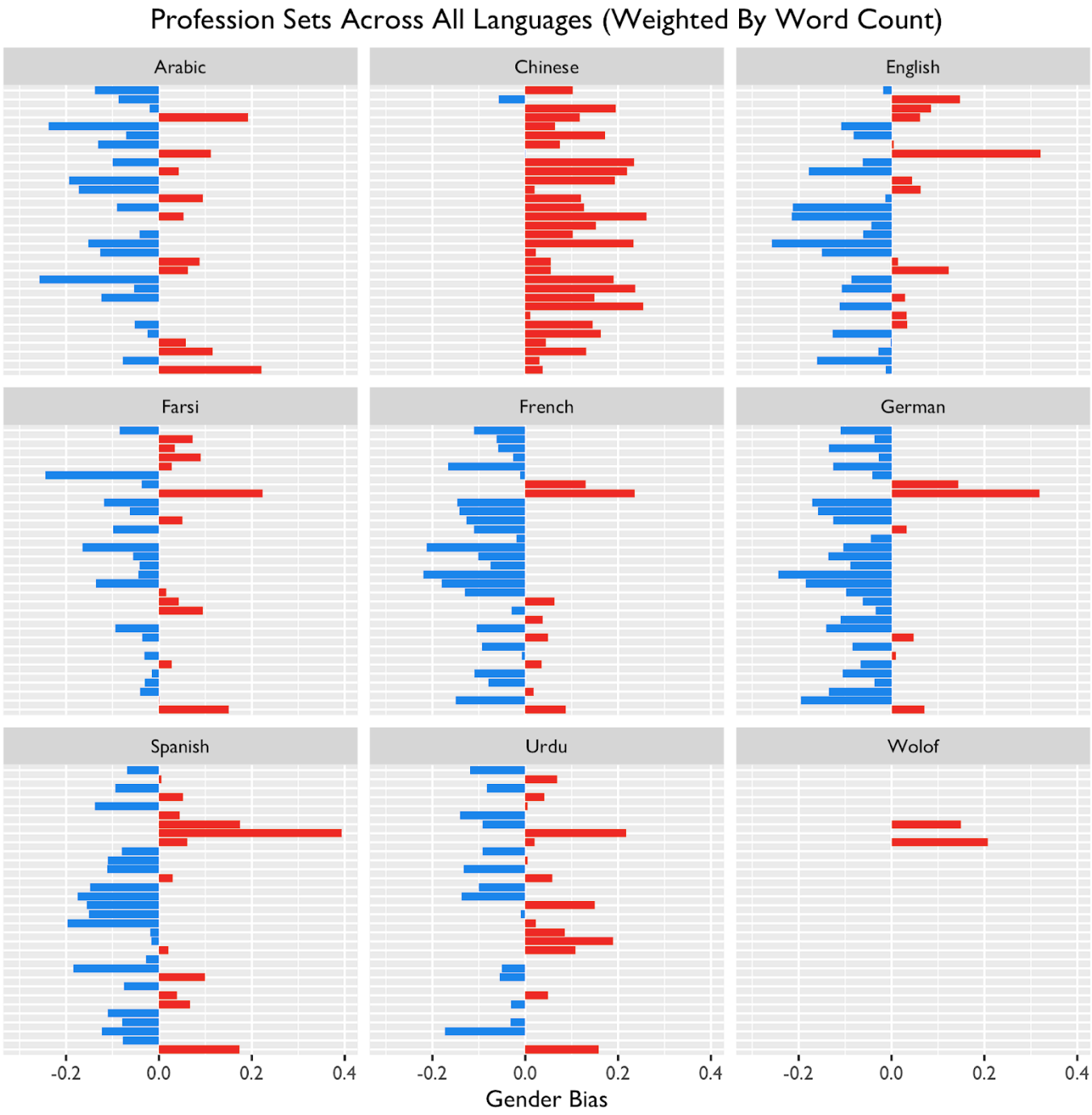
Profession Set

- Narrowed from 317 to 32
- Big differences in how often some profession words used in Wikipedia corpora across languages
- Many languages have male, female and sometimes neutral variants of profession words
  - Examples in Spanish: escritor/ escitora, periodista


Profession Set (English)
writer
worker
teacher
student
soldier
scientist
person
nurse
musician
manager
lawyer
judge
journalist
inventor
governor
filmmaker
farmer
engineer
driver
comedian
chef
biologist
banker
athlete
astronomer
astronaut
artist
analyst
ambassador
aide
adventurer
actor
-0.2
0.0
0.2
Gender Bias


Profession Set (Spanish)
Female
Male
Neutral
writer
worker
teacher
student
soldier
scientist
person
nurse
musician
manager
lawyer
judge
journalist
inventor
governor
filmmaker
farmer
engineer
driver
comedian
chef
biologist
banker
athlete
astronomer
astronaut
artist
analyst
ambassador
aide
adventurer
actor
-0.25
0.00
0.25
0.50
Gender Bias


Profession Set (Spanish)
writer
worker
teacher
student
soldier
scientist
person
nurse
musician
manager
lawyer
judge
journalist
inventor
governor
filmmaker
farmer
engineer
driver
comedian
chef
biologist
banker
athlete
astronomer
astronaut
artist
analyst
ambassador
aide
adventurer
actor
-0.2
0.0
0.2
0.4
Gender Bias


Profession Sets Across All Languages (Weighted By Word Count)
Arabic
Chinese
English
Farsi
French
German
Spanish
Urdu
Wolof
-0.2
0.0
0.2
0.4
Gender Bias

Lessons learned

- Big decisions about our lives made with NLP-guided software
  - What will it take to make those decisions fair, accountable, transparent? To make a future that works for individuals?
- Lots of work to do to make sure that our NLP-guided future reflects many languages and voices
- Biases we need to watch for! and
  - Gender bias in corpora
  - Over-representation of a digital text (not classics) and a few languages
- Reusing easy ingredients without much sense of the quality of those ingredients

Questions?

jnm@clarkson.edu

http://www.clarkson.edu/~jnm

@jeanna_matthews