Word Embeddings Quantify 100 years of Gender and Ethnic Stereotypes

Stats Reading Group

Ziyao Tian

Princeton University

1 May 2019

Today's Plan

- Overview
- Word Embedding
- Validation
- Quantify Gender Stereotypes
- Quantify Ethnic Stereotypes
- Discussions and Extensions

Overview

- Authors: Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. "Word embeddings quantify 100 years of gender and ethnic stereotypes." Proceedings of the National Academy of Sciences 115, no. 16 (2018): E3635-E3644.
- Language both reflects and perpetuates cultural stereotypes.
- Word embeddings can be used as a powerful tool to quantify historical trends and social change.
- The authors develop metrics based on word embeddings to characterize how gender stereotypes and attitudes toward ethnic minorities in the United States evolved during the 20th and 21st centuries starting from 1910.

Word Embedding

- word embeddings are a popular machine-learning method that represents each English word by a vector, such that the geometry between these vectors captures semantic relations between the corresponding words.
- trained on a large corpus of text

Word Embedding

- word embeddings are a popular machine-learning method that represents each English word by a vector, such that the geometry between these vectors captures semantic relations between the corresponding words.
- trained on a large corpus of text
- vectors being closer together has been shown to correspond to more similar words (e.g. XBox and PlayStation)

Word Embedding

- word embeddings are a popular machine-learning method that represents each English word by a vector, such that the geometry between these vectors captures semantic relations between the corresponding words.
- trained on a large corpus of text
- vectors being closer together has been shown to correspond to more similar words (e.g. XBox and PlayStation)
- can capture global relationships between words
  - London - England + France = ?

Word Embedding

- word embeddings are a popular machine-learning method that represents each English word by a vector, such that the geometry between these vectors captures semantic relations between the corresponding words.
- trained on a large corpus of text
- vectors being closer together has been shown to correspond to more similar words (e.g. XBox and PlayStation)
- can capture global relationships between words
  - London - England + France = Paris

Research Questions Rephrased

- Is word embedding useful in capturing stereotypes?
- If it is, what does it tell us about the changes of gender and ethnic stereotypes in the U.S. since 1910?
  - gender (exemplary): female, male
  - ethnic groups (exemplary): White, Asian, Hispanic

Validation: word embedding training

- contemporary: Google News word2vec
- historical: Google Books/Corpus of Historical American English (COHA)
  - led by BYU, 400 m words of text of American English from 1810 to 2009
  - the largest structured corpus of historical English
- robustness checks: GLoVe algorithm, New York Times Annotated Corpus 1988-2005, etc.

Validation: stereotype in word embedding

- stereotypes in (1) occupation and (2) personality trait (adjectives)
- collate several word lists to represent each gender and ethnicity, as well as neutral words.
  - gender: 20 related words (e.g. she, female, ... for "women")
    - *she, daughter, hers, her, mother, woman, girl, herself, female, sister, daughters, mothers, women, girls, femen, sisters, aunt, aunts, niece, niece*
  - ethnicity: last names (e.g. huang, chen ... for "Chinese")
  - occupation: available categories in U.S. census
  - adjectives: from previous studies on gender and ethnic stereotypes

Validation: simly embedding bias (e.g. Gender X Nurse)

- simly embedding bias
  - $\sum((-\|v_m - v_1\|_2) - (-\|v_m - v_2\|_2))$
  - negative norm distance captures similarity
  - v_1 is the average vector for women (i.e. average of 20 word vectors representing women)
  - v_m is a netural word (i.e. nurse) that belongs to set M ("occupation")
  - the more positive the difference, the more the target word (nurse) associates with women
- adaption for 3-ethnic-group comparison:
  $bias(hisp) = \sum[\frac{1}{2}(\|v_m - white\| + \|v_m - asian\|) - \|v_m - hisp\|]$

Validation: where is the “truth”

- occupation: finding the "unbiased" representation world
  - doable for occupation participation only via census data
  - occupation participant percent by gender and ethnic group
- personality traits: surveys of stereotypes
  - gender: John Williams and collegues (1977, 1990)
  - ethnic: Princeton Trilogy (1933, 1951, 1969)

Surveys of gender stereotypes

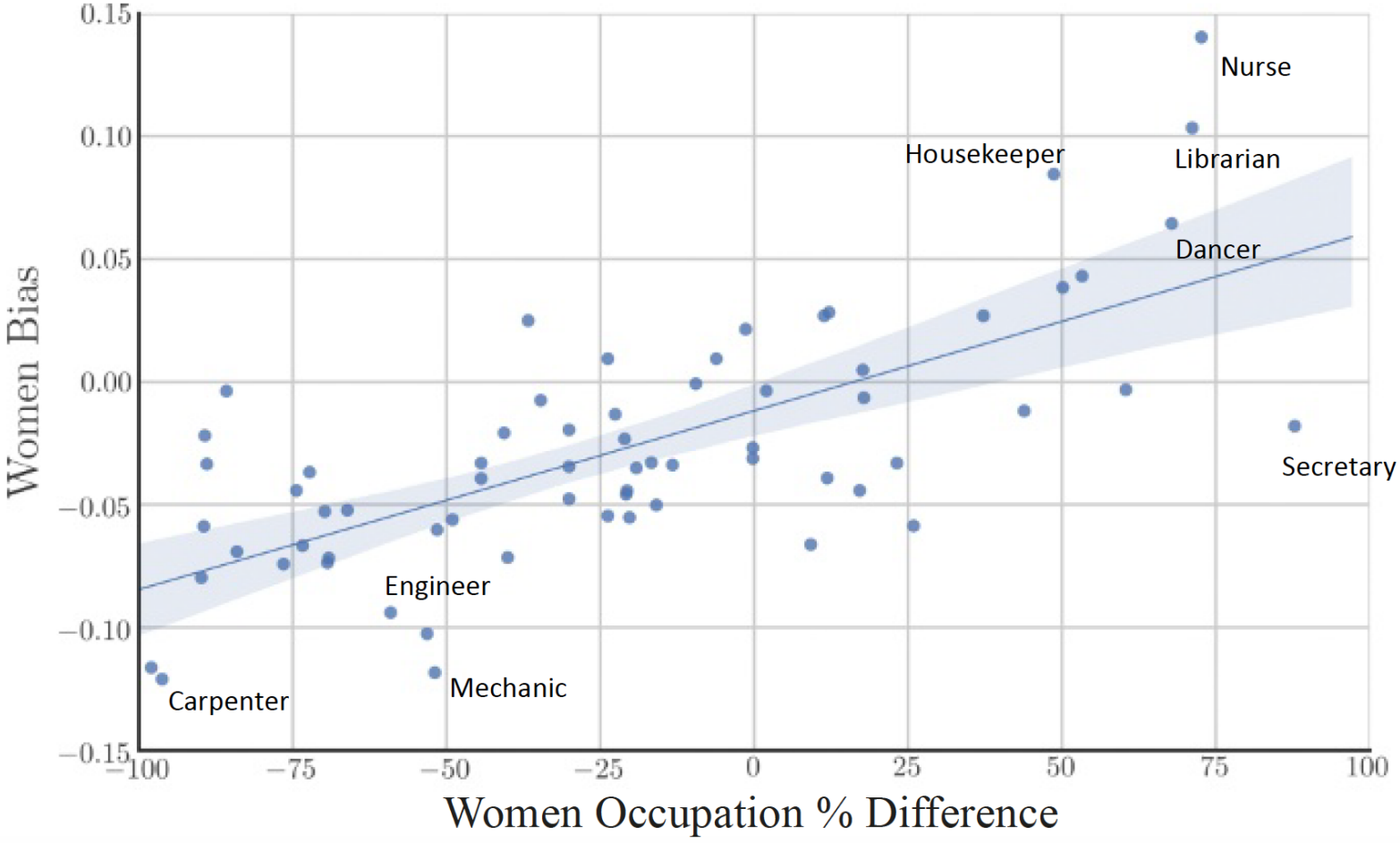
- Gender: asking students to consider each of the 300 adjectives on the Adjective Check List and to indicate whether it was more descriptive of men than women, more descriptive of women than men, or equally descriptive of both sexes. The male and female stereotypes were then defined by the selection of those adjectives which a majority of subjects of both sexes agreed were more characteristic of one sex than the other.

Surveys of gender stereotypes

- Gender: asking students to consider each of the 300 adjectives on the Adjective Check List and to indicate whether it was more descriptive of men than women, more descriptive of women than men, or equally descriptive of both sexes. The male and female stereotypes were then defined by the selection of those adjectives which a majority of subjects of both sexes agreed were more characteristic of one sex than the other.
- Ethnic: in 1933 (Katz and Braley), asking 100 male students from Princeton University to choose five traits that characterized different ethnic groups (for example Americans, Jews, Japanese, Negroes) from a list of 84 words; replicated in 1951 (Gilbert), many students expressed irritation at being asked to make generalizations at all and this could indicate a social change; replicated again in 1969 (Karlins et al.)

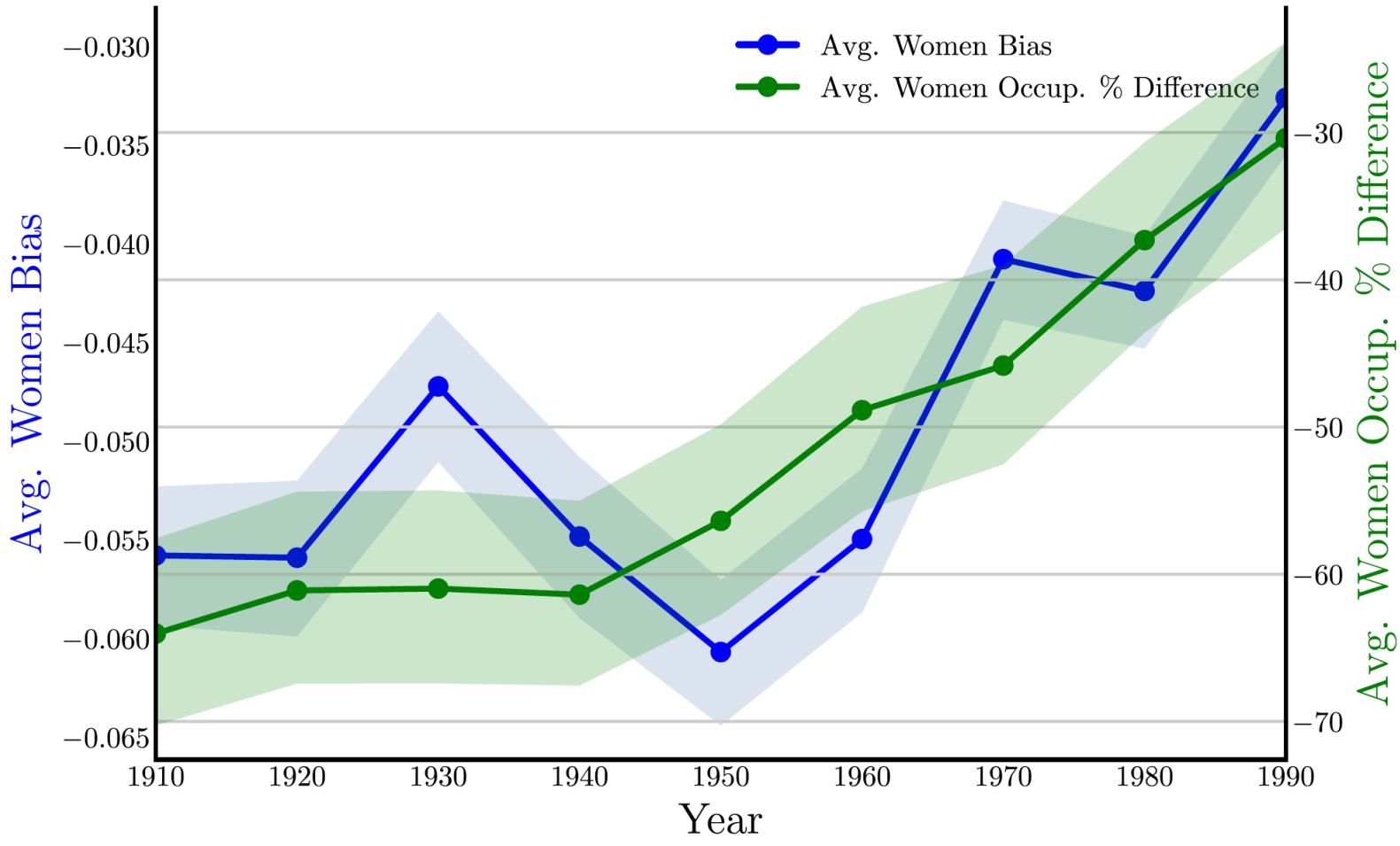
Validtion: how effective is "simply embedding bias"?

(1/2) Snapshot correlation between gender occupational bias and women's participation in 2015



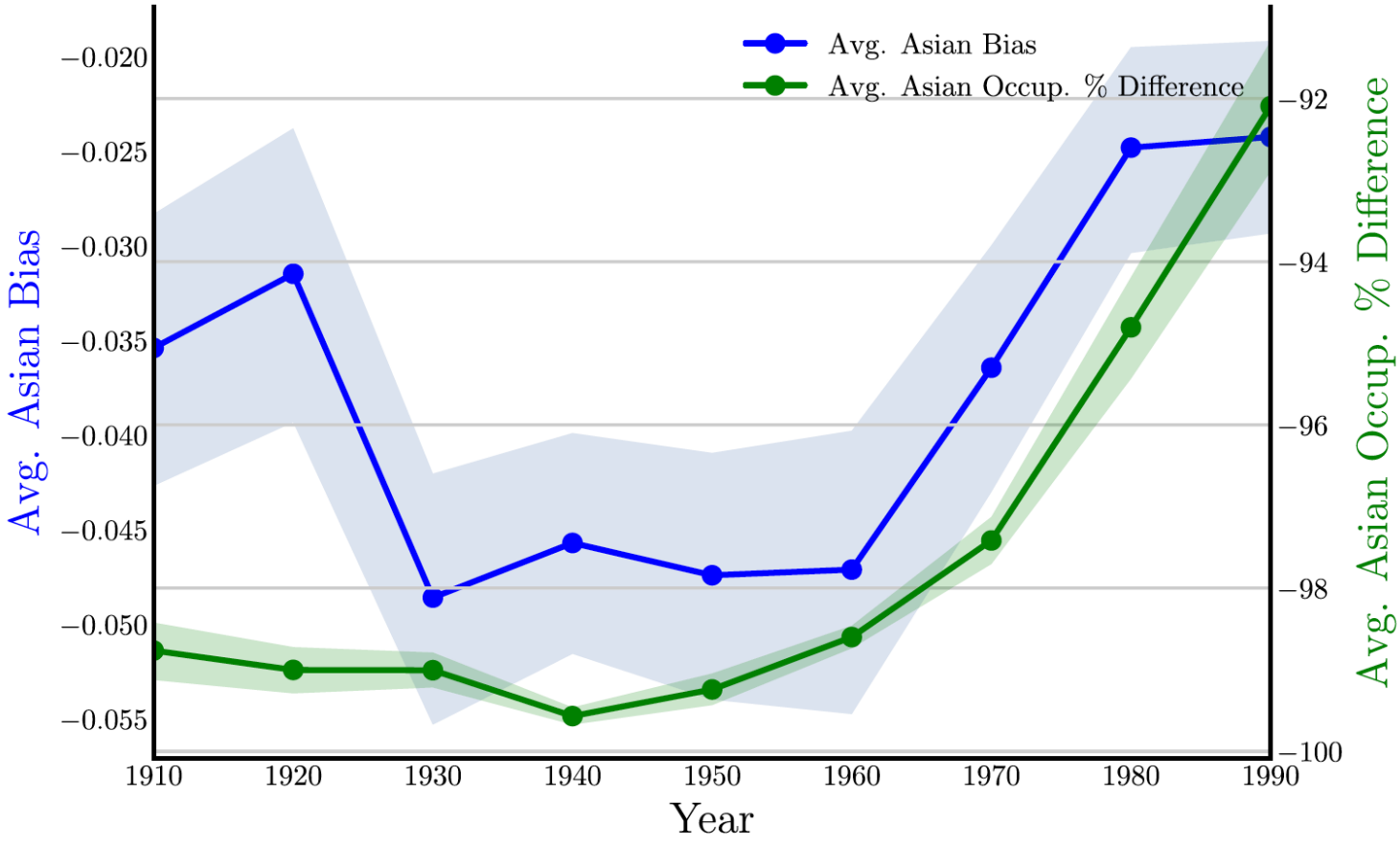
Validtion: how effective is "simply embedding bias"?

(2/2) Trends: the changes in embeddings over decades capture changes in the women's occupation participation



Validtion: how effective is "simply embedding bias"?

similar logic for ethnic occupation stereotypes



Validtion: how effective is "simply embedding bias"?

- Is this validation, or evidence to detect which occupations are more likely to be subject to gendered stereotypes? or both?
- For ethnic occupational stereotype, unlike roughly 50/50 share of female and male, should average percentage difference take into accout total population change?
  - from: $\frac{p_{min} - p_{white}}{p_{min} + p_{white}}$
  - to: $\frac{p_{min}}{P_{min}} - \frac{p_{white}}{P_{white}}$?

Validation: where is the “truth”?

- finding the "unbiased" representation of the world
- finding a good estimate of stereotypes documented via other ways
  - gender X occupation: MTurk done for contemporary
  - gender X personality trait: surveys
  - ethnic X occupation: ?
  - ethnic X personality trait: surveys

Validation in my view

- crowdsource in MTurk: crowdsource scores reflect aggregate human judgment as to whether an occupation is stereotypically associated with men or women.

```
lm1 <- simply embedding bias ~ occupation pct difference
lm2 <- crowdsource bias ~ occupation pct difference
```

- residual: the part of people's mindsets that cannot be justified on their "normal", "unbiased" observations of the world
- highly correlated: residual(lm1) and residual(lm2)

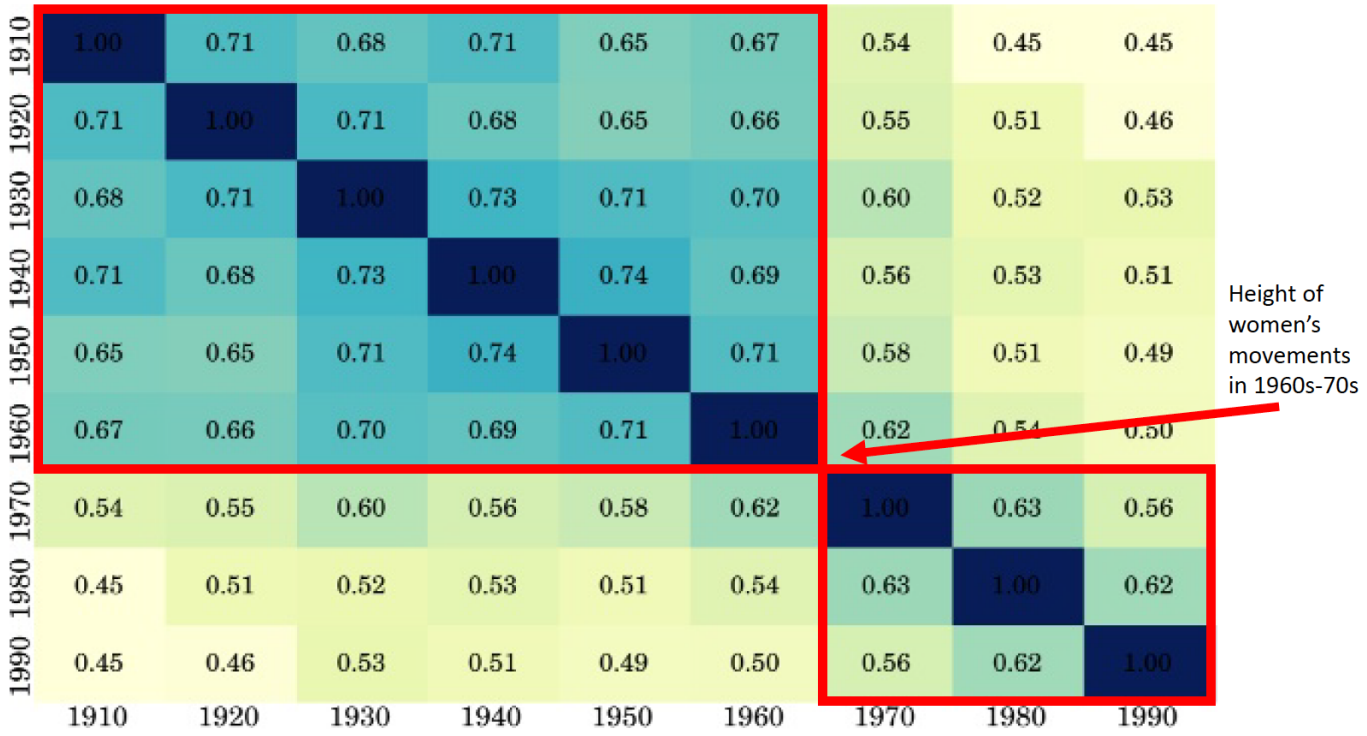
Quantify Gender Stereotypes 1/2: MTurk Contemporary Crowdsource Occupation Bias

"Bias beyond census data also appear in the embedding."

"Where such crowdsourcing is not possible, such as in studying historical biases, word embeddings can thus further serve as an effective measurement tool."

Quantify Gender Stereotypes 2/2: Changing Attitudes

- Heat map cell: $corr[dist(women, adj_i|t_m), dist(women, adj_i)|t_n]$
- sharp divide between the 1960s and 1970s
  - support a: smallest correlation in adjacent decades
  - support b: Kolmogorov-Smirnov test for phase change (appendix table B.23)



Quantify Gender Stereotypes 2/2: Extension

- How various narratives and descriptions of women developed and competed over time
  - e.g. competence, physical appearance
  - a society with decreasing but still significant gender biases.
- Discussion: do we see what we want to see? (1980-1990 support 1; 1940-50, 70-80 support 2)

Transition Interval	Test statistic	p value
1910-1920	0.449	0.1206
1920-1930	0.3265	0.4475
1930-1940	0.449	0.1206
1940-1950	0.5306	0.03958
1950-1960	0.3469	0.3714
1960-1970	0.8571	7.173e-05
1970-1980	0.551	0.0291
1980-1990	0.3265	0.4475

Table B.23: Kolmogorov-Smirnov tests for phase change for Figure 4.

Quantify Ethnic Stereotypes 1/2: Changing Attitudes

Similarly, trends in Asian stereotypes suggest "how external events changed attitudes"

Quantify Ethnic Stereotypes 2/2: Extension

- Applications:
  - Islam and Terrorism
  - Russian and Ethnic Adjectives
- Signifance: useful to "examine shifts in the attitudes toward other ethnic groups, especially around significant global events."
  - extension idea: immigration threat? Immigrants have fewer (good) jobs than what people assume?

Discussions

- Simple metrics
- Cautious authors
  - robustness checks: embedding algorithms, metric, corpus, etc.
  - linear association
  - the "black box" of word embedding
  - who's the authors of historical text
  - recall vs. precision
  - dependency on specific word list

Discussions

- Simple metrics
- Cautious authors
  - robustness checks: embedding algorithms, metric, corpus, etc.
  - linear association
  - the "black box" of word embedding
  - who's the authors of historical text
  - recall vs. precision
  - dependency on specific word list

Discussions

- Where is the truth? Recall vs. Precision
  - their metrics verify what previous studies see as potentially stereotypical
  - could be other candidates
  - the world could change

Discussions

- Where is the truth? Recall vs. Precision
  - their metrics verify what previous studies see as potentially stereotypical
  - could be other candidates
  - the world could change
- Dependency on specific word lists (can have traits)
  - consistent comparison
  - rank-rank slope, intergenerational elasticity

Discussions

- Where is the truth? Recall vs. Precision
  - their metrics verify what previous studies see as potentially stereotypical
  - could be other candidates
  - the world could change
- Dependency on specific word lists (can have traits)
  - consistent comparison
  - rank-rank slope, intergenerational elasticity
- "Language both reflects and perpetuates cultural stereotypes."
  - We see a number of good reflections
  - How can study the reproduction and perptuation of stereotypes? Do certain texts carry more weight than others?

Summary

- Word embedding can be a powerful tool in capturing human biases.
- These stereotypes are automatically learned by the embedding algorithm and could be problematic if the embedding is then used for sensitive applications such as search rankings, product recommendations, or translations.
- More space to take advantage of the tool, learn how it captures stereotypes, and how we can debias it.

References

- Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. "Word embeddings quantify 100 years of gender and ethnic stereotypes." Proceedings of the National Academy of Sciences 115, no. 16 (2018): E3635-E3644.
- Williams, John E., and Deborah L. Best. "Sex stereotypes and trait favorability on the Adjective Check List." Educational and Psychological Measurement 37, no. 1 (1977): 101-110.
- Williams, John E., and Deborah L. Best. Measuring sex stereotypes: A multination study, Rev. Sage Publications, Inc, 1990.
- Devine, Patricia G., and Andrew J. Elliot. "Are racial stereotypes really fading? The Princeton trilogy revisited." Personality and social psychology bulletin 21, no. 11 (1995): 1139-1150.
- "The Princeton Trilogy" https://www.appsychology.com/IB%20Psych/IBcontent/Studies/princeton_trilogy.htm