Relating Word Embedding Gender Biases to Gender Gaps: A Cross-Cultural Analysis

Scott Friedman, Sonja Schmer-Galunder, Anthony Chen, and Jeffrey Rye
SIFT, Minneapolis, MN USA
{friedman, sgalunder, achen, rye}@sift.net

Abstract

Modern models for common NLP tasks often employ machine learning techniques and train on journalistic, social media, or other culturally-derived text. These have recently been scrutinized for racial and gender biases, rooting from inherent bias in their training text. These biases are often sub-optimal and recent work poses methods to rectify them; however, these biases may shed light on actual racial or gender gaps in the culture(s) that produced the training text, thereby helping us understand cultural context through big data. This paper presents an approach for quantifying gender bias in word embeddings, and then using them to characterize statistical gender gaps in education, politics, economics, and health. We validate these metrics on 2018 Twitter data spanning 51 U.S. regions and 99 countries. We correlate state and country word embedding biases with 18 international and 5 U.S.-based statistical gender gaps, characterizing regularities and predictive strength.

1 Introduction

Machine-learned models are the *de facto* method for NLP tasks. Recently, machine-learned models that utilize *word embeddings* (i.e., vector-based representations of word semantics) have come under scrutiny for biases and stereotypes, e.g., in race and gender, arising primarily from biases in their training data (Bolukbasi et al., 2016). These biases produce systematic mistakes, so recent work has developed *debiasing* language models to improve NLP models' accuracy and and remove stereotypes (Zhao et al., 2018; Zhang et al., 2018).

Concurrently, other research has begun to characterize how biases in language models correspond to disparities in the cultures that produced the training text, e.g., by mapping embeddings to survey data (Kozlowski et al., 2018), casting analogies in the vector space to compute that "man is to woman as doctor is to nurse" (Bolukbasi et al., 2016), or varying the training text over decades and mapping each decade's model bias against its statistical disparities to capture periods of societal shifts (Garg et al., 2018).

Building on previous work, this paper presents initial work characterizing word embedding biases with statistical *gender gaps* (i.e., discrepancies in opportunities and status across genders). This is an important step in approximating cultural attitudes and relating them to cultural behaviors. We analyze 51 U.S. states and 99 countries, by (1) training separate word embeddings for each of these cultures from Twitter and (2) correlating the biases in these word embeddings with 5 U.S.-based and 18 international gender gap statistics.

Our claims are as follows: (1) some cultural gender biases in language are associated with gender gaps; (2) we can characterize biases based on strength and direction of correlation with gender gaps; and (3) themed word sets, representative of values and social constructs, capture different dimensions of gender bias and gender gaps.

We continue with a brief overview of gender gaps (Sec. 2) and then a description of our training data (Sec. 3) and four experiments (Sec. 4). We close with a discussion of the above claims and future work (Sec. 5).

2 Gender Gaps and Statistics

Within the social sciences, anthropologists often attempt to explain the asymmetrical valuations of the sexes across a range of cultures with respect to patterns of social and cultural experience (Rosaldo, 1974). This work contributes to this research by updating traditional qualitative approaches with computational methods.

The public sphere is often associated with male and agents traits (assertiveness, competitiveness)

Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing, pages 18–24
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

in domains like politics and executive rules at work. Private or domestic domains linked to family and social relationships are traditionally related to women, although social relationships are considered more important by people independent of gender (Friedman and Greenhaus, 2000). Gender gaps arise form these asymmetrical valuations, e.g., where men are typically over-represented and have higher salaries compared to women (Mitra, 2003; Vincent, 2013; Bishu and Alkadry, 2017).

We utilize diverse gender gap statistics in this work. For international data, we use 18 gender gap metrics comprising the Global Gender Gap Index (GGGI) originally compiled for the World Economic Forum's 2018 Gender Gap Report.[1] The GGGI measures clearly-defined dimensions for which reliable data in most countries was available (Hawken and Munck, 2013). For domestic data, we use a 2018 report from the U.S. Center for Disease Control (CDC) on male and female exercise rate (Blackwell and Clarke, 2018), wage gap and workforce data published by the U.S. Census Bureau in 2016, female percentages of math and computer science degrees from Society of Women Engineers,[2] and female percentages of each state's legislators from Represent Women's 2018 Gender Parity Report.[3]

3 Training Data

Our training data include public tweets from U.S. and international Twitter users over 100 days throughout 2018, including the first ten days of each of the first ten months. We use tweet's location property to categorize by location, and we include only English tweets in our dataset.

We filtered out all tweets with fewer than three words, and following other Twitter-based embedding strategies (e.g. Li et al., 2017), we replaced URLs, user names, hashtags, images, and emojis with other tokens. We divided the processed tweets into two separate datasets: (1) U.S. states and (2) countries. This helps us validate our approach with multiple granularities and datasets.

The international dataset contains 99 countries with varying number of tweets, ranging from 98K tweets (Mauritius) to 122M tweets (U.K). The U.S. states dataset contains 51 regions (50 states and Washington, D.C.) ranging from 450K tweets (Wyoming) to 65M tweets (California). For both datasets, we sampled 10 million tweets for all cultures that exceeded that number. These corpora are orders of magnitude smaller than other approaches for tweet embeddings (e.g., Li et al., 2017).

We use Word2Vec to construct word vectors for our experiments, but we compare Word2Vec with other algorithms in our analyses (Sec. 4.3).

4 Experiments

4.1 International Analysis

Our international and U.S.-based analyses have an identical experimental setup, varying only in the gender gap statistics and the word embeddings.

Our materials included word-sets based in part on survey data (Williams and Best, 1990) and recent work on word embeddings (Garg et al., 2018). These word-sets included (1) *female words* including female pronouns and nouns, (2) *male words*, including male pronouns and nouns, and (3) *neutral words* that were grouped thematically. For instance, we used *appearance* and *intellect* adjectives from (Garg et al., 2018), and we generated other thematic word sets representative of social constructs: *government* (democrat, republican, senate, government, politics, minister, presidency, vote, parliament, ...), *threat* (dangerous, scary, toxic, suspicious, threat, frightening ...), *communal* (community, society, humanity, welfare, ...), *criminal* (criminal, jail, prison, crime, corrupt, ...), *childcare* (child, children, parent, baby, nanny, ...), *excellent* (excellent, fantastic, phenomenal, outstanding, ...) and others.

We use the same male and female word sets for international and U.S. state analyses, and we compute per-gender vectors $\overrightarrow{female}$ and $\overrightarrow{male}$ by averaging the vectors of each constituent word, following (Garg et al., 2018). For any country or state's word embedding, we compute the *average axis projection* of a neutral word set W onto the male-female axis as:

$$avg_{w \in W}\left(\overrightarrow{w} \cdot \frac{\overrightarrow{female} - \overrightarrow{male}}{||\overrightarrow{female} - \overrightarrow{male}||_2}\right) \quad (1)$$

This average axis projection is our primary measure of gender bias in word embeddings.

For any neutral word list (e.g., government terms), we compute the average axis projection for all countries (or states) and compute its correlation to international (or U.S.) gender gaps. Fig. 2

[1] http://reports.weforum.org
[2] http://societyofwomenengineers.swe.org/
[3] http://www.representwomen.org

	govt	intellect	workplace	excellent	childcare	illness	communal	victim	*"pretty"*	r-1	r-2	r-3	r-4
Index: Overall Gender Gap	**.30**	.11	.17	.12	-.01	-.07	-.20	-.06	-.19	-.01	.02	.03	.02
Sex ratio at birth	.00	.01	.03	.00	-.02	-.01	.00	-.02	.00	.00	.03	-.04	.00
Index: Educational Attainment	.03	.05	.10	.03	-.18	-.12	-.19	**-.23**	-.07	-.04	.02	.00	.00
Literacy rate	.07	.08	.05	.07	-.18	-.13	-.21	**-.23**	-.08	-.03	.02	-.05	-.07
Enrollment tertiary education	.06	.10	.07	.02	**-.24**	-.20	-.12	-.11	-.06	-.01	-.01	-.08	-.03
Enrollment secondary education	.02	.01	.03	.00	-.08	-.01	**-.21**	-.13	-.02	.00	.04	-.02	-.01
Enrollment primary education	.01	.01	.05	.01	-.05	-.06	**-.15**	-.12	-.06	-.01	.04	.01	.03
Index: Political Empowerment	**.28**	.02	.10	.01	.04	.01	-.03	.04	-.14	.07	.01	-.04	.00
Women in ministerial positions	**.25**	.03	.17	.09	.00	-.02	-.02	.00	-.08	.04	.03	-.01	-.07
Women in parliament	**.16**	.04	.07	.02	.01	.03	-.02	.05	-.06	.02	.06	-.09	.00
% 50 years female head of state	.05	-.04	.01	-.07	.02	.01	.00	.00	-.06	.01	-.02	-.01	.01
Index: Economic Participation	.10	.04	.13	.10	-.02	-.12	**-.33**	-.09	-.10	-.07	.04	-.05	.04
Professional and technical workers	.20	.23	**.27**	.23	-.15	-.21	-.19	-.12	-.14	-.05	.08	-.08	.06
Legislators, officials, managers	.08	.15	.10	**.18**	-.05	-.14	**-.18**	-.10	-.01	-.05	.04	.00	.03
Labour force participation	.03	.04	.08	.02	-.03	-.09	**-.21**	-.04	-.14	-.09	.02	-.03	.01
Wage equality (survey)	-.02	-.04	-.02	-.02	.03	.04	.00	.00	-.02	.00	-.02	-.03	-.05
Index: Health and Survival	.03	.09	.08	.06	-.12	**-.13**	-.02	-.06	.00	-.02	.09	.01	.00
Healthy life expectancy	.06	.09	.12	.14	-.16	**-.31**	-.07	-.11	-.01	-.02	.07	.01	-.01

Figure 1: Correlation of themed neutral word sets' gender bias (columns) against categories of gender gaps from worldbank.org (rows). Values are R^2 coefficient of determination, where negation is added to indicate inverse correlation. The rightmost four word sets (*r-1 to r-4*) were randomly sampled from the vocabulary for comparison.

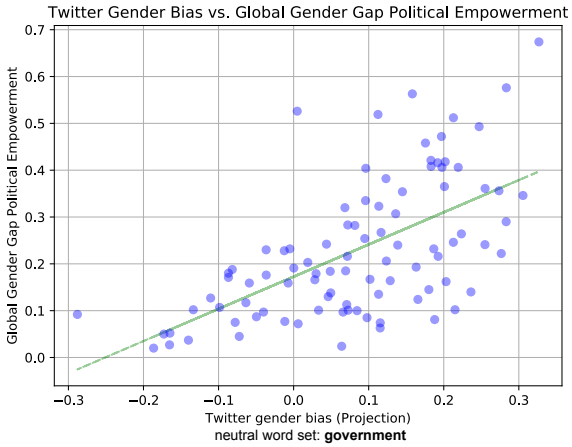

Figure 2: Correlation of country's gender bias of government words (x-axis; female association increases in positive direction) against the World Economic Forum's political empowerment gender gap index (y-axis; gender gap decreases in positive direction).

plots each country's government/political word bias against the World Economic Forum's Political Empowerment Gender Gap sub-index (from 0 to 1, where greater score indicates less gap). The value 0.0 on the x-axis indicates no gender bias, and female bias increases along the x-axis.

Consequently, Fig. 2 is consistent with the hypothesis that— globally, over our set of 99 countries— women's political influence and power increase (relative to men) as political language shows a more female bias.

We present results of each thematic word set regressed against all available international statistics. For each pair of themed word set and gender gap statistic, the algorithm (1) performs feature selection on 20% of the countries to optionally down-select from the set of words in the themed word set, (2) uses the down-selected word set to compute the R^2 determination against the full set of countries, and then (3) repeats a total of five times and averages the answers. Feature selection monotonically increases the R^2, and using 20% of countries helps prevent over-fitting.

Fig. 1 includes our results over this analysis, grouping gender gap sub-indices (bold) with their related statistics. This illustrates that different word sets vary in their correlation direction and strength across different statistic groups: the *political* set is positively correlated with the political empowerment subgroup and marginal on some economic statistics, but weak over health and education; intellectual and workplace terms positively correlate with economic statistics but are weak predictors otherwise; *illness* terms indirectly correlated with health and survival statistics, but are weak correlates elsewhere; and so-forth. The word *"pretty,"* shown in Fig. 1, was the single word with the strongest determination against the overall gender gap and other sub-indices. Fig. 1 also includes four randomly-generated word sets, which do not exceed $R^2 = 0.09$ for any gender gap.

The selective correlation of these thematic word sets with related gender gap statistics supports our claim that gender biases in word embeddings can help characterize and predict statistical gender gaps across cultures. Since we trained our embeddings on tweets alone— with as few as 98K tweets

	threat	unintelligent	criminal	persistent	excellent	stem-alum	childcare	victim	appearance	r-1	r-2	r-3	r-4
CDC Activity Proportion	.09	.05	-.03	**.41**	.07	-.04	-.02	.03	-.03	-.01	.07	.00	.00
Female State Legislators	-.16	-.22	**-.42**	.11	.11	.03	-.24	-.04	-.12	.00	-.08	-.03	-.06
Math & CS Degrees	-.15	-.27	-.02	-.07	.01	**.28**	.01	-.09	-.07	-.03	-.01	.00	-.01
Census Wage Gap	**-.51**	-.15	-.12	.04	.21	.11	-.06	-.17	-.15	.01	-.06	.02	.04
Census Workforce Ratio	-.06	**-.30**	-.06	-.04	.00	.03	.01	-.03	-.06	.01	-.03	-.03	-.04

Figure 3: Correlation of themed neutral word sets' gender bias (columns) against U.S. gender disparity statistics from CDC, US Census Bureau, and Represent Women's 2018 Gender Parity Report (rows). Values are R^2 coefficient of determination, where negation is added to indicate inverse correlation. The rightmost four word sets (*rand 1-4*) were randomly sampled from the vocabulary for comparison.

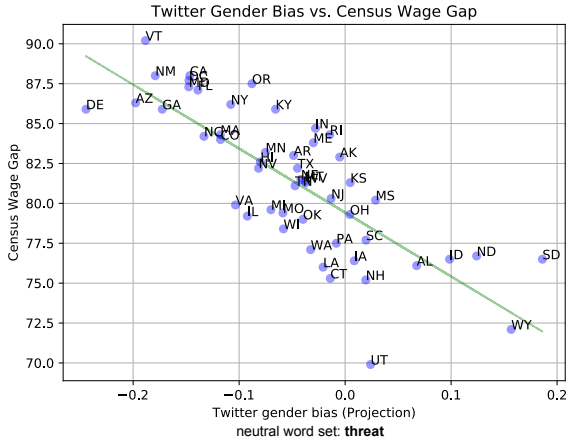

Figure 4: Correlation of states' gender bias of threat words' (x-axis; female association is positive direction) against the pay gap reported by U.S. Census Bureau in 2016 (y-axis; pay gap decreases in positive direction).

for some countries— this also supports our claim that social media is a plausible source to compute a culture's gender bias in language.

None of our themed word sets strongly correlated with: (1) sex ratio at birth, which was 1.0 for the vast majority of countries; (2) percentage of last 50 years with female head of state; and (3) survey-based wage equality. The latter two gender gaps may correlate with other themed word sets, or they may have a more complex or nonlinear relationship to a culture's gender bias in language.

4.2 U.S. State Analysis

Our analysis of 51 U.S. regions (50 U.S. states and Washington, D.C.) is analogous to our Sec. 4.1 international analysis; we only vary the word embeddings and the statistical gender gap data.

Fig. 4 shows an example of indirect correlation ($R^2 = 0.51$) of our *threat* word set (threat, dangerous, toxic, suspicious, scary, frightening, horrifying, ...) against U.S. Census Bureau data reported in 2016 on the gender pay gap. The y-axis indicates cents on the dollar earned by women for the same work as men, ranging from 69.9¢ (UT) to 91.2¢ (VT). This inverse correlation is consistent with the hypothesis that when masculinity is threatened in some cultures, men react by asserting dominance (Zuo and Tang, 2000; Schmitt and Branscombe, 2001).

Fig. 3 illustrates different word sets' determination on U.S. regions' statistical gender gaps. The word set describing persistence and devotion had strongest direct correlation with reduced gender gap in exercise. The word set for criminal behavior had strongest negative correlation with female proportion of state legislators. Words for STEM disciplines and alumni directly correlated with increased percentages of female math and CS degrees. Threat-based words negatively correlated with pay equality, and words for unintelligent and inept negatively correlated with female percentage of the workforce. Other word sets from the international analysis (e.g., childcare and victimhood) had less determination of gender gaps than in the international setting.

As with our international analysis, this domestic analysis supports our claim that gender biases in cultural language models can predict and characterize statistical gender gaps.

4.3 Algorithm Comparison

We compare four word embedding algorithms and three bias metrics using our gender gap statistics and word sets. We compare four algorithms: (1) GloVe, (2) Word2Vec (skip-gram), (3) CBOW Word2Vec, and (4) FastText (skip-gram). For each algorithm we utilize a window size 10, filter words that occur fewer than 5 times, and produce 200-dimension output vectors.

GloVe (Pennington et al., 2014) uses count-based vectorization to reduce dimensionality by minimizing reconstruction loss. The dot product of two GloVe vectors equals the log of the number of times those two words occur near each other.

Word2Vec (Mikolov et al., 2013) uses a predictive model to learn geometric encodings of words

Gender Gap	Word set	Axis Projection				Rel L2 Diff	Rel L2 Ratio
		w2v	w2v CBOW	GloVe	FastText	w2v	w2v
Census Wage Gap	threat	**-0.61**	-0.37	-0.30	-0.38	-0.52	-0.49
Female Legislators	criminal	**-0.49**	-0.33	-0.19	-0.17	-0.39	-0.38
Math & CS Degrees	stem-alum	**0.30**	0.28	**0.30**	0.26	0.28	0.29

Figure 5: Comparison of three bias metrics and four word embedding algorithms correlating themed word sets' gender bias with U.S. gender gap statistics. Unlike in Fig. 2, we perform feature selection using all countries.

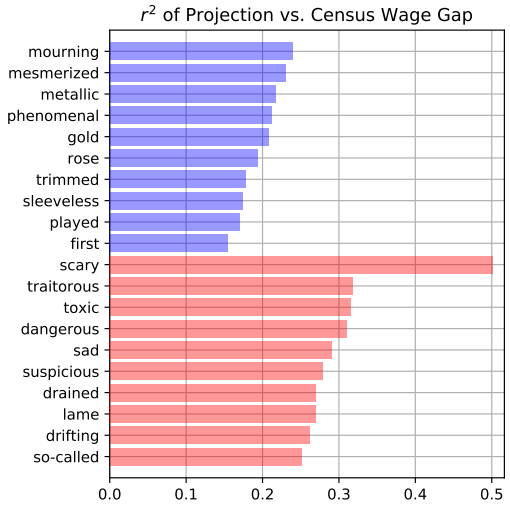

Figure 6: Ten adjectives with highest bias correlations to reduced pay gap (top, blue), and ten with highest correlation to increased pay gap (bottom, red).

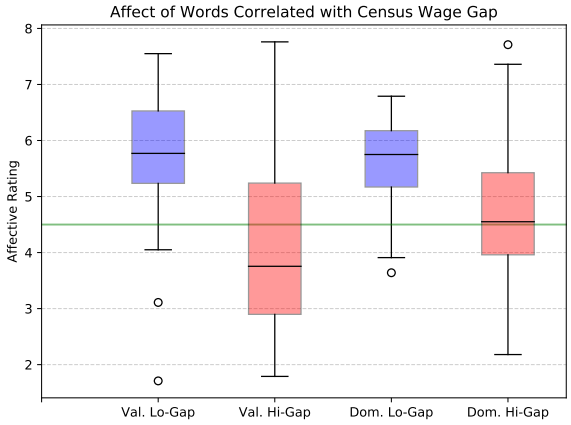

Figure 7: Valence and dominance scores for decreased pay gap words (blue) and increased pay gap words (red). Affect is neutral at 4.5 (plotted in green).

through a feed-forward neural network optimized by stochastic gradient descent. The Word2Vec *continuous bag-of-words* (CBOW) setting predicts the most probable word given a context. The Word2Vec *skip-gram* setting differs slightly by inputting a target word and predicting the context.

FastText (Joulin et al., 2016) characterizes each word as an n-gram of characters rather than an atomic entity. So each word vector is the sum of word vectors of the target word's n-gram (e.g. "app," "ppl," "ple" for "apple"). This is especially useful for rare words that might not exist in the corpus and accounting for misspellings.

Fig. 5 illustrates the above word embedding algorithms used on three different correlated word sets and statistics. In addition to comparing different word embedding algorithms, we also compare three different bias metrics on the Word2Vec algorithm: (1) the *axis projection* metric defined in Sec. 4.1; (2) the *relative L2 norm difference* (Garg et al., 2018); and the (3) *relative L2 norm ratio*. Unlike the axis projection, metrics (2) and (3) both compute the L2 norm from each word in the neutral word set to the $\overrightarrow{male}$ and $\overrightarrow{female}$ vectors, and then subtract or divide the two norms, respectively, returning the average over the word set.

The Fig. 5 results demonstrate that the gender bias is present in the product of all four word embedding algorithms, and is detectable with all three metrics. The Word2Vec approach with axis projection yields the highest coefficient of determination— for both direct and indirect correlation— across all three gender gap and word set pairs. This is the algorithm and bias metric that we use for all other experiments.

4.4 Valence and Dominance Analysis

Our Sec. 4.1 and Sec. 4.2 experiments specified word sets *a priori*, but we can also identify and analyze the individual words whose gender biases directly and indirectly correlate with statistical gender gaps to find trends and commonalities.

We identified all adjectives in the word embeddings using WordNet and then computed each adjective's R^2 score for direct or indirect correlation with each U.S. gender gap statistic. We filtered down the adjectives to those that correlate directly or indirectly with $R^2 > 0.1$. To illustrate, Fig. 6 plots ten highest R^2 words for direct (blue) and indirect (red) correlation against the pay gap, where blue adjectives' female bias correlates with *reduced* pay gap (higher wages) and red adjec-

tives' female bias correlates with *increased* pay gap (lower wages) in U.S. embeddings.

For each statistic, we measured the *valence* and the *dominance* of the directly- and indirectly-correlated adjectives using scores from Warriner et al. (2013). Fig. 7 shows a box plot of the valence and dominance of the reduced-gender-gap adjectives (blue) against increased-gender-gap adjectives (red) for the gender pay gap statistic, where the valence and dominance values for reduced gap (*Lo-Gap*) are significantly higher than the valence and dominance for increased gap (*Hi-Gap*) via t-test, where $p < 1.0e^{-7}$.

The same valence and dominance pattern held for adjectives directly and indirectly correlated with economic and educational gaps (i.e., Census Workforce Ratio, Female State Legislators, and Math & CS Degrees), where the valence and dominance of Lo-Gap words were significantly higher than Hi-Gap words with $p < .005$ throughout. The difference in valence and dominance for CDC Activity gap was not significant.

5 Conclusions

This paper characterized gender biases in Twitter-derived word embeddings from multiple cultures (99 countries and 51 U.S. regions) against statistical gender gaps in those cultures (18 international and 5 U.S.-based statistics).

We demonstrated that thematically-grouped word sets' gender biases correlate with gender gaps intuitively: word sets with a central topic or valence correlate with gender gaps of a similar topic, in a meaningful (positive or negative) direction. This supports our claims (from Sec. 1) that (1) cultural biases in language are correlated with cultural gender gaps and (2) we can characterize biases based on strength and direction of correlation with these gaps. We also demonstrated that these correlations are selective: not all topical word sets' biases correlate with all gender gaps, and random word sets do not correlate. This supports our claim that themed word sets capture different dimensions of gender bias and gender gaps.

Finally, we identified adjectives whose biases were highly correlated with increased and decreased gender gaps in education and economics, and we found that the adjectives correlated with *increased* gender gaps had statistically significantly lower valence and dominance than those correlated with *decreased* gender gaps. This is evidence of a cross-cutting attitude towards gender that we can characterize with future work.

The results of our three bias analyses are consistent with the social theory that differences in implicit gender valuation (e.g., linguistic gender bias) manifest in different gender opportunities and status (e.g., gender gaps) (Berger et al., 1972; Rashotte and Webster Jr, 2005). Specifically, when a culture attributes greater competence and social status to a gender, that gender receives higher rewards and evaluations (Dini, 2017).

Limitations and Future Work. Our use of English-only tweets facilitated comparison across embeddings, but it eliminates the native language of many countries and creates cultural blind-spots. Specifically, our use of English tweets does not capture the voices of those that (1) lack access to technology, (2) have poor knowledge of English, and (3) simply do not use Twitter. One might even argue that the gender bias effects may be even more pronounced off-line due to social desirability effects. Expanding to other languages presents additional challenges, e.g., gendered words and many-to-one vector mappings across languages, but recent language transformers facilitate this (Devlin et al., 2018). Incorporating additional languages and cultural texts are important next steps.

Previous Twitter word embedding approaches blend tweets with news or Wikipedia to improve NLP accuracy, using orders of magnitude more text per embedding (Li et al., 2017). Blending tweets with news may improve the embeddings' accuracy for NLP tasks, but it also risks diluting their implicit biases.

Finally, while our analyses illustrate correlations between gender biases and statistical gender gaps, they do not describe causality and they have limited interpretive power. We believe that integrating these methods with additional data and causal models (e.g., Dirichlet mixture models and Bayesian networks) will jointly improve interpretation and accuracy.

Acknowledgments

This research was supported by funding from the Defense Advanced Research Projects Agency (DARPA HR00111890015). The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

Joseph Berger, Bernard P Cohen, and Morris Zelditch Jr. 1972. Status characteristics and social interaction. *American Sociological Review*, pages 241–255.

Sebawit G Bishu and Mohamad G Alkadry. 2017. A systematic review of the gender pay gap and factors that predict it. *Administration & Society*, 49(1):65–104.

Debra L Blackwell and Tainya C Clarke. 2018. State variation in meeting the 2008 federal guidelines for both aerobic and muscle-strengthening activities through leisure-time physical activity among adults aged 18-64: United states, 2010-2015. *National health statistics reports*, (112):1–22.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Rachele Dini. 2017. *The second sex*. Macat Library.

Stewart D Friedman and Jeffrey H Greenhaus. 2000. *Work and family–allies or enemies?: what happens when business professionals confront life choices*. Oxford University Press, USA.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Angela Hawken and Gerardo L Munck. 2013. Cross-national indices with gender-differentiated data: what do they measure? how valid are they? *Social indicators research*, 111(3):801–838.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *CoRR*, abs/1607.01759.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2018. The geometry of culture: Analyzing meaning through word embeddings. *arXiv preprint arXiv:1803.09288*.

Quanzhi Li, Sameena Shah, Xiaomo Liu, and Armineh Nourbakhsh. 2017. Data sets: Word embeddings learned from tweets and general data. In *Eleventh International AAAI Conference on Web and Social Media*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Aparna Mitra. 2003. Establishment size, employment, and the gender wage gap. *The Journal of Socio-Economics*, 32(3):317–330.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Lisa Slattery Rashotte and Murray Webster Jr. 2005. Gender status beliefs. *Social Science Research*, 34(3):618–633.

Michelle Zimbalist Rosaldo. 1974. Woman, culture, and society: A theoretical overview. *Woman, culture, and society*, 21.

Michael T Schmitt and Nyla R Branscombe. 2001. The good, the bad, and the manly: Threats to one's prototypicality and evaluations of fellow in-group members. *Journal of Experimental Social Psychology*, 37(6):510–517.

Carole Vincent. 2013. Why do women earn less than men. *CRDCN Research Highlight/RCCDR en évidence*, 1(5):1.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

John E Williams and Deborah L Best. 1990. *Sex and psyche: Gender and self viewed cross-culturally.* Sage Publications, Inc.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2.

Jiping Zuo and Shengming Tang. 2000. Breadwinner status and gender ideologies of men and women regarding family roles. *Sociological perspectives*, 43(1):29–43.