Unsupervised Discovery of Gendered Language

ACL, 2019-07-29

Alexander Hoyle
University of Maryland[1]

Lawrence Wolf-Sonkin
Johns Hopkins University

Microsoft Research
Hanna Wallach
Microsoft Research

Isabelle Augenstein
University of Copenhagen

Ryan Cotterell
University of Cambridge[2]

[1] *Work undertaken at University College London* [2]*Moving to ETH Zurich.*

Background

Word choice is influenced by gender

Both the gender of the *speaker*

- Women more likely to use pronouns, emotion terms on Twitter; men use more curse words, proper nouns [1]

And of the *referent*

- Female infants rated as more *delicate* whereas male infants are *hardier* [2]

[1] Bamman et al. 2014

[2] Rubin et al 1974.

Gendered differences in language use can be...

...innocuous

“[H]e made a sign to a bearded man” [3]

...loaded

“[S]he moved from one posture to another ... growing more and more hysterical” [4]

[3] Dumas, A. 1901. Vaninka.

[4] Austen, J. 1811. Sense and Sensibility.

Corpus studies reveal gender stereotypes

"While men are evaluated in terms of their function and status in society, a woman is evaluated [...] in terms of her appearance and sexuality."[5]

"Boys are [...] energetic, playful, curious; [...] girls [...] are represented [...] with a focus on bodily appearance."[6]

[5] Norberg, C. 2016. Naughty Boys and Sexy Girls: The Representation of Young Individuals in a Web-Based Corpus of English

[6] Caldas-Coulthard, C., and Moon, R. 2010. Curvey, hunky, kinky: using corpora as tools for critical analysis.

Sociolinguistic approach uses gendered noun pairs

"man" →

...just what a young man ought to be...

...a single man in possession of a good fortune...

...most disagreeable man in the world...

"woman" →

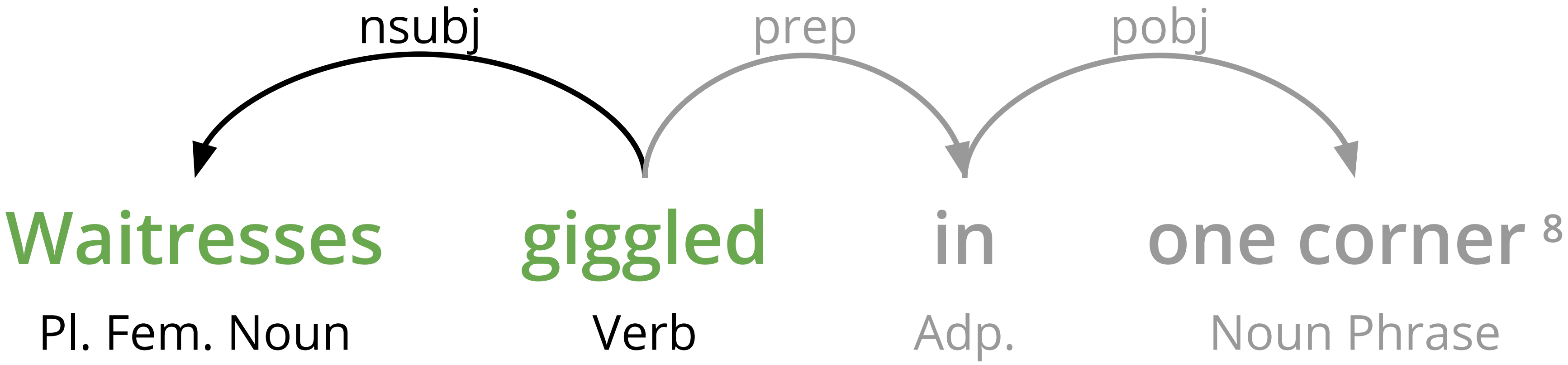
...a very good kind of woman...

...a sensible, intelligent young woman...

...I dare say she is a very agreeable woman...[7]

[7] *All quotes from* Austen, J. 1813. Pride and Prejudice.

Measure differences in syntactic collocations



nsubj

prep

pobj

Waitresses giggled in one corner[8]

Pl. Fem. Noun

Verb

Adp.

Noun Phrase

[8] *Paraphrase of* Orczy, B. 1908. The Old Man in the Corner.

This talk: solving issues in existing approach

Cannot compare across word pairs

Featurize gendered nouns, using multiple pairs

Some differences can be benign

Jointly model sentiment of attached words

Analysis of relative differences is qualitative

Make quantitative evaluation of differences

A teaser: stark differences that align with intuition

👨		👩
Hostile Violent Abusive Brutal	amod	Helpless Disagreeable Unmarried Widowed
Flourish Kill	nsubj	Giggle Gossip
Praise Kill	dobj	Eye Woo

Model

Model: a joint representation of nouns, adjectives or verbs, and sentiment

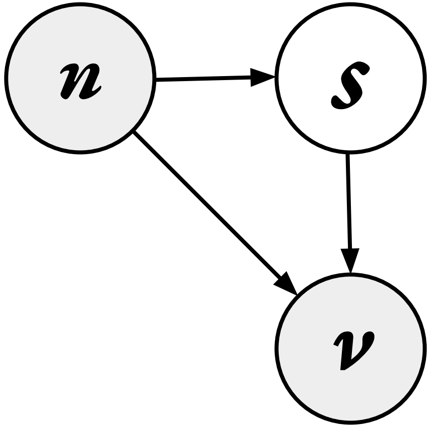
$$p(\, v, n, s \,) = p(\, v \mid n, s \,)\, p(\, s \mid n \,)\, p(\, n \,)$$

Corpus is that of Goldberg and Orwant (2013)

~3.5 million books

~11 billion words

Years 1900-2008



Components: a noun vector of lexical features

$$p(v, n, s) = \underline{p(v \mid n, s)\; p(s \mid n}\,)\; p(n)$$

$n \in \mathscr{G}$

$f_n \in \{0, 1\}^T$

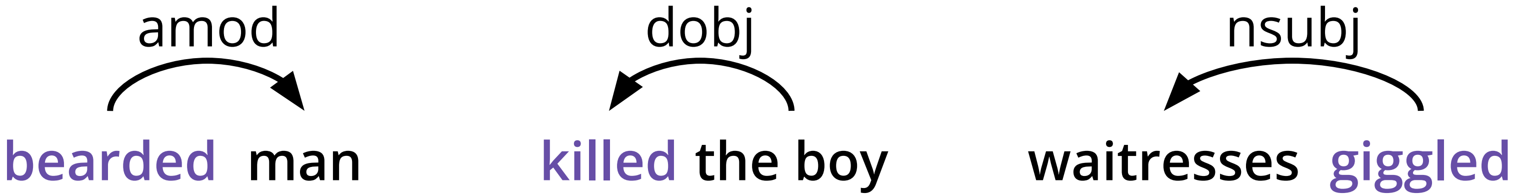
Waitresses → [WAITER, FEM, PL] → [..., *1*, *1*]

Waiter → [WAITER, MASC, S] → [..., *0*, *0*]

Components: neighbors and categorical sentiment

$$p(v, n, s) = \underline{p(v \mid n, s)\, p(s \mid n}\,)\, p(n)$$

$v \in \mathscr{V}$



amod

bearded man

dobj

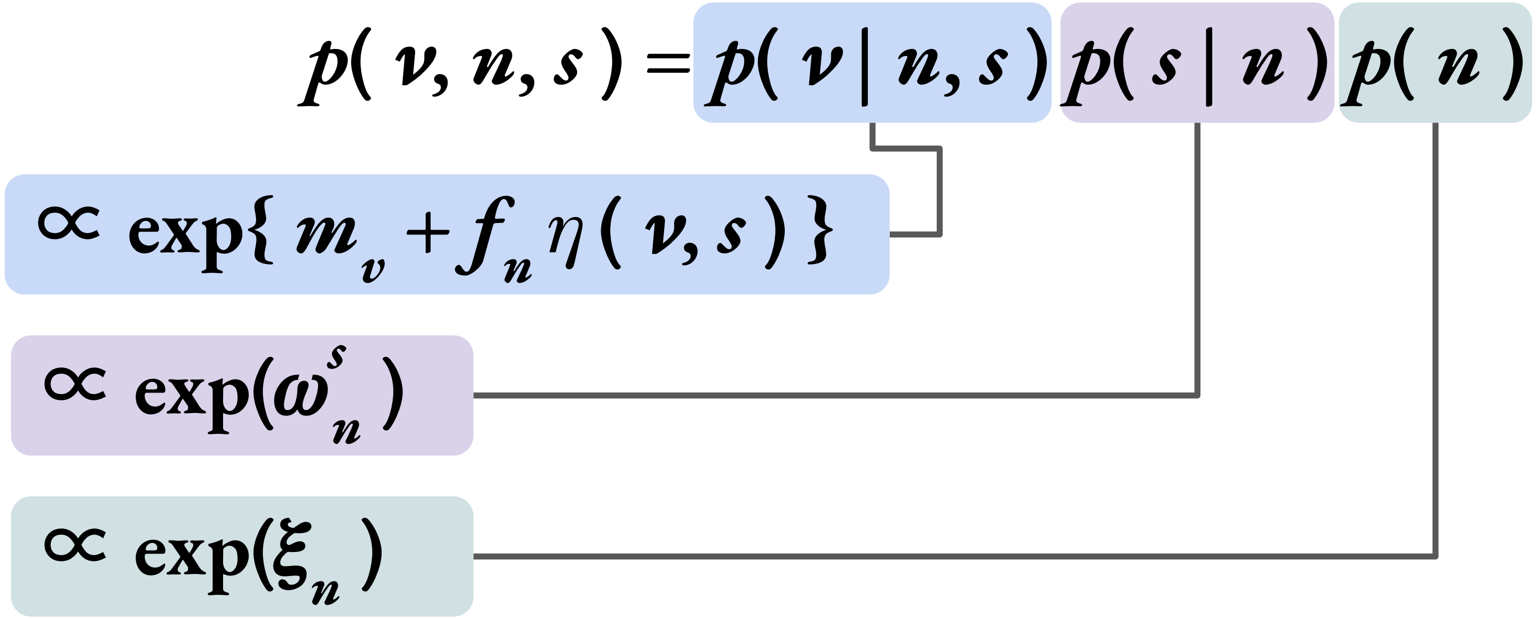
killed the boy

nsubj

waitresses giggled

$s \in \mathscr{S} = \{\text{POS, NEG, NEU}\}$

Probabilities are parameterized separately



$$p(\nu, n, s) = p(\nu \mid n, s)\, p(s \mid n)\, p(n)$$

$$\propto \exp\{ m_v + f_n \eta(\nu, s) \}$$

$$\propto \exp(\omega_n^s)$$

$$\propto \exp(\xi_n)$$

Log-linear model estimates neighbor probability

$$p(\nu \mid n, s) \propto \exp\{ \underbrace{m_v}_{\textit{Fixed Background Distribution}} + \underbrace{f_g^\top \eta_g(\nu, s) + f_{pl}^\top \eta_{pl}(\nu) + f_l^\top \eta_l(\nu)}_{\textit{Learned Deviation Terms}} \}$$

$m_{\text{CUTE}} \in \mathbb{R}$

$$\begin{array}{c} [\ \ldots,\ -9.5\ ,\ \ldots\ ] \\ \text{CUT, CUTE, CYCLIC} \end{array}$$

$\eta_g(\text{CUTE}, s) \in \mathbb{R}^T$

$$\begin{array}{c|cc} & \text{MASC} & \text{FEM} \\ \hline \text{POS} & 1.1, & 3.2 \\ \text{NEG} & -2.6, & 0.9 \\ \text{NEU} & -3.5, & 1.1 \end{array}$$

$\eta_l(\text{CUTE}) \in \mathbb{R}^T$

$$\begin{array}{c|c} \text{BOY} & 0.6 \\ \text{KING} & -6.8 \\ & \ldots \end{array}$$

Implication: obtain neighbors that modify nouns

$$\tau(\nu) \propto \exp\{\, f_{\text{FEM}}^{\top}\, \eta(\nu, \text{POS}) \,\}$$

m_ν		
-9.5	CUTE	$\begin{array}{ccc} & \text{MASC} & \text{FEM} \\ \text{POS} & [\;1.1, & 3.2\;] \end{array}$
-7.6	UGLY	$\begin{array}{ccc} & \text{MASC} & \text{FEM} \\ \text{POS} & [-4.6, & -0.7\;] \end{array}$
-6.1	INTELLIGENT	$\begin{array}{ccc} & \text{MASC} & \text{FEM} \\ \text{POS} & [\;1.1, & 0.6\;] \end{array}$

Problem: corpus does not label sentiment

$$p(\nu, n) = \sum_{s \in \mathcal{S}} p(\nu \mid n, s)\, p(s \mid n)\, p(n)$$

Objective:

$$\min_{\eta, \omega, \xi} \sum_{n \in \mathcal{G}} \sum_{\nu \in \mathcal{V}} \hat{p}(\nu, n) \log(p(\nu, n))$$

Solution: posterior regularization

$$p(s \mid \nu) = \sum_{n \in \mathcal{G}} p(\nu \mid n, s)\, p(s \mid n)\, p(n) \frac{1}{p(\nu)}$$

Objective:

$$\min_{\eta, \omega, \xi} \sum_{n \in \mathcal{G}} \sum_{\nu \in \mathcal{V}} \hat{p}(\nu, n) \log(p(\nu, n))$$

$$+ \beta\, \mathbf{KL}(q(s \mid \nu) \,||\, p(s \mid \nu))$$

$$+ \alpha ||\eta||_1$$

$q(s \mid \text{CUTE})$[8]	
Pos	0.68
Neg	0.14
Neu	0.17

[8] Hoyle et al, 2019

Results

Topics: 200 largest deviation terms for each gender-sentiment pair

$$\boldsymbol{\tau}(\boldsymbol{\nu}) \propto \exp\{\boldsymbol{f}_{\mathrm{FEM}}^{\top}\boldsymbol{\eta}(\boldsymbol{\nu}, \mathrm{POS})\}$$

Adjective Super-senses

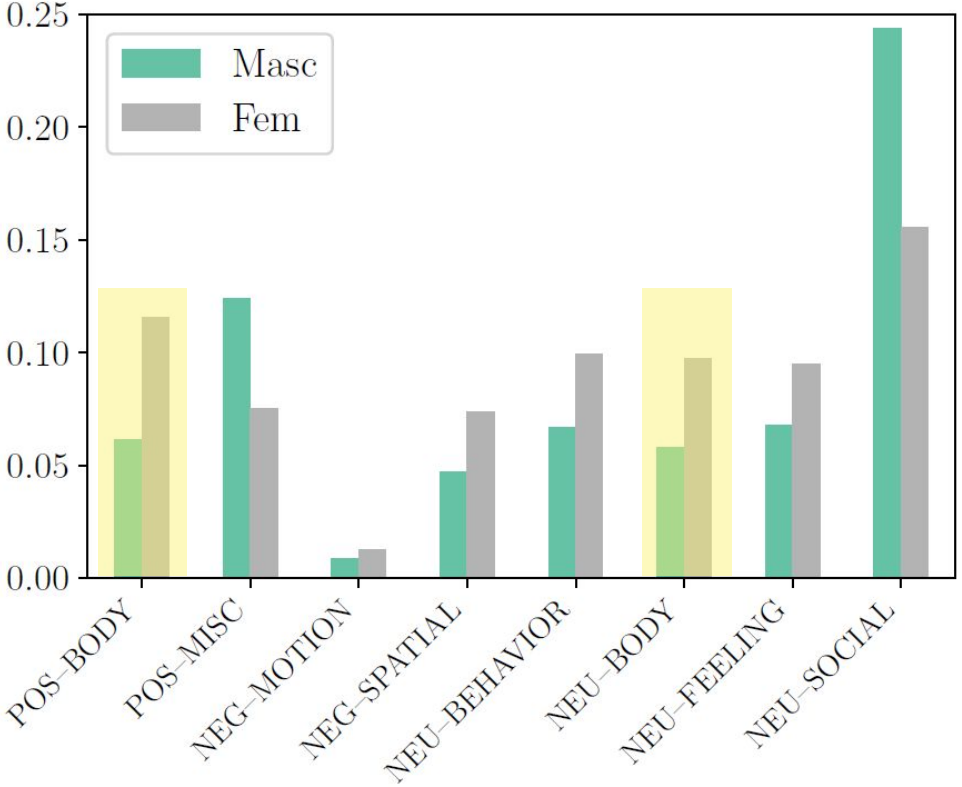
Verb Super-senses

Human Evaluation

Female bodies receive disproportionate attention

“Cute”[9]

BODY	0.78
FEELING	0.05
BEHAVIOR	0.04
SUBSTANCE	0.03
SOCIAL	0.02



[9] Tsvetkov et al, 2014

Positive “BODY” Adjectives

Fabulous
Chic
Sturdy
Manly

👍

Beautiful
Pretty
Lovely
Attractive
Gorgeous
Cute
Sexy
Topless
Blond
…

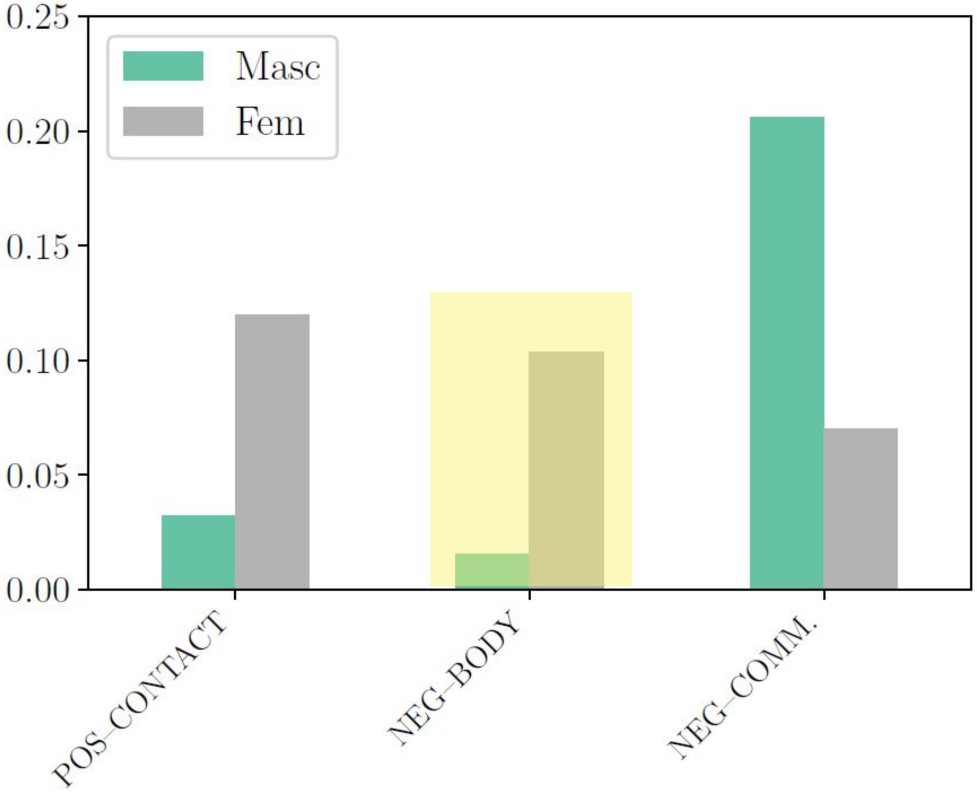
Negative “BEHAVIOR” Adjectives

Hostile
Rough
Abusive
Arrogant
Insane

👎

Shameless
Unprofessional
Crass
Bitchy
Crazy

“BODY” also a more likely NSUBJ verb category



“BODY” & “CONTACT” NSUBJ **Verbs**

Strike
Kill
Destroy
Violate
Choke

👎

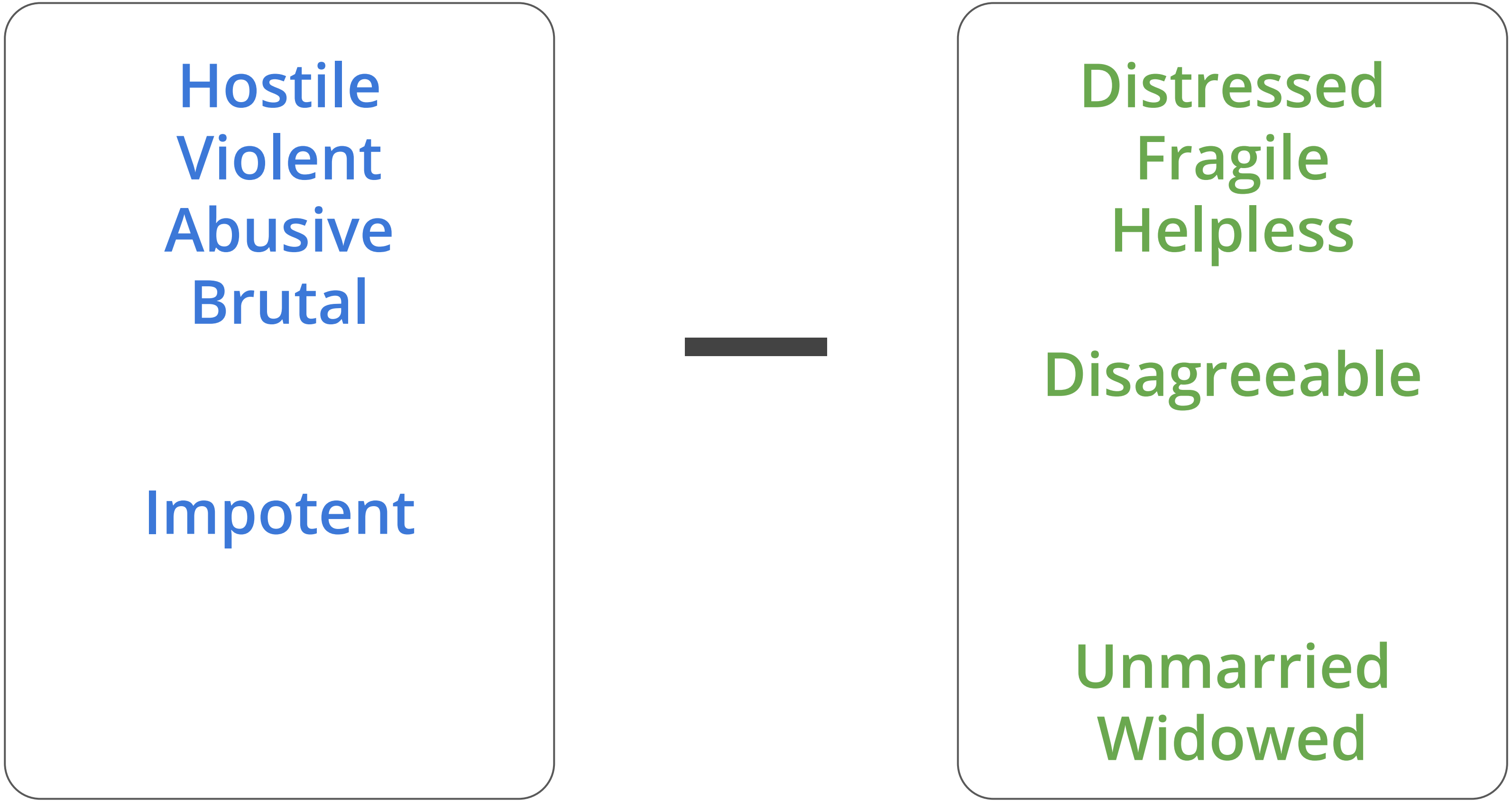
Weep
Cry
Frown
Gasp
Wreck

Embrace
Grin
Seize
Act
Force

👍 😑

Kiss
Attract
Wave
Gush
Dress

Negative Adjectives



Hostile
Violent
Abusive
Brutal

Impotent

−

Distressed
Fragile
Helpless

Disagreeable

Unmarried
Widowed

Verbs where Noun is Subject

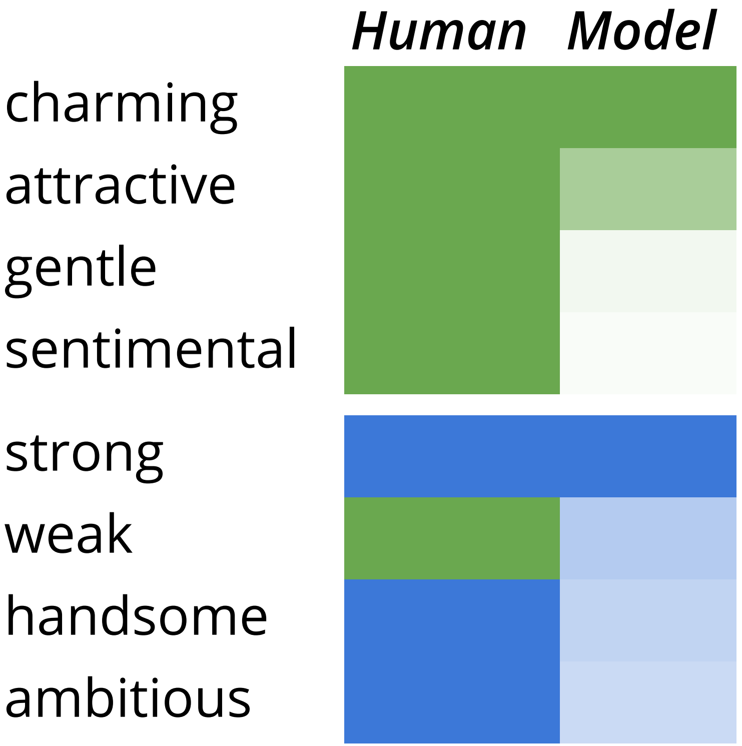
Succeed Flourish Protect Rescue	+	Giggle Kiss Smile Marry
Murder Fight Kill Threaten	−	Gossip Complain Weep Scream

Verbs where Noun is Object

Praise Reward Glorify Honor	+	Eye Escort Woo Protect
Mock Bully Kill Murder	−	Shame Forbid Drown Persecute

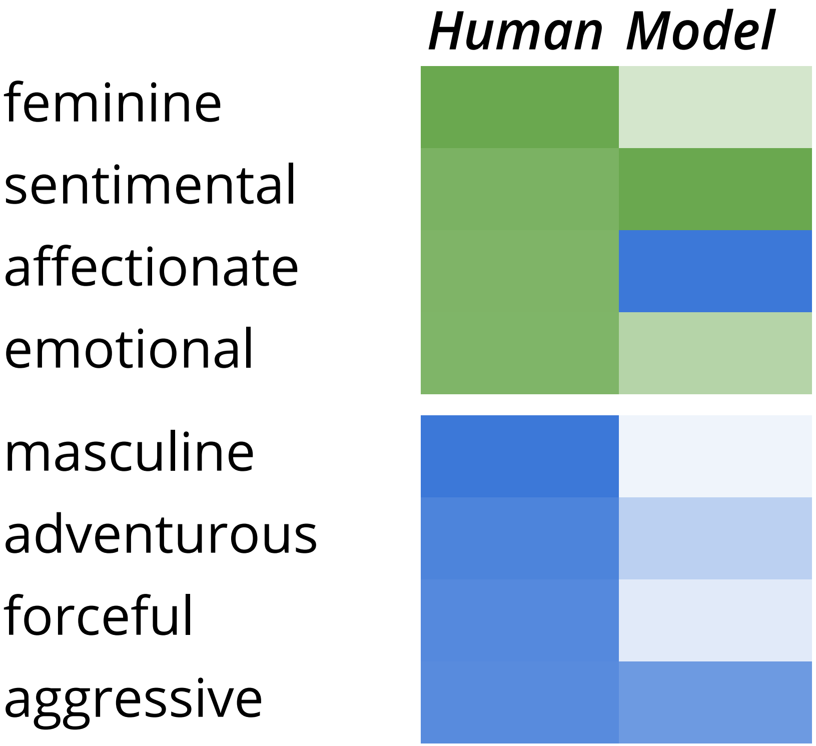
Correlation with human judgements

Williams and Bennet, 1975



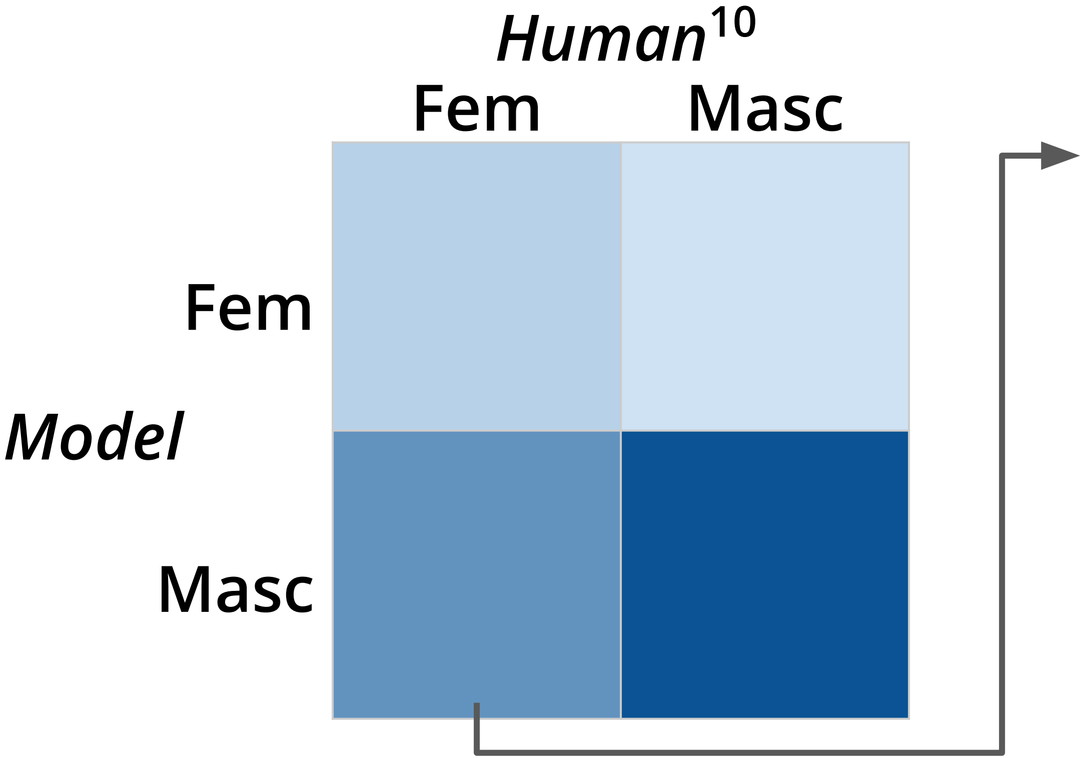
Spearman's ρ *0.59*

Williams and Best, 1977 & 1990



Spearman's ρ *0.33*

Male adjectives align with human judgements



Adjectives Misclassified as Masculine

Effeminate
Submissive
Cowardly
Weak
Timid

[10] Williams and Best, 1977 & 1990

Caveats

Ignore speaker & source (e.g., fiction or nonfiction)

Language changes over time, in particular that relating to gender[11]

Reporting bias ("Black sheep"[12])

Limited to binary gender

[11] Underwood et al. (2018)

[12] Meg Mitchell