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

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 12, 2018 (received for review November 22, 2017)

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 to England is as Paris 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.

Author contributions: N.G., L.S., D.J., and J.Z. designed research; N.G. and J.Z. performed research; and N.G. and J.Z. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Data and code related to this paper are available on GitHub (https://github.com/nikhgarg/EmbeddingDynamicStereotypes).

[1]To whom correspondence may be addressed. Email: nkgarg@stanford.edu or jamesz@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1720347115/-/DCSupplemental.

Published online April 3, 2018.

Our results demonstrate that word embeddings are a powerful lens through which we can systematically quantify common stereotypes and other historical trends. Embeddings thus provide an important quantitative metric which complements existing, more qualitative, linguistic and sociological analyses of biases. In *Embedding Framework Overview and Validations*, we validate that embeddings accurately capture sociological trends by comparing associations in the embeddings with census and other externally verifiable data. In *Quantifying Gender Stereotypes* and *Quantifying Ethnic Stereotypes* we apply the framework to quantify the change in stereotypes of women, men, and ethnic minorities. We further discuss our findings in *Discussion* and provide additional details in *Materials and Methods*.

Embedding Framework Overview and Validations

In this section, we briefly describe our methods and data and then validate our findings. We focus on showing that word embeddings are an effective tool to study historical biases and stereotypes by relating measurements from these embeddings to historical census and survey data. The consistent replication of such historical data, both in magnitude and in direction of biases, validates the use of embeddings in such work. This section extends the analysis of refs. 20 and 21 in showing that embeddings can also be used as a comparative tool over time as a consistent metric for various biases.

Summary of Data and Methods. We now briefly describe our datasets and methods, leaving details to *Materials and Methods* and *SI Appendix*, section A. All of our code and embeddings are available publicly[*]. For contemporary snapshot analysis, we use the standard Google News word2vec vectors trained on the Google News dataset (24, 25). For historical temporal analysis, we use previously trained Google Books/Corpus of Historical American English (COHA) embeddings, which are a set of nine embeddings, each trained on a decade in the 1900s, using the COHA and Google Books (26). As additional validation, we train, using the GLoVe algorithm (27), embeddings from the *New York Times* Annotated Corpus (28) for every year between 1988 and 2005. We then collate several word lists to represent each gender[†] (men, women) and ethnicity[‡] (White, Asian, and Hispanic), as well as neutral words (adjectives and occupations). For occupations, we use historical US census data (29) to extract the percentage of workers in each occupation that belong to each gender or ethnic group and compare it to the bias in the embeddings.

Using the embeddings and word lists, one can measure the strength of association (embedding bias) between neutral words and a group. As an example, we overview the steps we use to quantify the occupational embedding bias for women. We first compute the average embedding distance between words that represent women—e.g., she, female—and words for occupations—e.g., teacher, lawyer. For comparison, we also compute the average embedding distance between words that represent men and the same occupation words. A natural metric for the embedding bias is the average distance for women minus the average distance for men. If this value is negative, then the embedding more closely associates the occupations with men. More generally, we compute the representative group vector by taking the average of the vectors for each word in the given gender/ethnicity group. Then we compute the average Euclidean distance between each representative group vector and each vector in the neutral word list of interest, which could be occupations or adjectives. The difference of the average distances is our metric for bias—we call this the relative norm difference or simply embedding bias.

We use ordinary least-squares regressions to measure associations in our analysis. In this paper, we report r^2 and the coefficient P value for each regression, along with the intercept confidence interval when relevant.

Validation of the Embedding Bias. To verify that the bias in the embedding accurately reflects sociological trends, we compare the trends in the embeddings with quantifiable demographic trends in the occupation participation, as well as historical surveys of stereotypes. First, we use women and minority ethnic participation statistics (relative to men and Whites, respectively) in different occupations as a benchmark because it is an objective metric of social changes. We show that the embedding accurately captures both gender and ethnic occupation percentages and consistently reflects historical changes.

Next, we validate that the embeddings capture personality trait stereotypes. A difficulty in social science is the relative dearth of historical data to systematically quantify stereotypes, which highlights the value of our embedding framework as a quantitative tool but also makes it challenging to directly confirm our findings on adjectives. Nevertheless, we make use of the best available data from historical surveys, gender stereotypes from 1977 and 1990 (6, 7) and ethnic stereotypes from the Princeton trilogy from 1933, 1951, and 1969 (8–10).

Comparison with women's occupation participation. We investigate how the gender bias of occupations in the word embeddings relates to the empirical percentage of women in each of these occupations in the United States. Fig. 1 shows, for each occupation, the relationship between the relative percentage (of women) in the occupation in 2015 and the relative norm distance between words associated with women and men in the Google News embeddings. (Occupations whose 2015 percentage is not available, such as midwife, are omitted. We further note that the Google News embedding is trained on a corpus

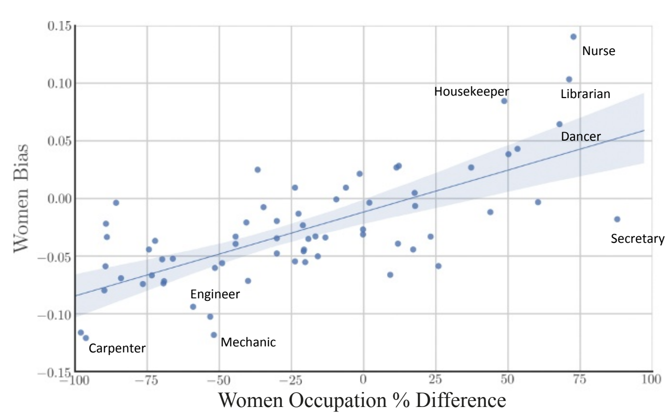

Fig. 1. Women's occupation relative percentage vs. embedding bias in Google News vectors. More positive indicates more associated with women on both axes. $P < 10^{-10}$, $r^2 = 0.499$. The shaded region is the 95% bootstrapped confidence interval of the regression line. In this single embedding, then, the association in the embedding effectively captures the percentage of women in an occupation.

[*]All of our own data and analysis tools are available on GitHub at https://github.com/nikhgarg/EmbeddingDynamicStereotypes. Census data are available through the Integrated Public Use Microdata Series (29). We link to the sources for each embedding used in *Materials and Methods*.

[†]There is an increasingly recognized difference between sex and gender and thus between the words male/female and man/woman, as well as nonbinary categories. We limit our analysis to the two major binary categories due to technical limitations, and we use male and female as part of the lists of words associated with men and women, respectively, when measuring gender associations. We also use results from refs. 6 and 7 which study stereotypes associated with sex.

[‡]When we refer to Whites or Asians, we specifically mean the non-Hispanic subpopulation. For each ethnicity, we generate a list of common last names among the group. Unfortunately, our present methods do not extend to Blacks due to large overlaps in common last names among Whites and Blacks in the United States.