Unsupervised Discovery of Gendered Language

ACL, 2019-07-29

Alexander Hoyle  
University of Maryland

Lawrence Wolf-Sonkin  
Johns Hopkins University

Isabelle Augenstein  
University of Copenhagen

Ryan Cotterell  
University of Cambridge

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”

...loaded

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

---

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.”

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

---

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 All quotes from Austen, J. 1813. Pride and Prejudice.
Measure differences in syntactic collocations

Paraphrase of Orczy, B. 1908. The Old Man in the Corner.

Waitresses
Pl. Fem. Noun
giggled
Verb
in
Adp.
one corner
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

<table>
<thead>
<tr>
<th>Hostile</th>
<th>Violent</th>
<th>Abusive</th>
<th>Brutal</th>
<th>amod</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flourish</td>
<td>Kill</td>
<td></td>
<td></td>
<td>nsubj</td>
</tr>
<tr>
<td>Praise</td>
<td>Kill</td>
<td></td>
<td></td>
<td>dobj</td>
</tr>
<tr>
<td>Helpless</td>
<td>Disagreeable</td>
<td>Unmarried</td>
<td>Widowed</td>
<td></td>
</tr>
<tr>
<td>Giggle</td>
<td>Gossip</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eye</td>
<td>Woo</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Model
Model: a joint representation of nouns, adjectives or verbs, and sentiment

\[ p(\nu, n, s) = p(\nu | n, s) p(s | n) p(n) \]

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

\[ p(\nu, n, s) = p(\nu|n, s)p(s|n)p(n) \]

\[ n \in \mathcal{G} \quad f_n \in \{0, 1\}^T \]

\[
\begin{align*}
\text{Waitresses} & \quad \rightarrow \quad [\text{WAITER, FEM, PL}] & \quad \rightarrow \quad [\ldots, 1, 1] \\
\text{Waiter} & \quad \rightarrow \quad [\text{WAITER, MASC, S}] & \quad \rightarrow \quad [\ldots, 0, 0]
\end{align*}
\]
Components: neighbors and categorical sentiment

\[ p(\nu, n, s) = p(\nu | n, s) p(s | n) p(n) \]

\( \nu \in \mathcal{V} \)

\( s \in \mathcal{P} = \{\text{POS, NEG, NEU}\} \)

- bearded man
- killed the boy
- waitresses giggled
Probabilities are parameterized separately

\[ p(\nu, n, s) = p(\nu | n, s) \cdot p(s | n) \cdot p(n) \]

\[ \propto \exp\{m_\nu + f_n \eta(\nu, s)\} \]

\[ \propto \exp(\omega^s_n) \]

\[ \propto \exp(\xi_n) \]
Log-linear model estimates neighbor probability

\[ p(\nu | n, s) \propto \exp\{m_c + \eta_g(\nu, s) + \eta_{pl}(\nu) + \eta_l(\nu)\} \]

Fixed Background Distribution

\[ m_{\text{CUTE}} \in \mathbb{R} \]

[ ..., -9.5, ... ]

CUT, CUTE, CYCLIC

Learned Deviation Terms

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

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

**POS**

<table>
<thead>
<tr>
<th>MASC</th>
<th>FEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>3.2</td>
</tr>
</tbody>
</table>

**NEG**

<table>
<thead>
<tr>
<th>MASC</th>
<th>FEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.6</td>
<td>0.9</td>
</tr>
</tbody>
</table>

**NEU**

<table>
<thead>
<tr>
<th>MASC</th>
<th>FEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3.5</td>
<td>1.1</td>
</tr>
</tbody>
</table>

**BOY**

<table>
<thead>
<tr>
<th>MASC</th>
<th>FEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td></td>
</tr>
</tbody>
</table>

**KING**

<table>
<thead>
<tr>
<th>MASC</th>
<th>FEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>-6.8</td>
<td></td>
</tr>
</tbody>
</table>
Implication: obtain neighbors that modify nouns

\[ \tau(\nu) \propto \exp\{ f^{\top}_{\text{FEM}} \eta(\nu, \text{POS}) \} \]

\[ m_{\nu} \]

-9.5

-7.6

-6.1

CUTE

UGLY

INTELLIGENT

MASC  FEM

POS [ 1.1, 3.2 ]

POS [ -4.6, -0.7 ]

POS [ 1.1, 0.6 ]
Problem: corpus does not label sentiment

\[ p(\nu, n) = \sum_{s \in \mathcal{G}} p(\nu | n, s) p(s | n) p(n) \]

Objective:

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

\[
\min_{\eta, \omega, \xi} \sum_{n \in \mathcal{G}} \sum_{\nu \in \mathcal{V}} \hat{p}(\nu, n) \log(p(\nu, n)) + \beta \text{KL}( q(s \mid \nu) \parallel p(s \mid \nu) ) + \alpha \|\eta\|_1
\]

\(^8\) Hoyle et al, 2019
Results
Topics: 200 largest deviation terms for each gender-sentiment pair

$$\tau(\nu) \propto \exp\{f_{\text{FEM}}^\top \eta(\nu, \text{POS})\}$$
Adjective Super-senses
Verb Super-senses
Human Evaluation
Female bodies receive disproportionate attention

“Cute”\textsuperscript{9}

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BODY</td>
<td>0.78</td>
</tr>
<tr>
<td>FEELING</td>
<td>0.05</td>
</tr>
<tr>
<td>BEHAVIOR</td>
<td>0.04</td>
</tr>
<tr>
<td>SUBSTANCE</td>
<td>0.03</td>
</tr>
<tr>
<td>SOCIAL</td>
<td>0.02</td>
</tr>
</tbody>
</table>

\textsuperscript{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

<table>
<thead>
<tr>
<th>Strike</th>
<th>Weep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kill</td>
<td>Cry</td>
</tr>
<tr>
<td>Destroy</td>
<td>Frown</td>
</tr>
<tr>
<td>Violate</td>
<td>Gasp</td>
</tr>
<tr>
<td>Choke</td>
<td>Wreck</td>
</tr>
<tr>
<td>Embrace</td>
<td>Kiss</td>
</tr>
<tr>
<td>Grin</td>
<td>Attract</td>
</tr>
<tr>
<td>Seize</td>
<td>Wave</td>
</tr>
<tr>
<td>Act</td>
<td>Gush</td>
</tr>
<tr>
<td>Force</td>
<td>Dress</td>
</tr>
</tbody>
</table>

👍

👎

😑
Negative Adjectives

Hostile
Violent
Abusive
Brutal
Impotent

Distressed
Fragile
Helpless
Disagreeable
Unmarried
Widowed
Verbs where Noun is Subject

Succeed  
Flourish  
Protect  
Rescue  

Murder  
Fight  
Kill  
Threaten  

Giggle  
Kiss  
Smile  
Marry  

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

- charming
- attractive
- gentle
- sentimental
- strong
- weak
- handsome
- ambitious

\[ \text{Spearman’s } \rho \] 0.59

Williams and Best, 1977 & 1990

- feminine
- sentimental
- affectionate
- emotional
- masculine
- adventurous
- forceful
- aggressive

\[ \text{Spearman’s } \rho \] 0.33

Williams and Bennet, 1975

- charming
- attractive
- gentle
- sentimental
- strong
- weak
- handsome
- ambitious

\[ \text{Spearman’s } \rho \] 0.59

Williams and Best, 1977 & 1990

- feminine
- sentimental
- affectionate
- emotional
- masculine
- adventurous
- forceful
- aggressive

\[ \text{Spearman’s } \rho \] 0.33
Male adjectives align with human judgements

- Model
- Fem
- Masc

\[ \text{Human}^{10} \]

- Fem
- Masc

Adjectives Misclassified as Masculine
- Effeminate
- Submissive
- Cowardly
- Weak
- Timid

\[ ^{10} \text{Williams and Best, 1977 & 1990} \]
Caveats

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

Language changes over time, in particular that relating to gender\textsuperscript{11}

Reporting bias (“Black sheep”\textsuperscript{12})

Limited to binary gender

\textsuperscript{11} Underwood et al. (2018)
\textsuperscript{12} Meg Mitchell