Word embeddings quantify 100 years of gender and ethnic stereotypes

Nikhil Garg[a,1], Londa Schiebinger[b], Dan Jurafsky[c,d], and James Zou[e,f,1]

[a]Department of Electrical Engineering, Stanford University, Stanford, CA 94305; [b]Department of History, Stanford University, Stanford, CA 94305; [c]Department of Linguistics, Stanford University, Stanford, CA 94305; [d]Department of Computer Science, Stanford University, Stanford, CA 94305; [e]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305; and [f]Chan Zuckerberg Biohub, San Francisco, CA 94158

Word embeddings are a powerful machine-learning framework that represents each English word by a vector. The geometric relationship between these vectors captures meaningful semantic relationships between the corresponding words. In this paper, we develop a framework to demonstrate how the temporal dynamics of the embedding helps to quantify changes in stereotypes and attitudes toward women and ethnic minorities in the 20th and 21st centuries in the United States. We integrate word embeddings trained on 100 y of text data with the US Census to show that changes in the embedding track closely with demographic and occupation shifts over time. The embedding captures societal shifts—e.g., the women's movement in the 1960s and Asian immigration into the United States—and also illuminates how specific adjectives and occupations became more closely associated with certain populations over time. Our framework for temporal analysis of word embedding opens up a fruitful intersection between machine learning and quantitative social science.

word embedding | gender stereotypes | ethnic stereotypes

The study of gender and ethnic stereotypes is an important topic across many disciplines. Language analysis is a standard tool used to discover, understand, and demonstrate such stereotypes (1–5). Previous literature broadly establishes that language both reflects and perpetuates cultural stereotypes. However, such studies primarily leverage human surveys (6–16), dictionary and qualitative analysis (17), or in-depth knowledge of different languages (18). These methods often require time-consuming and expensive manual analysis and may not easily scale across types of stereotypes, time periods, and languages. In this paper, we propose using word embeddings, a commonly used tool in natural language processing (NLP) and machine learning, as a framework to measure, quantify, and compare beliefs over time. As a specific case study, we apply this tool to study the temporal dynamics of gender and ethnic stereotypes in the 20th and 21st centuries in the United States.

In word-embedding models, each word in a given language is assigned to a high-dimensional vector such that the geometry of the vectors captures semantic relations between the words—e.g., vectors being closer together has been shown to correspond to more similar words (19). These models are typically trained automatically on large corpora of text, such as collections of Google News articles or Wikipedia, and are known to capture relationships not found through simple co-occurrence analysis. For example, the vector for France is close to vectors for Austria and Italy, and the vector for XBox is close to that of PlayStation (19). Beyond nearby neighbors, embeddings can also capture more global relationships between words. The difference between London and England—obtained by simply subtracting these two vectors—is parallel to the vector difference between Paris and France. This pattern allows embeddings to capture analogy relationships, such as London is to England is as Paris is to France.

Recent works demonstrate that word embeddings, among other methods in machine learning, capture common stereotypes because these stereotypes are likely to be present, even if subtly,

in the large corpora of training texts (20–23). For example, the vector for the adjective honorable would be close to the vector for man, whereas the vector for submissive would be closer to woman. 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. An important direction of research is to develop algorithms to debias the word embeddings (20).

In this paper, we take another approach. We use the word embeddings as a quantitative lens through which to study historical trends—specifically trends in the gender and ethnic stereotypes in the 20th and 21st centuries in the United States. We develop a systematic framework and metrics to analyze word embeddings trained over 100 y of text corpora. We show that temporal dynamics of the word embedding capture changes in gender and ethnic stereotypes over time. In particular, we quantify how specific biases decrease over time while other stereotypes increase. Moreover, dynamics of the embedding strongly correlate with quantifiable changes in US society, such as demographic and occupation shifts. For example, major transitions in the word embedding geometry reveal changes in the descriptions of genders and ethnic groups during the women's movement in the 1960s–1970s and Asian-American population growth in the 1960s and 1980s. We validate our findings on external metrics and show that our results are robust to the different algorithms for training the word embeddings. Our framework reveals and quantifies how stereotypes toward women and ethnic groups have evolved in the United States.

Significance

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. We demonstrate that word embeddings can be used as a powerful tool to quantify historical trends and social change. As specific applications, we 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. Our framework opens up a fruitful intersection between machine learning and quantitative social science.

COMPUTER SCIENCES

SOCIAL SCIENCES

Replication Exercise

PPOL-6801

Chris Cheng

&

Muhammad Saad

Paper Overview

Word embeddings quantify 100 years of ethnic and gender stereotypes, 2018

Key Takeaway

Word embeddings trained on text data from the 20th and 21st century can be used to quantify/track socioeconomic trends towards gender and ethnic minorities in the US
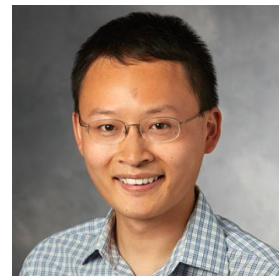
Authors

Nikhil Garg

Londa Schienberger

Dan Jurafsky

James Zou

*Collaborative effort b/w Dept. of **Engineering**, **Computer Science**, **Linguistics**, and **Data Science** at **Stanford University***

Research Framing Intuition

Do historical biases measured through embeddings match those measured through historical census/survey data in the US?

Embedding Bias

Calculate *distance between words that represent category of interest (gender, ethnicity) with words for occupations and stereotypes*

➡️

Societal Trends

Compare *with independent demographic trends related to occupation rates and historical surveys*

➡️

Demonstrate(!)

Show *that embedding bias accurately captures gender/ethnic occupation rates and historical changes*

Data & Embedding Algorithms

Use pre-trained word embeddings on Google, NYT data for text analysis. Validate through census and survey data

Text Data

Google News
(current)

Google Books & COHA
(1910-1990)

New York Times
(1988 - 2005)

➡️

Embeddings

Pre-trained word embeddings

Word2Vec(Skim-gram Negative Sampling)
GloVe

➡️

Validation

U.S. Census Data
(Occupation Rates)

Historical Surveys such as Williams & Best 1977/1990, Princeton Trilogy 1933–1969
(Stereotypes)

Analytical Pipeline

Use pre-trained word embeddings

⬇

Build word lists for groups (men/women and White/Asian/Hispanic) and neutral categories (occupations, adjectives)

⬇

Compute embedding bias metric using the cosine similarity metric

$$\text{Bias} = \sum (\|v_m - v_{\text{men}}\| - \|v_m - v_{\text{women}}\|)$$

⬇

Compare the bias captured by the embeddings against the census and survey results

Initial Replication Challenges

Data Unavailability

Resource Does Not Exist: The New York Times Annotated Corpus

This corpus is no longer available.

LDC Catalog No.:	LDC2008T19
ISBN:	1-58563-486-7
ISLRN:	429-488-225-160-9
DOI:	https://doi.org/10.35111/77ba-9x74
Citation:	Sandhaus, Evan. The New York Times Annotated Corpus LDC2008T19. Web Download. Philadelphia: Linguistic Data Consortium, 2008.

NYT Annotated Corpus no longer available (!)
No back-up with authors

Pre-trained Embeddings Mismatch

word2vec

Google Word2Vec embeddings used by authors trained up till 2015. Our case, till 2023.

Result 1

Google News: Gender Embeddings Bias vs Gender Occupation Rate (Overall snapshot)
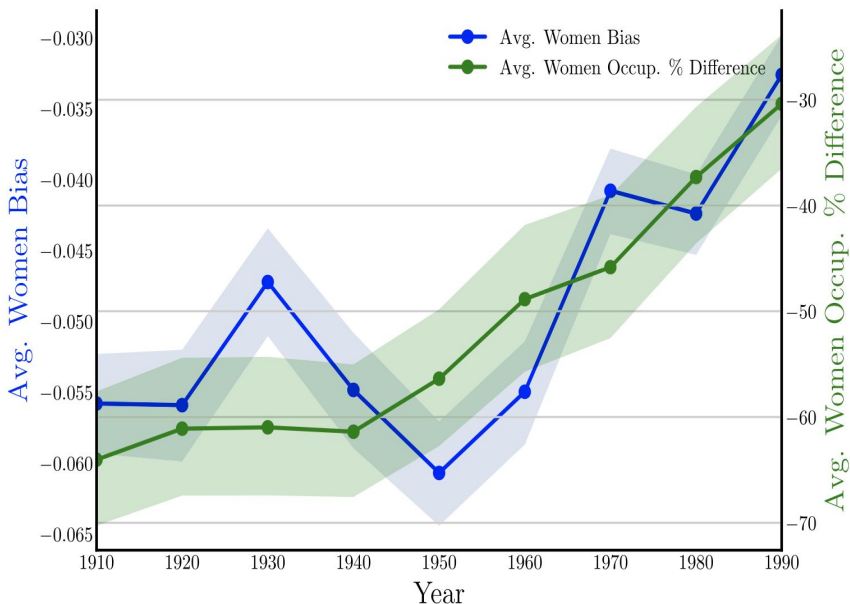
Study

Replication
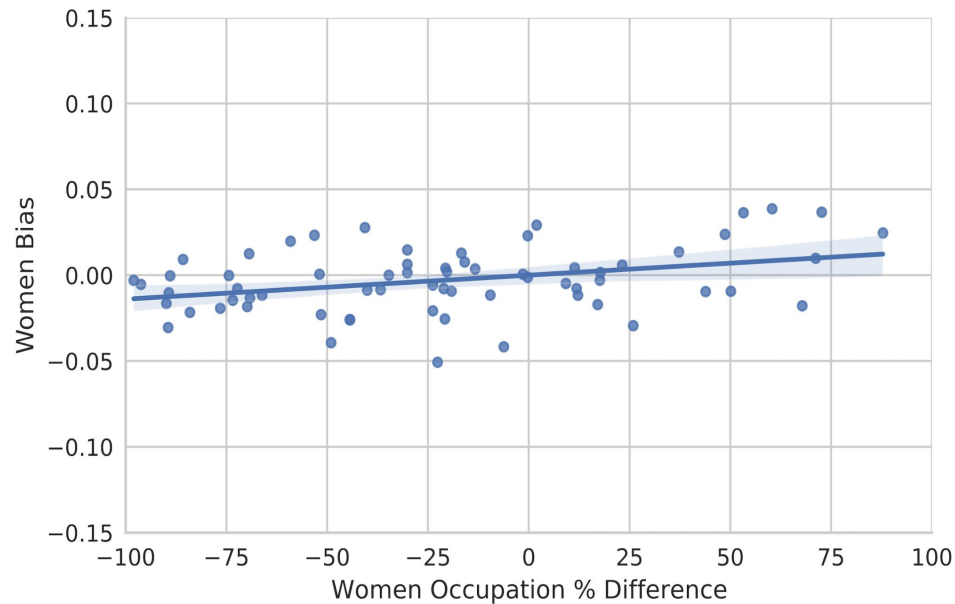
Mismatch: Lower variance in embedding bias compared to study
Reason: Possibly data. Google News data updated from 2015 to 2023; some words might not exist in the 2023 training data
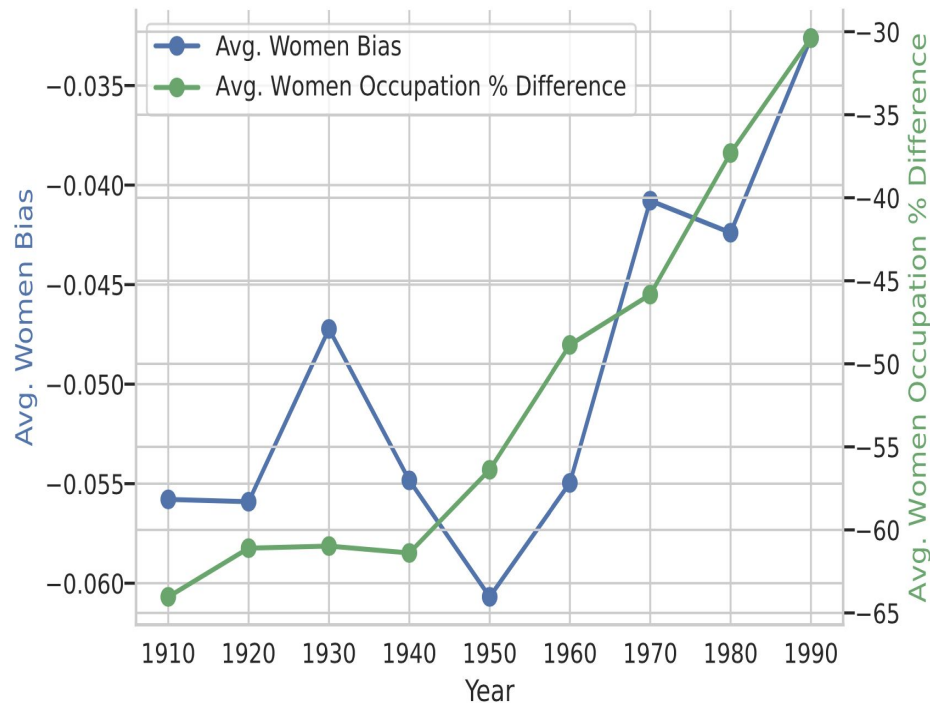
*COHA: **Gender Average Embedding Bias** vs Average Occupation Rate (Temporal 1910-90)*
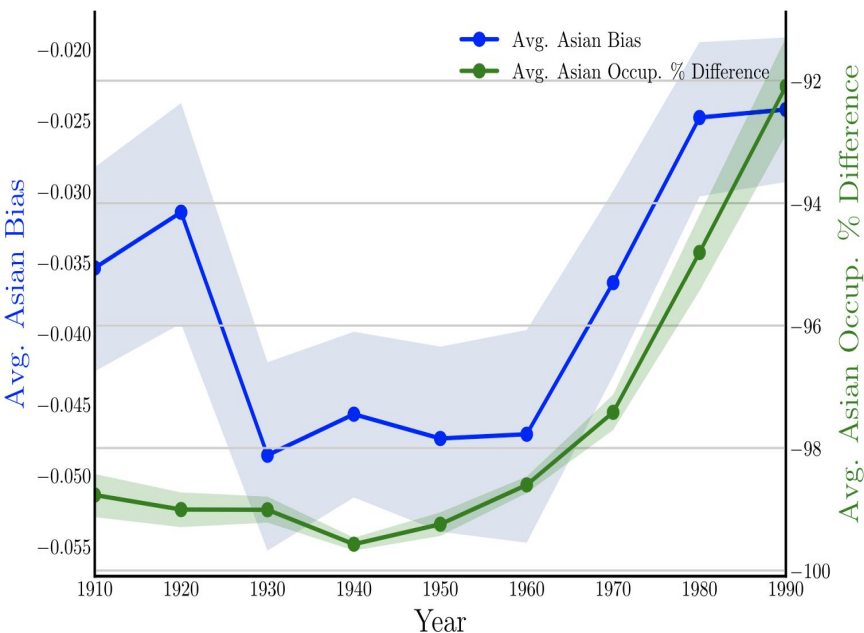
Study

Replication

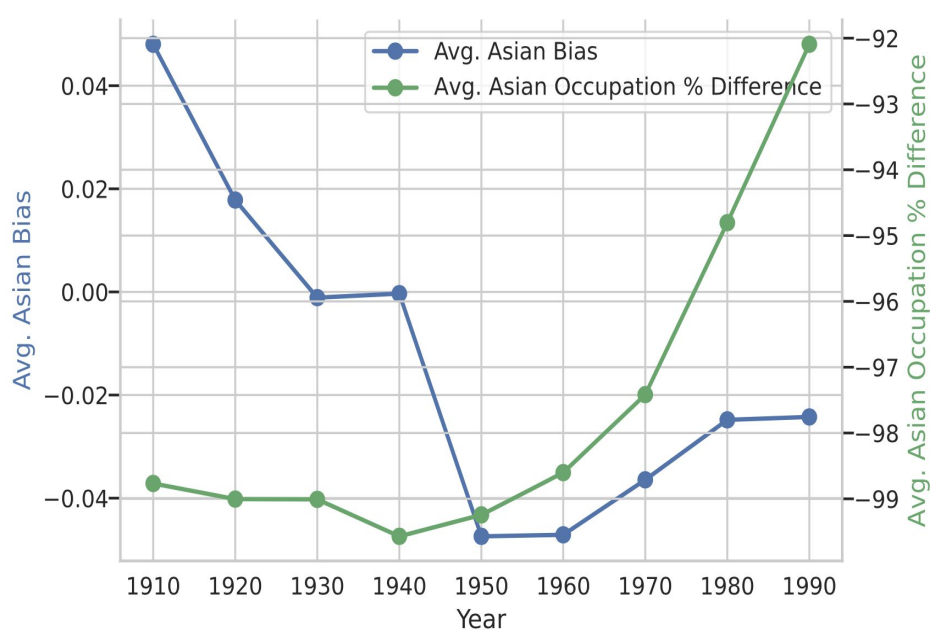**Perfect replication!**

Result 3

*COHA: **Ethnicity Average Embedding Bias** vs Average Occupation Rate (Temporal 1910-90)*
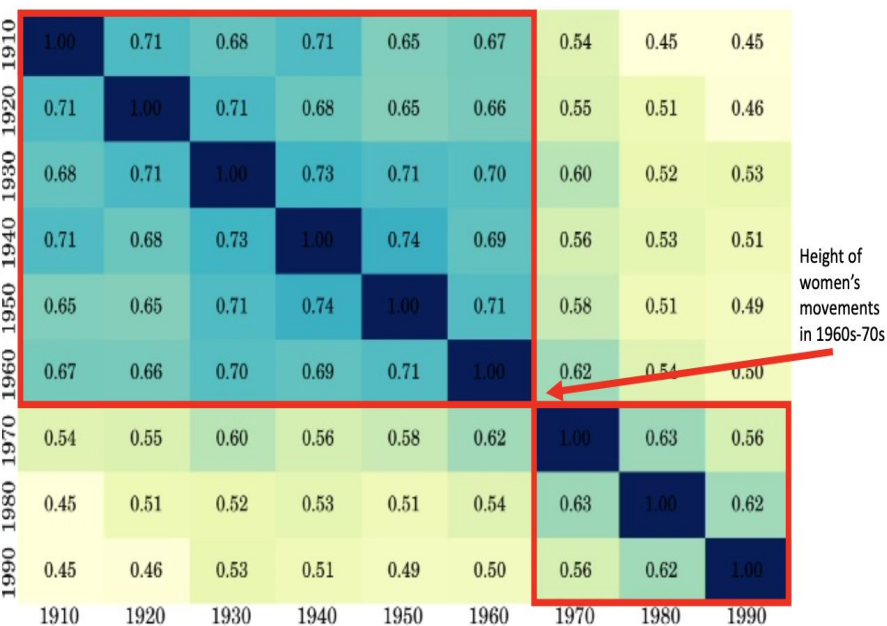
Study

Replication

Mismatch: Embedding bias does not match the original study for the replication
Reason: The authors likely performed some preprocessing or filtering that they did not include in the repository
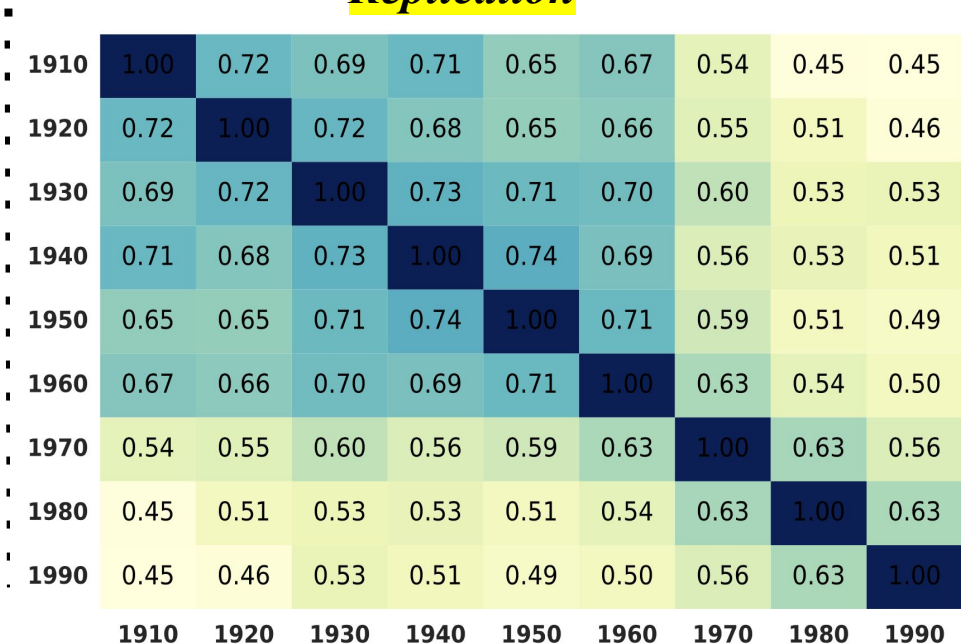
Result 4

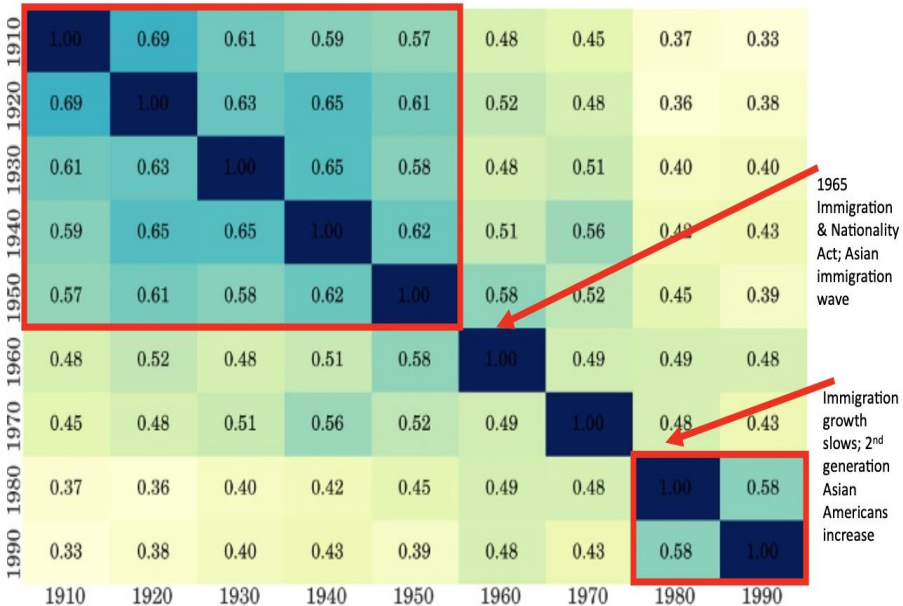*COHA: Correlation of Embeddings Bias for* **Women Related Adjectives**
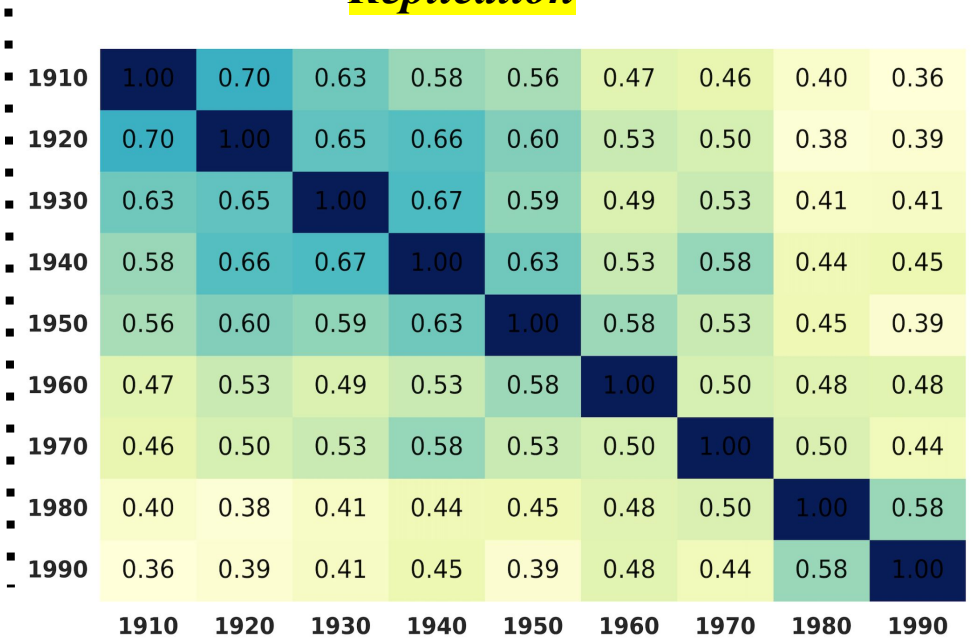
Study

Replication

Height of women's movements in 1960s-70s

Near perfect replication…

Result 5

*COHA: Correlation Embeddings Bias Score for **Asian Related Adjectives***

Study

Replication

1965 Immigration & Nationality Act; Asian immigration wave

Immigration growth slows; 2nd generation Asian Americans increase
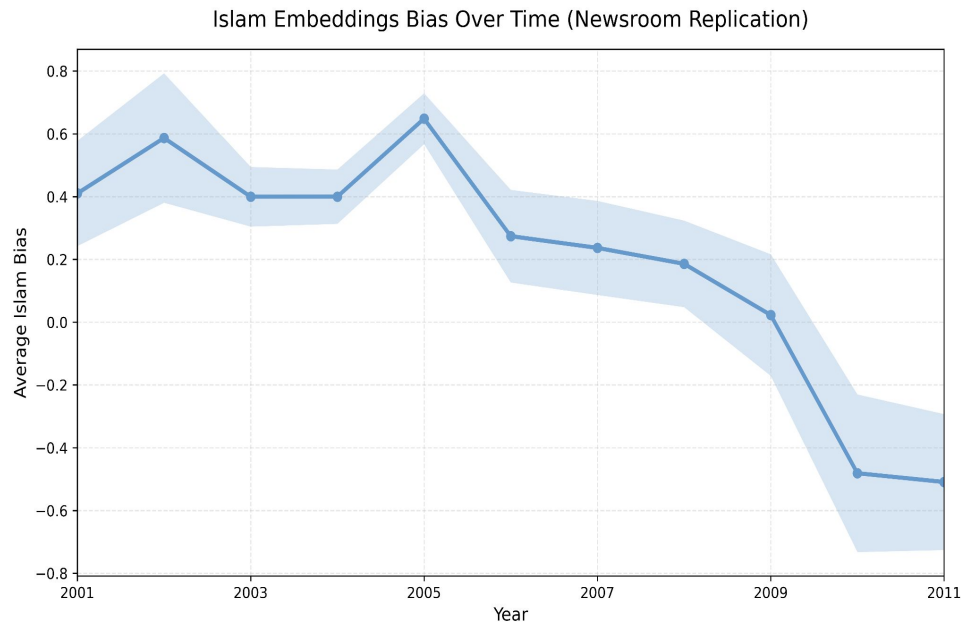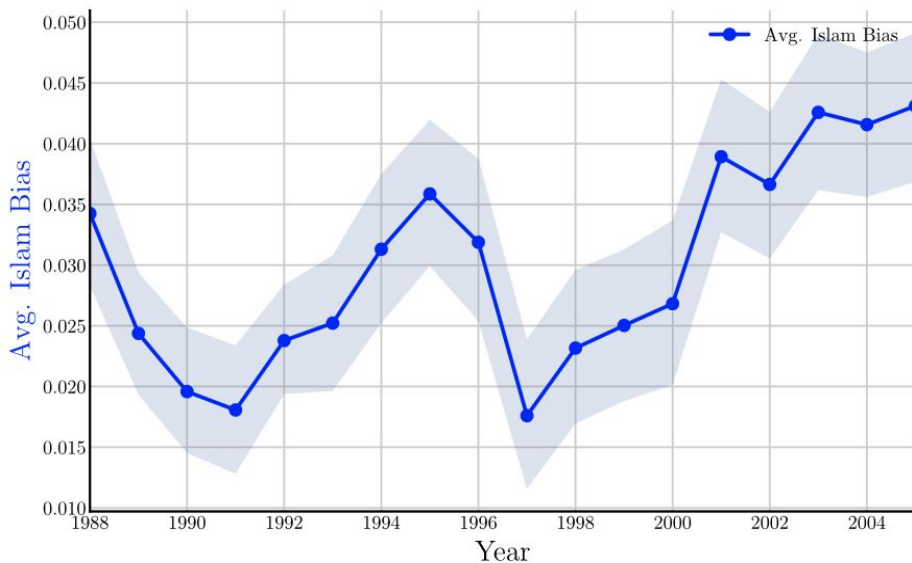
Near perfect replication…

Result 6:

Islam Bias Score: NYT (Study) vs Newsroom (Replication extension)

Study   *Extension*

Mismatch: Used Newsroom data; only available 1998 onwards
Close correspondence b/w overall trend near 2001 for NYT and Newsroom.
Another peak in bias around 2005 (London bombings)

Replication Autopsy

What worked ✅

The overall **intuition** and **logical flow was clear**. Easy replicate aspects like bias calculations and validation checks.

Most of the **data was publicly available**.

What did <u>not</u> work ❌

Repo Organization:
- Incomplete scripts for preprocessing
- Creation of intermediate folders not mentioned
- The read_me was quite concise

Use of python2 vs python3
- Study ran in 2015. A lot of the packages are slightly outdated wrt to updated python compatibility.
- Had to update packages/use alternatives based on newer versions.

Study Extension

We recommend using publicly accessible text data, updated embeddings and human validation as extension

Open Access Data

Replace NYT Annotated Corpus with reproducible news datasets

We used Cornell NewsRoom. Other options include: HuffPost, etc

+

Updated Embeddings

Use Transformers based embeddings (BERT) instead of SGNS

Can more accurately capture semantic meaning without word lists

+

Human Validation

Use of words lists

Key step compilation of word lists representative of different categories. Process is arbitrary and could use human reviewers.

Thank you!

Questions…

Replication problems

1.Common problems:
　　1)bad organize of project directory
　　2)not complete scripts for preprocessing
　　3)python2

2.A little mistake we made — used the wrong raw data
　　In essay the author used so called "Google Books/Corpus of Historical American English (COHA)", and referred this as "COHA" in the main body, while the real data author used is called "eng-all" on the download page.

1.Common extensions:

    1)include other topics that changes across a century, like certain kind of impression on a specific country.

    2)include more word embedding sources for robustness test.

2.Using transformers embedding instead of SGNS

    With SGNS, to get a less biased word vector, it require a word lists that contain synonyms and calculate the average of their vectors, while the word might be uncommon or not appear in all decades. Transformer-based word embedding method can more accurately capture the semantic meaning of words without the need of a word list.