Word embeddings quantify 100 years of gender and ethnic stereotypes

•••

Simrin Kalkat, Marc Huo

Sections (for our use) - discussion question at end of section

- Introduction (Marc)
  - Word embeddings briefly
  - Historical context for stereotyping
  - Interactive activity
  - Discussion question:
- Methods (Simrin)
  - Embedding Framework Overview and Validations
  - Real-world implications
- Quantifying Gender Stereotypes (Simrin)
  - Broader context + analyzing specific results
  -
- Quantifying Ethnic Minority Stereotypes (Marc)
- Limitations (Marc)
- Broader research context (Simrin)
  - Analyzing other literature in the space

Introduction

- Previous studies used human surveys and manual dictionary/qualitative analysis to measure stereotypes - not robust and scalable across history
    - Discussion (2 min): What factors may prevent the scalability and robustness of qualitative methods in examining language throughout history?

Introduction

- Previous studies used human surveys and manual dictionary/qualitative analysis to measure stereotypes - not robust and scalable across history
  - Discussion (2 min): What factors may prevent the scalability and robustness of qualitative methods in examining language throughout history?
- Study proposes use of word embeddings to study semantic relations between words, even if they're implicit or subtle biases
  - e.g. honorable closer to man, submissive closer to woman
  - https://projector.tensorflow.org/

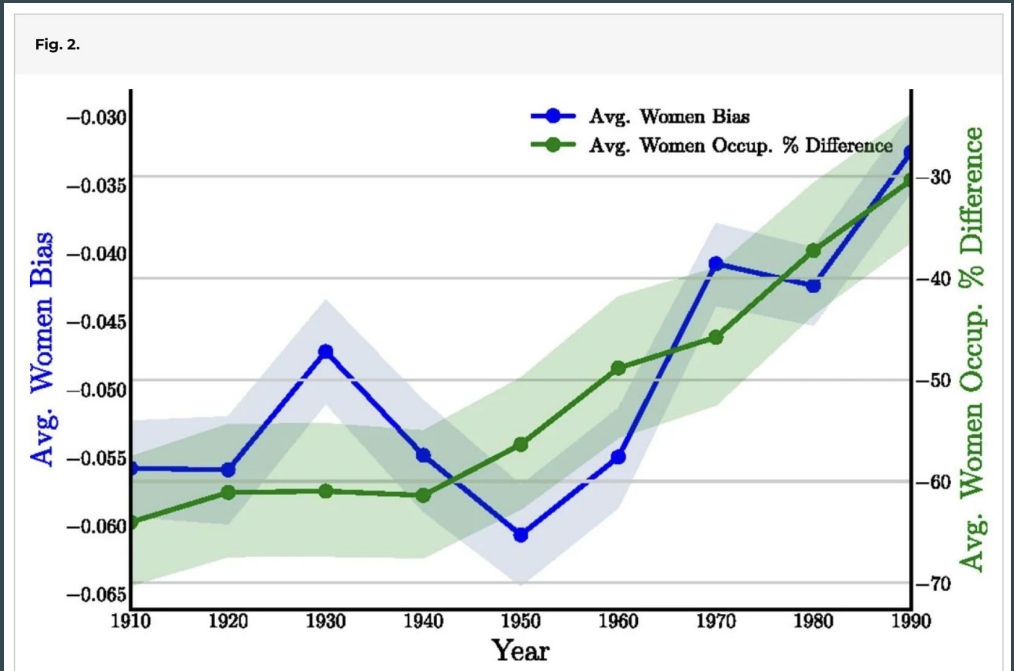
Introduction

- Previous studies used human surveys and manual dictionary/qualitative analysis to measure stereotypes - not robust and scalable across history
  - Discussion (2 min): What factors may prevent the scalability and robustness of qualitative methods in examining language throughout history?
- Study proposes use of word embeddings to study semantic relations between words, even if they're implicit or subtle biases
  - e.g. honorable closer to man, submissive closer to woman
  - https://projector.tensorflow.org/
- Study specifically focused on women's movement in the 1960s-1970s and the Asian-American population growth in the 1960s and 1980s
  - Discussion (2 min): What other historical trends and movements would you like to examine

Methods

- Data
  - Google News word2vec vectors for contemporary snapshot analysis
  - Previously trained Google Books/Corpus of Historical American English (COHA) embeddings for historical temporal analysis
  - Embeddings trained using the GLoVe algorithm on the New York Times Annotated Corpus.
- Word lists
  - Gender – lists representing men and women
  - Ethnicity – white, asian, hispanic
  - Neutral words – adjectives and occupations
    - Historical census data for occupations
- Embedding Bias
  - Computed average embedding distance between words that represent women and words for occupations. They also computed the average embedding distance between words that represent men and the same occupation words. The difference of the average distances was their metric for bias.
  - Eg: avg1 = she -> teacher, avg2 = he -> teacher, bias = avg1 - avg2

Quantifying Gender Stereotypes



We see that bias steadily decreases as we see more representation of women in the other occupancies:

DISCUSSION QUESTION:

- How might the use of embedding bias analysis help organizations and policymakers promote greater gender diversity in professional occupations?
- What are some limitations of current diversity programs in big tech/finance and how might we mitigate them in the future?

Quantifying Gender Stereotypes

Looking at words describing perceived competence and physical appearance

1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

DISCUSSION QUESTION (1):

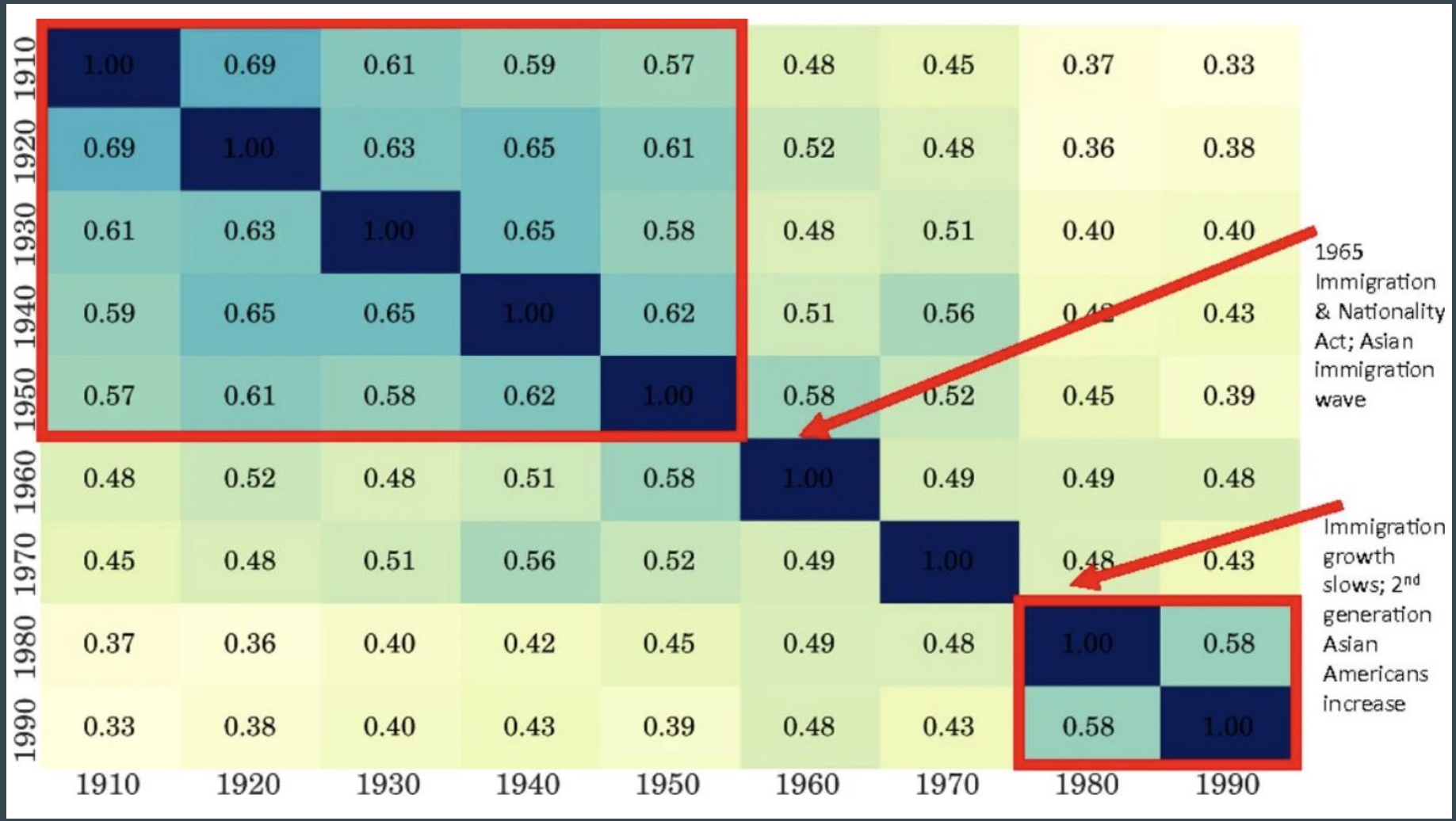
What other components of gender stereotypes would you have wanted to analyze if the study were conducted in the 2010s?

DISCUSSION QUESTION (2):

How has our narrative of what femininity and womanhood means changed over the last few decades? How do you see the women's movement in 60s/70s contributing to the change in word embeddings between 1950 and 1990

Quantifying Ethnic Minority Stereotypes

- Study focused primarily on Asian-American immigration from the 1960's-1980's using corpus with distinctly "Asian" last names - results in a list of 20 last names that are primarily but not exclusively Chinese last names
  - Discussion (2 min): How do you feel about the sampling method used for extraction of "Asian" last names?
- Two phase changes:
  - 1965 Immigration and Nationality Act coincide with sharp rise in Asian immigration in US
  - 1980s second-generation Asian-American population


1910 1.00 0.69 0.61 0.59 0.57 0.48 0.45 0.37 0.33
1920 0.69 1.00 0.63 0.65 0.61 0.52 0.48 0.36 0.38
1930 0.61 0.63 1.00 0.65 0.58 0.48 0.51 0.40 0.40
1940 0.59 0.65 0.65 1.00 0.62 0.51 0.56 0.43
1950 0.57 0.61 0.58 0.62 1.00 0.58 0.52 0.45 0.39
1960 0.48 0.52 0.48 0.51 0.58 1.00 0.49 0.49 0.48
1970 0.45 0.48 0.51 0.56 0.52 0.49 1.00 0.48 0.43
1980 0.37 0.36 0.40 0.42 0.45 0.49 0.48 1.00 0.58
1990 0.33 0.38 0.40 0.43 0.39 0.48 0.43 0.58 1.00
1910 1920 1930 1940 1950 1960 1970 1980 1990
1965 Immigration & Nationality Act; Asian immigration wave
Immigration growth slows; 2nd generation Asian Americans increase

Quantifying Ethnic Minority Stereotypes

- Biased Asian adjectives extracted went from strongly negative words describing outsiders (pre-1950s) such as “barbaric, hateful, monstrous, bizarre, and cruel” to more stereotypical adjectives such as “sensitive, passive, complacent, active, and hearty”
- Discussion (2 min): It seems as if there was a complete 180 in adjectives used to describe Asian immigrants - why is this the case (ignorance, outsider bias, etc.)?
- Discussion: How would you be able to more closely measure “othering” and dehumanization of immigrants over time?

Quantifying Ethnic Minority Stereotypes

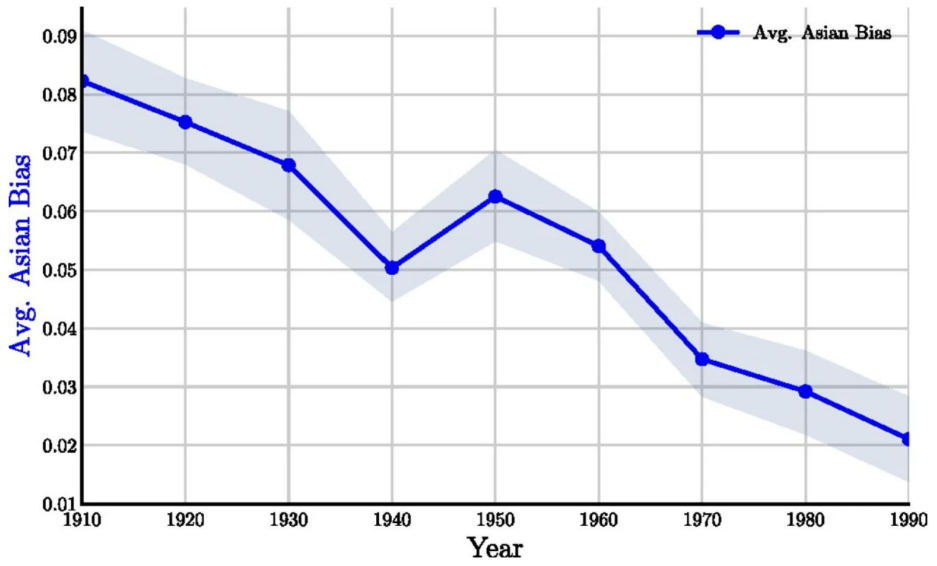
- Islam (vs. Christianity) associated with terrorism-related words such as terror, bomb, and violence
  - Increase in association seen after 1993 WTC bombing and 9/11
- Hispanic name stereotyping increased gradually and steadily over time in contrast to Asian and East European stereotypes, due to lack of any major historical event
- Discussion: What potential differences may exist between short-term (catalyzed by specific historical events) and long-term stereotypic rhetoric and what recent events have you seen in history that have contributed to stereotyping?

Quantifying Ethnic Minority Stereotypes

Fig. 7.

Religious (Islam vs. Christianity) bias score over time for words related to terrorism in New York Times data. Note that embeddings are trained in 3-y windows, so, for example, 2000 contains data from 1999–2001. The shaded region is the bootstrap SE interval.

Fig. 6.



Asian bias score over time for words related to outsiders in COHA data. The shaded region is the bootstrap SE interval.

Broader research context

Limitations

- Discussion: What limitations have you noticed in the text?

Limitations

- Linear models used to fit relationship between embedding bias and various external metrics, whereas true relationships may be nonlinear
- Generalizability of the specific word lists used and the recall of the models used in capturing human biases
- Historical texts do not actually reflect the popular opinions at that time
    - Discussion: Which historical texts are more susceptible to not reflecting popular opinions of society and which historical texts may actually be a good indicator of public opinion?
- Embeddings used are fully “black box”, limiting causal explanations of how stereotypes appear in text. What structural features of words actually confer “stereotyping”?
    - Discussion: What efforts can be made to actually understand the semantic underpinnings of how stereotypes appear in text?