Gender Gaps Correlate with Gender Bias in Social Media Word Embeddings

Scott Friedman, Sonja Schmer-Galunder, Anthony Chen, Robert Goldman, Michelle Ausman
{sfriedman, sgalunder, achen, rpgoldman, mausman}@sift.net
SIFT, LLC, 319 N 1st Ave
Minneapolis, MN 55401 USA

Abstract

Gender status, gender roles, and gender values vary widely across cultures. Anthropology has provided qualitative accounts of economic, cultural, and biological factors that impact social groups, and international organizations have gathered indices and surveys to help quantify gender inequalities in states. Concurrently, machine learning research has recently characterized pervasive gender biases in AI language models, rooting from biases in their textual training data. While these machine biases produce sub-optimal inferences, they may help us characterize and predict statistical gender gaps and gender values in the culture(s) that produced the training text, thereby helping us understand cultural context through big data. This paper presents an approach to (1) construct word embeddings (i.e., vector-based lexical semantics) from a region's social media, (2) quantify gender bias in word embeddings, and (3) correlate biases with survey responses and statistical gender gaps in education, politics, economics, and health. We validate this approach using 2018 Twitter data spanning 143 countries and 51 U.S. territories, 23 international and 7 U.S. gender gap statistics, and seven international survey results from the World Value Survey. Integrating these heterogeneous data across cultures is an important step toward understanding (1) how biases in culture might manifest in machine learning models and (2) how to estimate gender inequality from big data.

Keywords: gender bias; gender gaps; word embeddings; NLP

Introduction

As social media becomes available across the world, we have new opportunities to observe and interpret what we call *implicit cultural data*, comprising biases and themes couched in language. Thus, biases inherent in the local use of language can be seen as a window into the collective cognitive model of a specific culture. Such approach highlights two important implications: (1) Bias (in language) is a reflection of the lived experience (and not a product of machine learning models) and (2) attempts to *de-bias* data conceals inherent systemic inequalities (e.g. gender inequalities) (Gonen & Goldberg, 2019). Instead, understanding the cultural context (e.g. salient local differences in cognition or perception) can help explain which conclusions based on observed biases are meaningful.

Recently, machine-learned models that utilize *word embeddings* (i.e., vector-based representations of word semantics) have been shown to contain implicit racial and gender biases, arising primarily from biases in their training data. For example, using machine-learned word embeddings have produced analogies containing stereotypes such as *"man is to woman as doctor is to nurse"* (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016). These biases are sub-optimal, so recent work has developed *debiasing* techniques to improve accuracy and remove stereotypes (Bolukbasi et al., 2016; Zhao, Wang, Yatskar, Ordonez, & Chang, 2018; Zhang, Lemoine, & Mitchell, 2018).

In parallel with efforts to improve and de-bias these machine-learned language models, other research has begun to utilize biased models for prediction and diagnosis of present and historical social inequalities. For instance, biases of different cultures' text can correlate with survey responses of said cultures (Kozlowski, Taddy, & Evans, 2018), and biases in word embeddings trained over different decades can capture periods of societal shift, such as 1960s feminism (Garg, Schiebinger, Jurafsky, & Zou, 2018). This recent work provides numerical metrics of the gender biases in word embeddings and evidence that word embedding biases are indicators of social or cultural shifts.

This paper builds upon previous work to integrate word embedding bias within a larger context to help understand group bias. We integrate three types of data and use the following terminology throughout this paper:

1. **Implicit cultural data**: language bias computed from machine-learned word embeddings. These data represent systemic language biases, learned from a large volume of a culture's text (e.g., public social media posts).
2. **Explicit cultural data**: objective statistics about economic, educational, political, or developmental factors of a culture. These include statistical *gaps* (i.e., discrepancies in opportunity and status across groups).
3. **Survey data**: subjective answers to survey questions, aggregated on a per-culture basis.

Integrating these data, characterizing their combined value and understanding causal relationships between them is an important step in approximating cultural attitudes and relating them to cultural behaviors.

In this work, we focus on the topic of gender across countries and across U.S. states. Our implicit cultural data includes tweets from 143 countries and 51 U.S. territories that we use to build per-country and per-state embeddings. We assess each country's gender bias across multiple themes. Our explicit cultural data includes 23 international gender gap statistics and seven U.S. gender gap statistics from multiple sources. Our survey data includes eight questions about the

value of men and women in economic and university settings from the World Value Survey (WVS) (Inglehart et al., 2014).

The primary claim of this paper is that implicit gender biases correlate selectively and intuitively with relevant explicit data (i.e., statistical gender gaps) and survey data, across cultures. Our empirical results support this claim. Importantly, since our training data is entirely English social media data, we have a population bias for (1) literate, English-speaking individuals in predominantly non-English-speaking countries and (2) individuals with enough resources— and enough interest— to share their thoughts on social media. This population bias might be improved in future work, as we note in our conclusion.

We continue with a brief overview of gender gaps and then a description of our training data and experiments. We close with a discussion of the above claims and future work.

Methods and Materials

We describe the explicit, implicit, and survey data used in our experiments, as well as the methodology for computing gender bias in a high-dimension word embedding vector space.

Gender Gaps and Gender Valuation Surveys

Research in anthropology suggests that the public sphere (e.g., politics and economics) is often associated with the male gender and traits of assertiveness and competitiveness (Butler, 2011). Conversely, private or domestic spheres (e.g., domains of family and social relationships) are traditionally related to women (De Beauvoir & Parshley, 1953), although social relationships are considered more important by people independent of gender (Friedman & Greenhaus, 2000). Surveys such as the World Values Survey (WVS) (Inglehart et al., 2014) capture different countries' *valuations* of gender, and indices such as the Global Gender Gap (GGG) Report capture different countries' *outcomes* and concerning gender. These gender valuations and gender gap outcomes are highly related. For instance, in cultures where men tend to be over-represented economically and politically, men have higher salaries compared to women (Mitra, 2003; Vincent, 2013; Bishu & Alkadry, 2017). In this paper, our analyses are agnostic about the direction of causality, since asymmetrical gender valuations might cause gender gaps, and gender gaps might reinforce asymmetrical gender valuations.

Textual Training Data

Our implicit cultural data includes biases derived from word embeddings trained on different regions' Twitter posts. Our training data include public, geo-tagged tweets from Twitter users throughout 2018. 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 (Li, Shah, Liu, & Nourbakhsh, 2017), we replaced URLs, user names, images, and emojis with other tokens.

The dataset contains 143 international territories and 51 U.S. territories ranging from 310K tweets (Kosovo) to 1.8B tweets (all of USA). We sampled 25 million tweets for all territories that exceeded that number. These corpora are orders of magnitude smaller than other approaches for tweet embeddings (Li et al., 2017). We use Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) skip-gram algorithm to construct separate word embeddings for each region.

Word-Sets for Thematic Bias

Our materials included word-sets based in part on survey-based experiments (Williams & 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) *thematic words* about a shared theme but with no explicit gender ascription.

Our female and male word-sets were derived from previous work (Garg et al., 2018) and extended to add additional nouns found in tweets (e.g., girlfriend, boyfriend, wife, husband, mom, dad, mama, papa). We seeded our thematic word sets as possible from (Garg et al., 2018) and we generated other thematic word-sets to represent social constructs: *politics* (democrat, republican, senate, government, politics, minister, presidency, vote, parliament, ...), *communal* (community, society, humanity, welfare, ...), *victim* (victim, vulnerable, abused, survivor, ...), *childcare* (child, children, parent, baby, nanny, ...), *excellent* (excellent, fantastic, phenomenal, outstanding, ...), *workforce* (market, job, salary, pay, wage, career, boss, ...), and others. Each thematic word-set (e.g., *politics*) comprises the identical set of words in our U.S. and international experiments below, but we did not use every word-set for both analyses.

Axis Projection as Gender Bias

In our experiments, we compute per-gender vectors $\overrightarrow{female}$ and $\overrightarrow{male}$ by averaging the vectors of each constituent word, following Garg et al. (2018). Within a region's word embedding, we compute the region's gender bias of a thematic word-set W as an *average axis projection* of the 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 projects each thematic word's vector $\overrightarrow{w}$ onto the gender axis, which is computed as the gender difference vector $\overrightarrow{female} - \overrightarrow{male}$ scaled by the L2 norm $||\overrightarrow{female} - \overrightarrow{male}||_2$. The bias of theme W is the average of each word $w \in W$. This is our primary measure of thematic gender bias in implicit cultural data.

For any thematic word-set (e.g., *politics*), we compute the average axis projection for all countries and compute its correlation to international gender gaps. For instance, Fig. 1 plots each country's *politics* word-set bias against the z-normalized GGG statistic "Women in Parliament" (where greater score indicates greater share of women in parliament) with $r^2 = 0.29$. Female bias increases along the x-axis, where

0.0 indicates no bias. We revisit this specific result and describe others in our experimental analysis below.

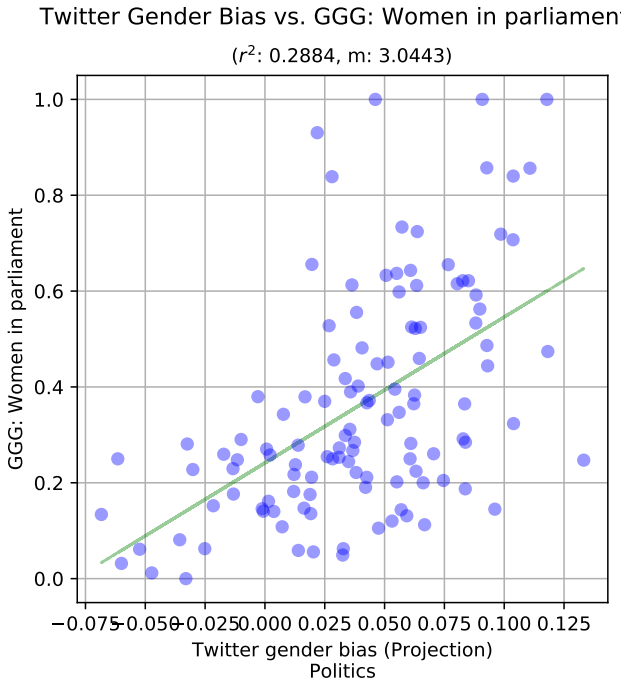

Figure 1: Correlation of countries' gender bias of political words (x-axis; female bias increases in positive direction) against the GGG "Women in parliament" statistic (y-axis; female participation increases in positive direction).

Experiment 1: International Biases and Gaps

This analysis characterizes the relationship between (1) implicit gender biases from word embeddings and (2) statistical gender gaps and survey data.

This experiment utilizes 23 gender gap statistics from the World Economic Forum's 2018 Global Gender Gap (GGG) report,[1] United Nations Human Development Indices,[2], the Georgetown Institute on Women, Peace, and Security (GIWPS) Index,[3] and 8 survey questions concerning gender valuation from the WVS. We have WVS responses for 55 of the 143 countries with word embeddings, so we use that 55-country subset when correlating against WVS data.

Fig. 1 shows a single correlation between a GGG political gender gap statistic and countries' Twitter gender bias on the *politics* theme: as the politics themes increase in female bias, women have a larger percentage of seats in their countries' parliaments. This is consistent with our claim that implicit gender biases in word embeddings correlate meaningfully with gender gaps.

We ran similar analyses of eight themed word-sets and two randomly-generated word-sets against all international gender gap statistics and 8 WVS survey results. For each pair of word-set and gender gap, the algorithm (1) performs feature selection to optionally down-select to at least three words and then (2) uses the down-selected word set to compute the *r*-value for that pair. We plot this in Fig. 3, with bold horizontal lines separating different datasets and dotted horizontal lines (and *eq*↑ and *eq*↓ to indicate positive and negative indicators of gender equality, respectively). Each table cell is a distinct *r*-value (i.e., correlation coefficient).

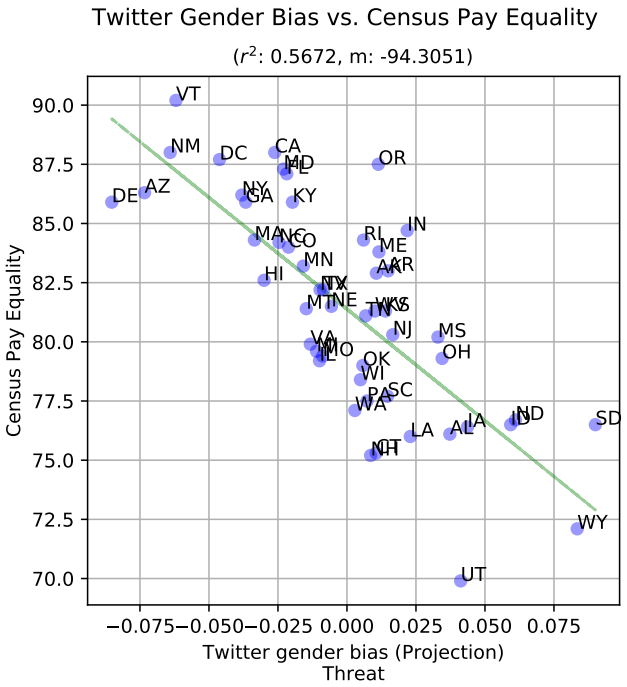

Figure 2: Correlation of U.S. states 2016 women's wages (cents on the dollar per men's wages; y-axis) against each state's Twitter gender bias of the *threat*-themed word-set.

As shown in Fig. 3, themed word-sets vary in their direction and strength of correlation across different statistics. Female *politics* and *workforce* biases correlate strongest with the political and economic empowerment indices, and also *negatively* with WVS agreement that women have value in political and economic positions. Female *community* and *childcare* and *illness* biases correlate with low literacy, educational enrollment, share of professional and technical positions, and Gender Development Index, and a high agreement with the survey question that men are more fitting for business, politics, and university, and that women should not earn more than their husbands. Female *attractive* and *intelligent* biases generally correlated in opposite directions on most dimensions, and high *attractive* bias is the highest indicator of low contraceptive prevalence. The random word-sets in Fig. 3 are substantially weaker in correlation than any thematic columns.

Experiment 2: U.S. Biases and Gaps

We used the same experimental setup as above on 51 U.S. territories (50 states and Washington, D.C.). We used

[1] https://www.weforum.org/reports/the-global-gender-gap-report-2018

[2] http://hdr.undp.org/en/content/human-development-index-hdi

[3] https://giwps.georgetown.edu/the-index/

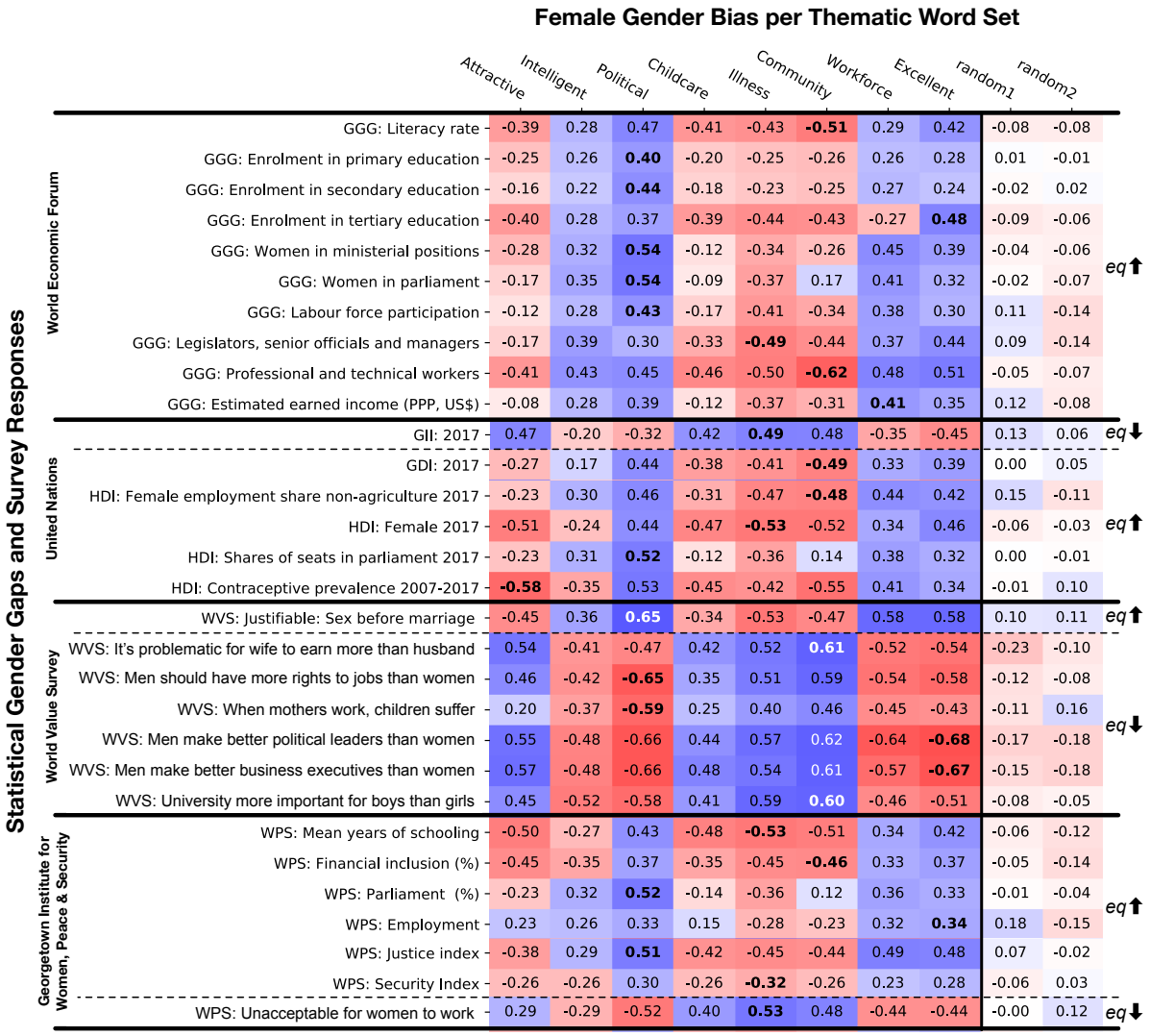

Figure 3: Correlation of themed word sets' gender bias (columns) against international gender gap statistics and WVS survey responses about gender (rows). Values are *r*-values (correlation coefficients), where negative indicates inverse correlation. The two word sets *rand1* and *rand2* were randomly sampled from the embeddings for comparison.

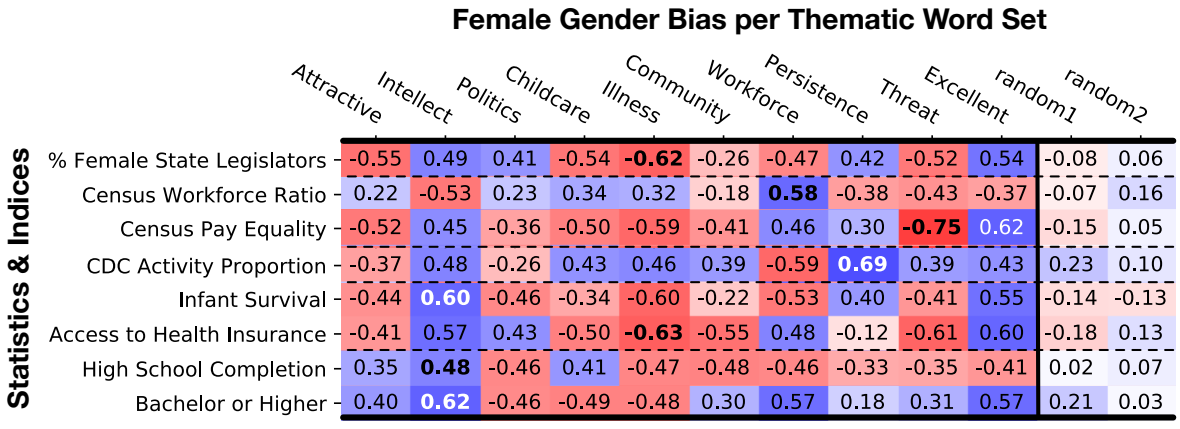

Figure 4: Correlation of themed word sets' gender bias (columns) against U.S. gender gap statistics (rows). Values are *r*-values (correlation coefficients), where negative indicates inverse correlation. The two word sets *rand1* and *rand2* were randomly sampled from the embeddings for comparison.

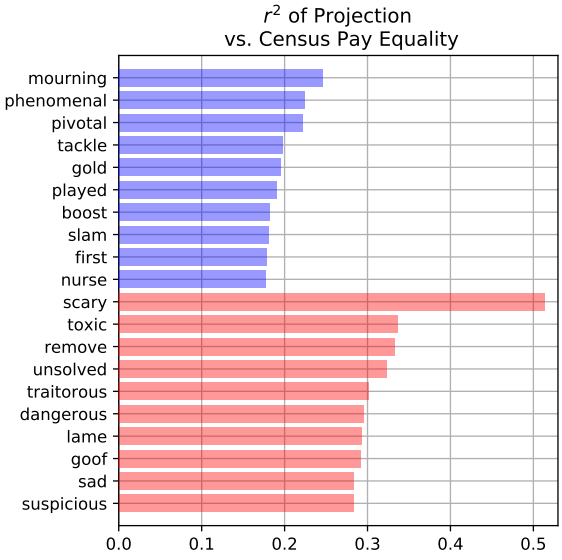

Figure 5: Ten words whose female bias correlates most positively with female wage equality (top, blue), and ten words whose female bias correlates most negatively (bottom, red). r^2 values are reported for compact comparison of positively- and negatively-correlated words, but we cluster these into positive and negative correlation direction (i.e., positive and negative r-values).

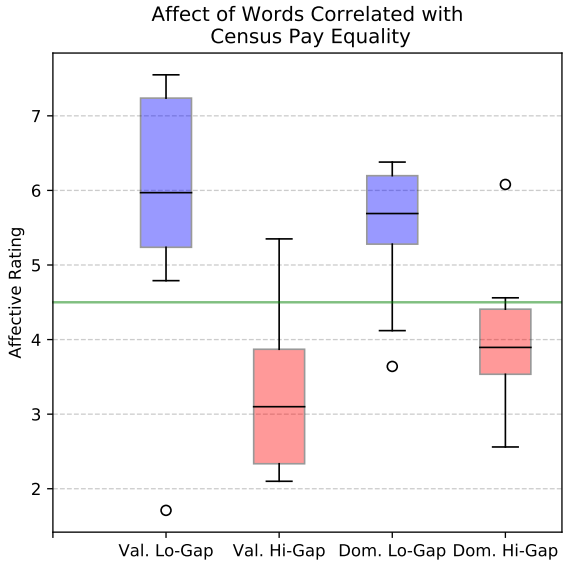

Figure 6: Valence ("Val") and dominance ("Dom") scores for the Fig. 5 wage equality words (blue) and wage inequality words (red). Affect is neutral at 4.5 (plotted in green).

geo-tagged tweets in conjunction with eight gender-relevant statistics from the U.S. census, U.S. Center for Disease Control (CDC), and other sources.

A scatter-plot of the strongest correlation is shown in Fig. 2, where the female bias of the *threat* theme (including adjectives such as *"scary," "toxic," "threat,"* and *"dangerous"*) is inversely correlated with pay equality. The the nature of the threat (e.g., whether women are threatened or threatening) and the presence of causality is not clear from this high-level analysis, and we revisit these questions in the conclusion.

As with the international analysis, we plotted each theme with each statistic, as shown in Fig. 4. Each table cell is a distinct r-value (i.e., correlation coefficient). Notably, female *intellect* bias is most highly correlated with the two educational outcomes, female *illness* bias is inversely correlated with female access to health insurance and with female legislature seats, female *workforce* bias is most highly correlated with women in the workforce, and female *persistence* bias is most highly correlated with CDC activity proportion (where women meet exercise guidelines relative to men). The random word-sets in Fig. 4 do not meaningfully correlate.

In both the international and U.S. analyses, the selective correlation of thematic word-sets with gender gaps and survey responses supports our claim that implicit gender biases— as captured in word embeddings from countries' social media— correlate selectively and intuitively with relevant gender gaps and survey data.

Experiment 3: Valence and Dominance

Our first two 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.

For each of the eight U.S. gender-based statistics, we rank the WordNet adjectives' correlation that directly and indirectly correlate against them based on r scores. To illustrate with a single statistic, Fig. 5 plots ten highest r words for positive (blue) and ten lowest r words for negative (red) correlation with pay equality.

For each statistic, we measured the *valence* and the *dominance* of the positively- and negatively-correlated adjectives using scores from Warriner, Kuperman, and Brysbaert (2013). Fig. 6 shows a box plot of the valence and dominance of the adjectives in Fig. 5: the blue (positively-correlated adjectives) are significantly higher valence and higher dominance than the red (negatively-correlated adjectives) via t-test ($p < 1.0e^{-4}$ and $p < 5.0e^{-4}$, respectively).

Over all eight U.S. gender-based statistics, dominance and valence averages were higher for adjectives that positively correlate with gender equality than for adjectives that negatively correlate, except for "High School Completion." The differences in valence and dominance were significant for "Census Wage Equality" (described above), "% Female State Legislators" ($p < 0.005$), and "Infant Survival" ($p < 0.005$).

These results across gender gap statistics suggest that *gaps* in gender opportunity correlate with implicit biases in lower-valence, lower-dominance concepts, and gender *equality* in status and opportunity correlates with implicit biases in higher-valence, higher-dominance concepts.

Conclusions

This paper demonstrated that gender biases in Twitter-derived word embeddings from 143 countries and 51 U.S. territories correlate meaningfully with gender gap statistics and survey questions about gender valuation.

Our international and U.S. analyses demonstrate that implicit cultural data— computed as vector-space gender biases over thematic word-sets— correlates with statistical gender gaps intuitively. Different word-sets' gender biases correlated with statistical gender gaps and survey data of a similar theme, in a meaningful (positive or negative) direction. Not all thematic word-sets' biases correlate with all gender gaps, and random word sets do not correlate. This supports our claim that implicit gender biases correlate selectively and intuitively with relevant explicit data and survey data.

All of our empirical results are consistent with the social science research that gender biases manifest in implicit ways and that differences in implicit gender bias (e.g., linguistic gender bias) are associated with gender valuations (assessed via survey responses) and metrics that quantify gender opportunities and status (i.e., gender gaps) (Berger, Cohen, & Zelditch Jr, 1972; Rashotte & Webster Jr, 2005). Thus, quantifying biases inherent in large data in order to facilitate comparisons between nations, can help capture variables that may cause structural barriers for women, and in turn help inform global gender equality policies. Our work is a first step in mapping the global gender landscape based on unstructured data and help solidify existing measures, thus providing more validity to existing measures of gaps in opportunities for women worldwide. Next, we plan to develop causal models to make this mapping more dynamic and generalizable.

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., with additional gendered words and many-to-one vector mappings across languages, but recent language transformers facilitate this (Devlin, Chang, Lee, & Toutanova, 2018). Consequently, incorporating additional languages and cultural texts are important next steps.

Finally, while our analyses illustrate correlations between gender biases and statistical gender gaps, they do not describe causality and they have limited interpretive power. Integrating our existing 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

Berger, J., Cohen, B. P., & Zelditch Jr, M. (1972). Status characteristics and social interaction. *American Sociological Review*, 241–255.

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

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems* (pp. 4349–4357).

Butler, J. (2011). *Gender trouble: Feminism and the subversion of identity*. routledge.

De Beauvoir, S., & Parshley, H. M. (1953). *The second sex*. Vintage books New York.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., ... others (2014). World values survey: Round six-country-pooled datafile 2010-2014. *JD Systems Institute, Madrid*.

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2018). The geometry of culture: Analyzing meaning through word embeddings. *arXiv preprint arXiv:1803.09288*.

Li, Q., Shah, S., Liu, X., & Nourbakhsh, A. (2017). Data sets: Word embeddings learned from tweets and general data. In *Eleventh international aaai conference on web and social media.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in*

neural information processing systems 26 (pp. 3111–3119).

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

Rashotte, L. S., & Webster Jr, M. (2005). Gender status beliefs. *Social Science Research*, *34*(3), 618–633.

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

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, *45*(4), 1191–1207.

Williams, J. E., & Best, D. L. (1990). *Sex and psyche: Gender and self viewed cross-culturally.* Sage Publications, Inc.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 aaai/acm conference on ai, ethics, and society* (pp. 335–340).

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (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* (Vol. 2).