Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, James Zou (2017)

Claudia Rebmann Mingyang He
Embeddings

Institut für Computerlinguistik
Ruprecht-Karls-Universität Heidelberg

16.07.2019

Stereotype

- Deutsche sind pünktlich

Stereotype

- Deutsche sind pünktlich
- Franzosen sind romantisch

Stereotype

- Deutsche sind pünktlich
- Franzosen sind romantisch
- Griechen sind faul und können nicht mit Geld umgehen

Überblick

Motivation

- Geschlechter- und ethnische Stereotype sind ein wichtiges Thema in vielen Disziplinen
- Die Sprachanalyse ist ein Standardwerkzeug zur Demonstration eines Stereotyps
- Frühere Studien: Nutzen in erster Linie menschliche Umfragen, Wörterbuch- und qualitative Analysen oder „in-depth knowledge" verschiedener Sprachen
- Diese Methoden erfordern oft eine zeitaufwendige und teure manuelle Analyse und lassen sich möglicherweise nicht einfach über Stereotypen, Zeiträume und Sprachen hinweg skalieren

Neue Methode

- NLP und Machine Learning
- Neueste Arbeiten im Bereich des maschinellen Lernens zeigen, dass Word Embedding auch Stereotype erfassen
- Bolukbasi et al.,2016; Caliskan, Bryson,and Narayanan,2017; Zhao et al.,2017; van Miltenburg,2016
- Honorable-Männer $\leftrightarrow$ Submissive-Frauen

Neuer Versuch

- Word Embeddings als quantitative Linse zur Untersuchung historischer Trends
- Systematisches Framework und Metriken zur Analyse von Word Embeddings, die in über 100 Jahren Textkorpora trainiert wurden
- Trends in Geschlechter- und ethnischen Stereotypen im 20. und 21. Jahrhundert in den Vereinigten Staaten.

Überblick

Embeddings

- Contemporary snapshot analysis: Google News word2vec Vectors trainiert auf dem Google News Dataset [1] [2]
- Historical temporal analysis: vortrainierte Google Books/COHA embeddings [3]
- zusätzliche Validierung: New York Times Annotated Corpus mit GLoVe-Algorithmus für jedes Jahr zwischen 1988 und 2005 [4]

---

[1] Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient estimation of word representations in vector space.arXiv preprint arXiv:1301.3781

[2] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed representations of words andphrases and their compositionality. InAdvances in neural information processing systems, 3111–3119

[3] Davies, M. 2010. The 400 million word corpus of historical American English (18102009)

[4] Sandhaus, E. 2008. The New York Times Annotated Corpus

Überblick

Wortlisten

- jedes Geschlecht (Männer, Frauen)

Wortlisten

- jedes Geschlecht (Männer, Frauen)
- Für Männer: he, son, his, him, father, man, boy, himself

Wortlisten

- jedes Geschlecht (Männer, Frauen)
- Für Männer: he, son, his, him, father, man, boy, himself
- Für Frauen: z.B. she, daughter, hers, her, mother, woman, girl

Wortlisten

- jede ethnische Zugehörigkeit (Weiße, Asiaten und Spanier [5]) [6]

[5] gemeint sind immer Spanier und Lateinamerikaner

[6] available https://raw.githubusercontent.com/fivethirtyeight/data/master/most-common-name/surnames.csv

Wortlisten

- jede ethnische Zugehörigkeit (Weiße, Asiaten und Spanier [5]) [6]
- Weiße Nachnamen: harris, nelson, robinson, thompson, moore, wright, anderson

[5] gemeint sind immer Spanier und Lateinamerikaner

[6] available https://raw.githubusercontent.com/fivethirtyeight/data/master/most-common-name/surnames.csv

Wortlisten

- jede ethnische Zugehörigkeit (Weiße, Asiaten und Spanier [5]) [6]
- Weiße Nachnamen: harris, nelson, robinson, thompson, moore, wright, anderson
- Spanische Nachnamen: ruiz, alvarez, vargas, castillo, gomez, soto

[5] gemeint sind immer Spanier und Lateinamerikaner

[6] available https://raw.githubusercontent.com/fivethirtyeight/data/master/most-common-name/surnames.csv

Wortlisten

- jede ethnische Zugehörigkeit (Weiße, Asiaten und Spanier [5]) [6]
- Weiße Nachnamen: harris, nelson, robinson, thompson, moore, wright, anderson
- Spanische Nachnamen: ruiz, alvarez, vargas, castillo, gomez, soto
- Asiatische Nachnamen: cho, wong, tang, huang, chu, chung, ng, wu, liu

[5] gemeint sind immer Spanier und Lateinamerikaner

[6] available https://raw.githubusercontent.com/fivethirtyeight/data/master/most-common-name/surnames.csv

Wortlisten

- Beschäftigungen: janitor, statistician, midwife, bailiff, auctioneer, photographer, geologist, shoemaker, athlete, cashier, dancer, housekeeper
- Adjektive (Williams and Best,1977,1990): headstrong, thankless, tactful, distrustful, quarrelsome, effeminate, ckle, talkative, dependable, resentful, sarcastic
- Auch Teilmenge von diesen neutralen Wörtern: professionelle Berufe, intellektuelle Adjektive[7], Adjektive zu physischem Aussehen [8]

[7] mostly from https://www.e-education.psu.edu/writingrecommendationlettersonline/node/151,https://www.macmillandictionary.com/us/thesaurus-category/american/words-used-to-describe-intelligent-or-wise-people

[8] mostly from http://usefulenglish.ru/vocabulary/appearance-and-character, http://www. sightwordsgame.com/parts-of-speech/adjectives/appearance/, http://www.stgeorges.co.uk/blog/physical-appearance-adjectives-the-bald-and-the-beautiful

Überblick

Word Similarity

- Wenn zwei Vektoren gegeben sind, kann ihre Ähnlichkeit entweder durch die negative Differenznorm oder die Konsinus-Ähnlichkeit gemessen werden
- neg-norm-dif$(u, v) = -||u - v||_2$
- cos-sim$(u, v) = u \cdot v$

Bias in den Embeddings

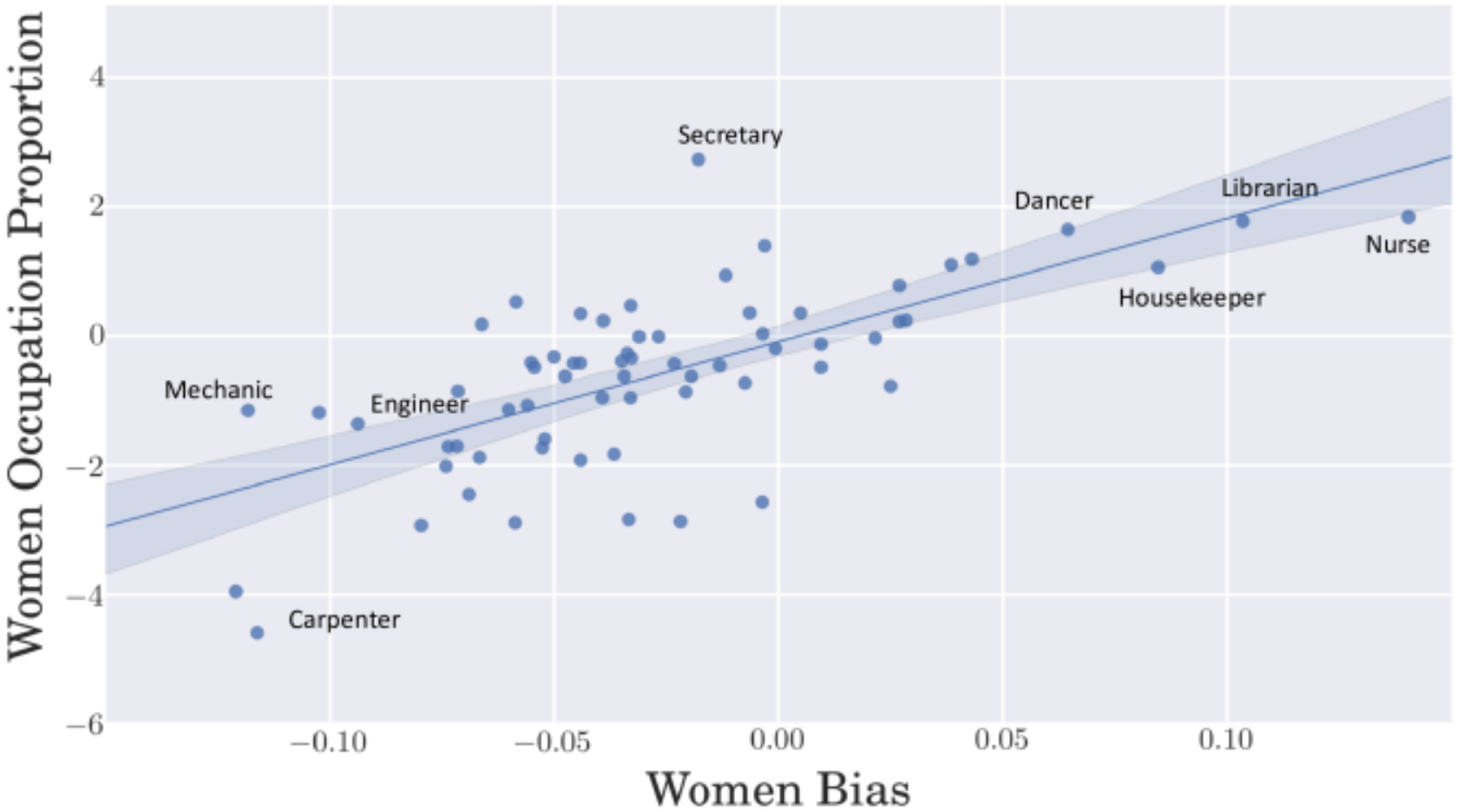
- Bias in dem Embedding: Relative norm difference
- Repräsentativer Gruppenvektor: Der Durchschnitt der Vektoren für jedes Wort in der gegebenen Geschlecht- / Ethnizitätsgruppe;
- Die durchschnittliche L2-Norm der Differenzen zwischen jedem repräsentativen Gruppenvektor und jedem Vektor in der neutralen Wortliste wird berechnet
- Die relative Normdifferenz ist die Differenz der durchschnittlichen L2-Normen

Bias in the embeddings

- Relative norm distance = $\sum_{v_m \in M} ||v_m - v_1||_2 - ||v_m - v_2||_2$
- M ist eine Menge neutraler Wortvektoren
- v1 ist der Durchschnittsvektor für Gruppe eins
- v2 ist der Durchschnittsvektor für Gruppe zwei

Überblick

Beschäftigung Bias



Woman occupation proportion vs embedding bias in Google News vectors. More positive indicates more women biased on both axes. $p < 10^{-9}$ r-squared=0.462

Beschäftigung Bias

- Historical U.S. census data [9] vs Word Embedding

[9] Steven Ruggles; Katie Genadek; Ronald Goeken; Josiah Grover; and Matthew Sobek. 2015. Integrated Public UseMicrodata Series: Version 6.0 [dataset]

[10] where $p = \%$ of woman in occupation

Beschäftigung Bias

- Historical U.S. census data [9] vs Word Embedding
- log-prop(p) = $log\frac{p}{1-p}$ [10]

[9] Steven Ruggles; Katie Genadek; Ronald Goeken; Josiah Grover; and Matthew Sobek. 2015. Integrated Public UseMicrodata Series: Version 6.0 [dataset]

[10] where $p = \%$ of woman in occupation

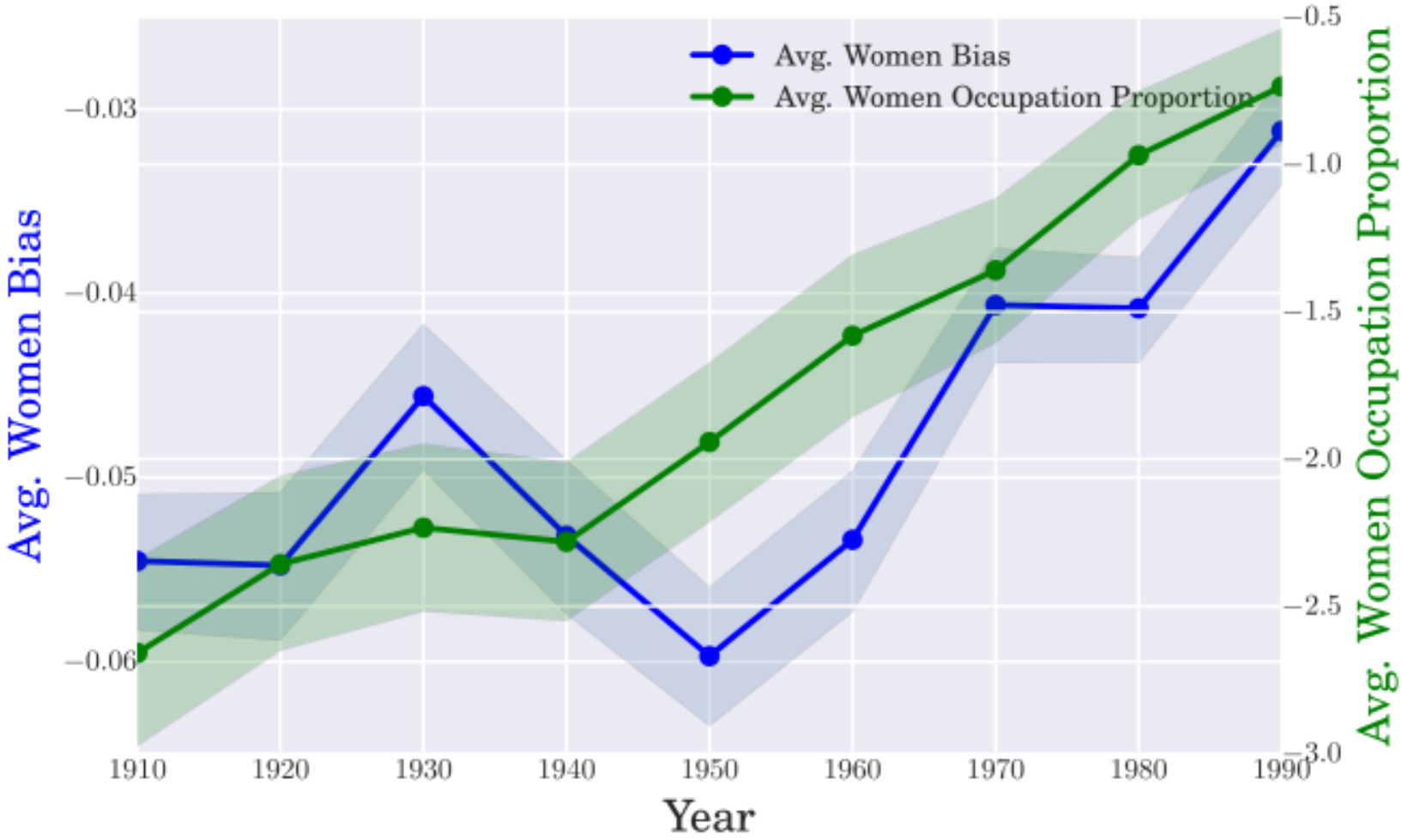
Beschäftigung Bias

- Historical U.S. census data [9] vs Word Embedding
- log-prop(p) = $log\frac{p}{1-p}$ [10]
- Regression durch (0,0): Beschäftigungen, deren geschlechtsspezifische Beteiligung genau mittig (50:50) liegt, weisen keinen messbare Embedding Bias auf

[9] Steven Ruggles; Katie Genadek; Ronald Goeken; Josiah Grover; and Matthew Sobek. 2015. Integrated Public UseMicrodata Series: Version 6.0 [dataset]

[10] where p = % of woman in occupation

Beschäftigung Bias



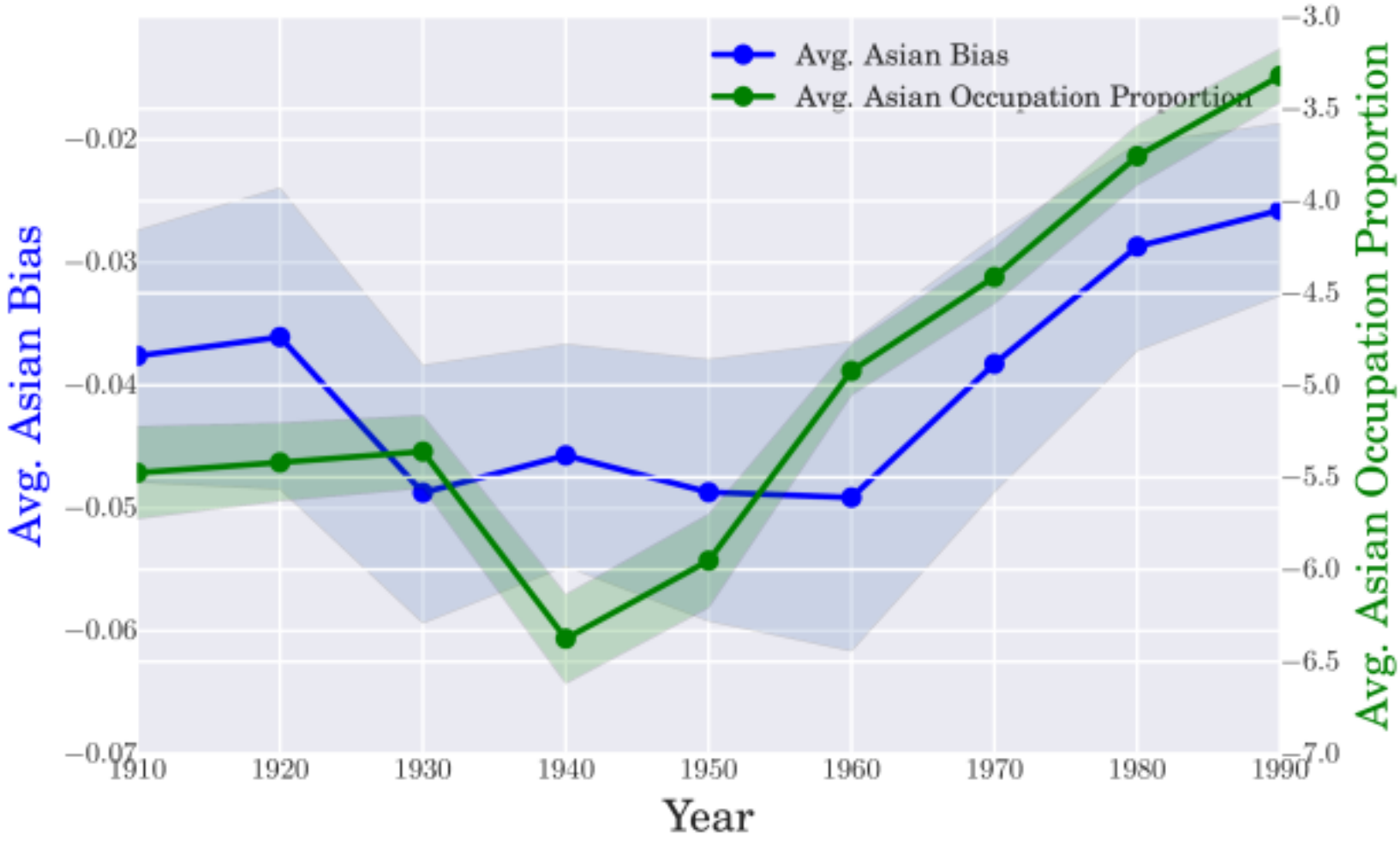
Average gender bias score over time in COHA embeddings in occupations vs the average log proportion. In blue is relative women bias in the embeddings, and in green is the average log proportion of women in the same occupations.

Beschäftigung Bias

Hispanic	Asian	White
housekeeper	professor	smith
mason	official	blacksmith
artist	secretary	surveyor
janitor	conductor	sheriff
dancer	physicist	weaver
mechanic	scientist	administrator
photographer	chemist	mason
baker	tailor	statistician
cashier	accountant	clergy
driver	engineer	photographer

(c) The top ten occupations most closely associated with each ethnic group in the Google News embedding.

Beschäftigung Bias



Average ethnic (Asian vs White) bias score over time for occupations in COHA (blue) vs the average conditional log proportion (green).

Beschäftigung Bias

- cond-log-prop(group 1,group 2) $= log \frac{p}{1-p}$ [11]

[11] where $p = \frac{\%\ of\ group\ 1}{\%\ of\ group\ 1\ +\ \%\ of\ group\ 2}$

Überblick

Adjektive

- Wie hat sich die Darstellung von Frauen über die Jahre verändert?

[12]Williams, J. E., and Best, D. L. 1977. Sex Stereotypes and Trait Favorability on the Adjective Check List.Educationaland Psychological Measurement37(1):101–110

[13]Williams, J. E., and Best, D. L. 1990.Measuring sex stereotypes: A multination study, Rev. Sage Publications, Inc

Adjektive

- Wie hat sich die Darstellung von Frauen über die Jahre verändert?
  $\rightarrow$ Adjektive

[12]Williams, J. E., and Best, D. L. 1977. Sex Stereotypes and Trait Favorability on the Adjective Check List.Educationaland Psychological Measurement37(1):101–110

[13]Williams, J. E., and Best, D. L. 1990.Measuring sex stereotypes: A multination study, Rev. Sage Publications, Inc

Adjektive

- Wie hat sich die Darstellung von Frauen über die Jahre verändert?
  $\rightarrow$ Adjektive
- wenige systematische und quantitative Metriken für Adjektiv Bias in der Literatur

[12]Williams, J. E., and Best, D. L. 1977. Sex Stereotypes and Trait Favorability on the Adjective Check List.Educationaland Psychological Measurement37(1):101–110

[13]Williams, J. E., and Best, D. L. 1990.Measuring sex stereotypes: A multination study, Rev. Sage Publications, Inc

Adjektive

- Wie hat sich die Darstellung von Frauen über die Jahre verändert?
  $\rightarrow$ Adjektive
- wenige systematische und quantitative Metriken für Adjektiv Bias in der Literatur
- Set mit 230 Adjektiven von Menschen nach Geschlechterstereotypen annotiert [12] [13]

[12] Williams, J. E., and Best, D. L. 1977. Sex Stereotypes and Trait Favorability on the Adjective Check List.Educationaland Psychological Measurement37(1):101–110

[13] Williams, J. E., and Best, D. L. 1990.Measuring sex stereotypes: A multination study, Rev. Sage Publications, Inc

Adjektive

- Wie hat sich die Darstellung von Frauen über die Jahre verändert? $\rightarrow$ Adjektive
- wenige systematische und quantitative Metriken für Adjektiv Bias in der Literatur
- Set mit 230 Adjektiven von Menschen nach Geschlechterstereotypen annotiert [12] [13]
- Korrelation mit Embedding Bias ($p < .0002$)

[12] Williams, J. E., and Best, D. L. 1977. Sex Stereotypes and Trait Favorability on the Adjective Check List.Educationaland Psychological Measurement37(1):101–110

[13] Williams, J. E., and Best, D. L. 1990.Measuring sex stereotypes: A multination study, Rev. Sage Publications, Inc

Adjektive

- Wie hat sich die Darstellung von Frauen über die Jahre verändert? $\rightarrow$ Adjektive
- wenige systematische und quantitative Metriken für Adjektiv Bias in der Literatur
- Set mit 230 Adjektiven von Menschen nach Geschlechterstereotypen annotiert [12] [13]
- Korrelation mit Embedding Bias ($p < .0002$)
- geschlechtsneutrale Adjektive $\rightarrow$ unbiased

[12] Williams, J. E., and Best, D. L. 1977. Sex Stereotypes and Trait Favorability on the Adjective Check List.Educationaland Psychological Measurement37(1):101–110

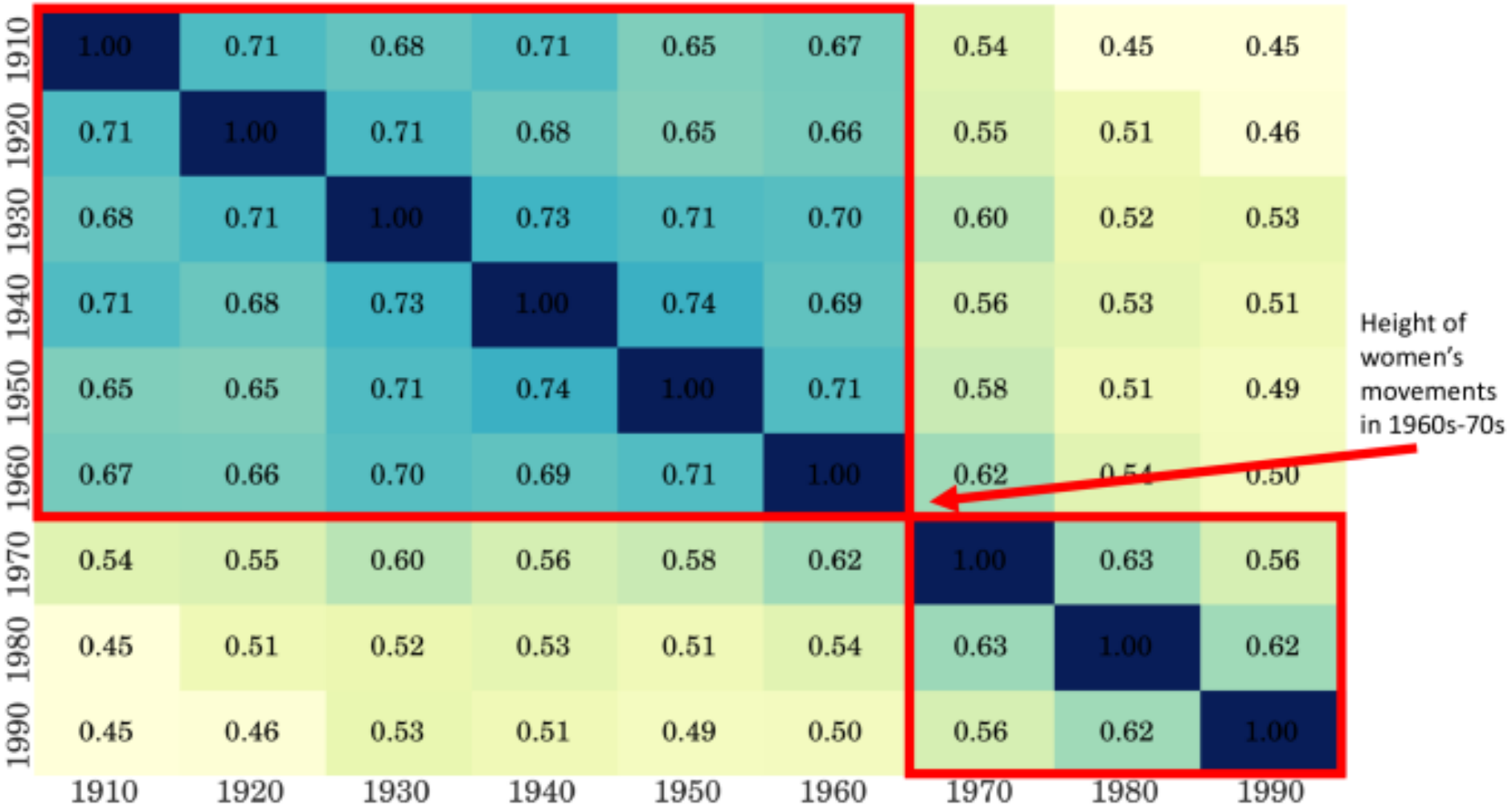
[13] Williams, J. E., and Best, D. L. 1990.Measuring sex stereotypes: A multination study, Rev. Sage Publications, Inc

Frauen und Adjektive

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

Top adjectives associated with women in 1910, 1950, and1990 by relative norm difference in the COHA embedding.

Frauen und Adjektive



Pearson correlation in embedding bias scores for adjectives over time between embeddings for each decade. The phase shiftin the 1960s-70s corresponds to the U.S. women's movement.

Beobachtungen

Teilmengen von Adjektiven:

- Intelligenz (intelligent, logical, thoughtful...)
  - Assoziation mit Frauen steigt
  - starker positiver Trend nach den 1960ern
- Aussehen (attractive, ugly, fashionable...)
  - keine signifikante Veränderung des Bias

Beobachtungen

Teilmengen von Adjektiven:

- Intelligenz (intelligent, logical, thoughtful...)
  - Assoziation mit Frauen steigt
  - starker positiver Trend nach den 1960ern
- Aussehen (attractive, ugly, fashionable...)
  - keine signifikante Veränderung des Bias

Individuelle Adjektive:

- hysterisch
  - bis Anfang 20. Jahrhundert psychischen Erkrankung von Frauen
  - 1920: Top 5 woman-biased
  - 1990: nicht in Top 100
- emotional
  - Assoziation mit Frauen steigt
  - spiegelt aktuellen Stand wider

Asiaten und Adjektive

1910	1950	1990
irresponsible	disorganized	inhibited
envious	outrageous	passive
barbaric	pompous	dissolute
aggressive	unstable	haughty
transparent	effeminate	complacent
monstrous	unprincipled	forceful
hateful	venomous	fixed
cruel	disobedient	active
greedy	predatory	sensitive
bizarre	boisterous	hearty

Top Asian (vs White) Adjectives in 1910, 1950, and 1990by relative norm difference in the COHA embedding.

Asiaten und Adjektive

Pearson correlation in embedding Asian bias scores for adjectives over time between embeddings for each decade.

Ethnische und Kulturelle Stereotype

- vor 1950: stark abwertende Adjektive, Beschreibung von Außenseitern
- ab 1950 und besonders ab 1980: Stereotype von heutigen Asian-Americans

Ethnische und Kulturelle Stereotype

- vor 1950: stark abwertende Adjektive, Beschreibung von Außenseitern
- ab 1950 und besonders ab 1980: Stereotype von heutigen Asian-Americans

- Russische Namen:
  - 1910-1920er: Russische Revolution $\rightarrow$ schwacher Wandel
  - 1950er: Kalter Krieg $\rightarrow$ starker Wandel
- Spanische Namen:
  - stetiger Wandel
  - kein großes Ereigniss, eher viele kleine

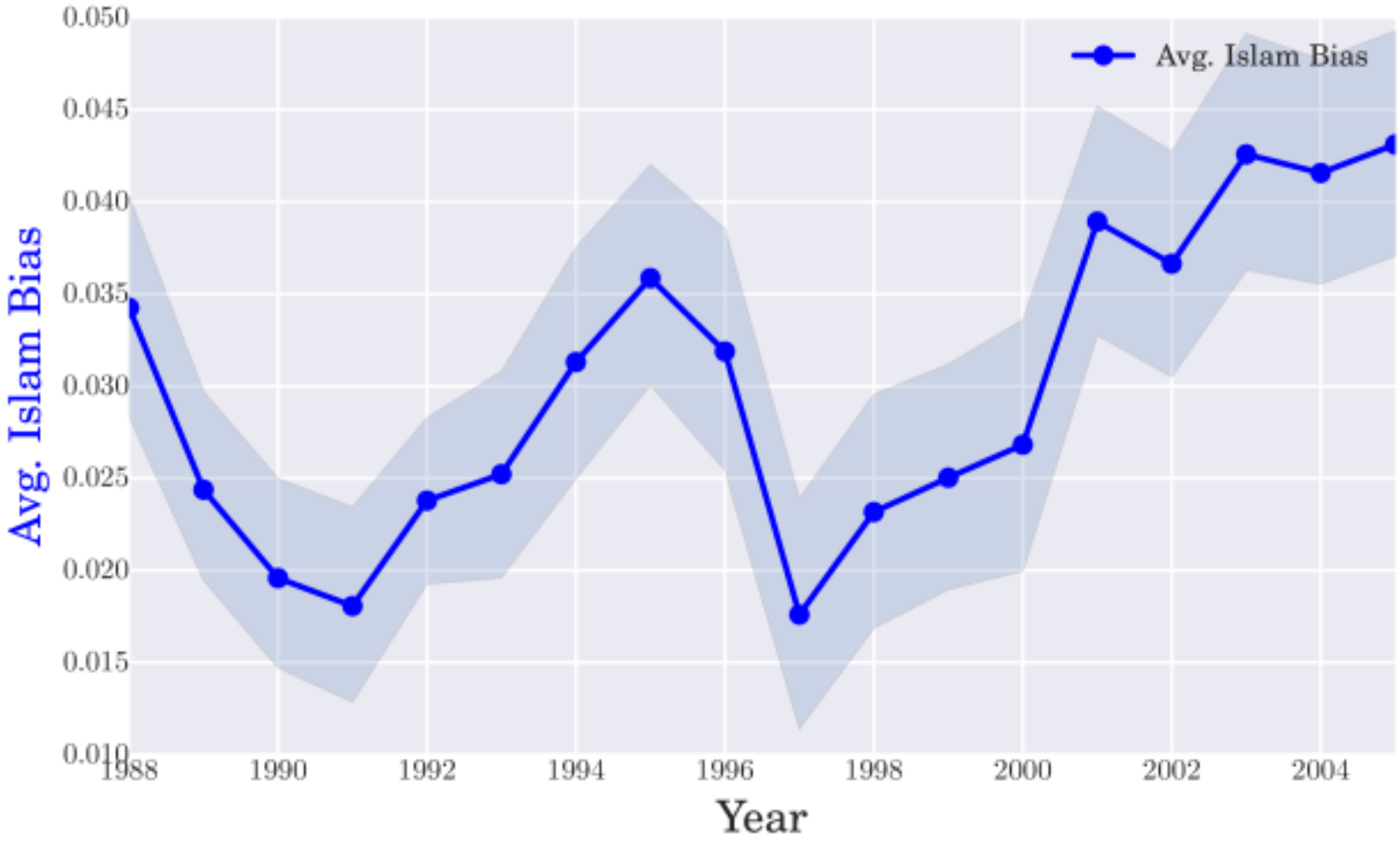
Ethnische und Kulturelle Stereotype

- vor 1950: stark abwertende Adjektive, Beschreibung von Außenseitern
- ab 1950 und besonders ab 1980: Stereotype von heutigen Asian-Americans

- Russische Namen:
  - 1910-1920er: Russische Revolution $\rightarrow$ schwacher Wandel
  - 1950er: Kalter Krieg $\rightarrow$ starker Wandel
- Spanische Namen:
  - stetiger Wandel
  - kein großes Ereigniss, eher viele kleine

$\Longrightarrow$ Embedding Bias beinhaltet Informationen über die Haltung gegenüber ethnischen Gruppen, insbesondere rund um globale Ereignisse

Ethnische und Kulturelle Stereotype



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 year windows, so, for example, 2000 contains data from 1999-2001.

Überblick

Zusammenfassung

- Vergleich von Word Embeddings mit dem demographischen Wandel im Bezug auf Geschlechter- und ethnische Stereotypen
- Quantifizierung eines *embedding biases* für Beschäftigungen und Adjektive

Zusammenfassung

- Vergleich von Word Embeddings mit dem demographischen Wandel im Bezug auf Geschlechter- und ethnische Stereotypen
- Quantifizierung eines *embedding biases* für Beschäftigungen und Adjektive
- Ergebnisse:
  - vorhergesagte Beschäftigungen folgen der Realität
  - Adjektive zeigen wie verschiedene Personengruppen über die Zeit betrachtet werden

Herausforderungen und Ausblick

- Robustheit abhängig von Daten und Metriken
  (Alternative Metriken: Caliskan, Bryson, und Narayanan (2017) und Bolukbasi et al. (2016))
- Abhängigkeit von Wortlisten
  $\rightarrow$ Vergleich Beschäftigungen vs. professionelle Beschäftigungen
  $\rightarrow$ verschiedene Adjektiv-Listen
- geschriebene Texte können die soziale Haltung nicht komplett reflektieren
- Dimensionen der Embeddings haben keine Bedeutung
  (Besser: Rothe and Schtze, 2016)
- separate Embeddings pro Jahrzehnt
  (Vereint: Rudolph et al., 2017; Rudolph und Blei, 2017)

Kritik

- Wortlisten:
  - Asiaten = Chinesen ?
  - keine Variationen der Gruppen-Wortlisten
  - Bedeutung/Einfluss spezifischer Worte
- Wandel des Embedding Bias ohne globale Ereignisse
- Abweichungen in den Ergebnissen bei anderer Textgrundlage (z.B. Wikipedia GloVe)

Diskussion

Vielen Dank für eure Aufmerksamkeit!