Relating Linguistic Gender Bias, Gender Values, and Gender Gaps: An International Analysis*

Scott Friedman, Sonja Schmer-Galunder, Jeffrey Rye,
Robert Goldman, and Anthony Chen

SIFT, Minneapolis MN 55401, USA
{friedman, sgalunder, rye, rpgoldman, achen}@sift.net

Abstract. Recent research in machine learning has shown that many machine-learned language models contain pervasive racial and gender biases, rooting from biases in their textual training data. While these biases produce sub-optimal parsing and 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) quantify gender bias in word embeddings (i.e., vector-based lexical semantics), (2) correlate gender biases with survey responses and statistical gender gaps in education, politics, economics, and health, and (3) integrate numerical biases and statistics to model different cultures' survey results more accurately than either in isolation. We validate this approach using 2018 Twitter data spanning 99 countries, 18 Global Gender Gap statistics from the World Economic Forum, and 8 international survey results from the World Value Survey. Integrating these heterogeneous data across cultures is an important step toward building computational models to understand group bias.

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

1 Introduction

Machine-learned models that utilize *word embeddings* (i.e., vector-based representations of word semantics) are presently under scrutiny for biases and stereotypes, e.g., in race and gender, 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"* [3]. These biases are sub-optimal, so recent work has developed *debiasing* techniques to improve accuracy and remove stereotypes [3, 19, 18].

In parallel with efforts to improve and debias these machine-learned language models, other research has begun to utilize biased models for prediction and

* 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.

diagnosis of present and historical social inequalities. For instance, by comparing word embeddings trained on different cultures' text, we can correlate biases to survey data [11], or by comparing embeddings trained over different decades, we can capture periods of societal shift, such as 1960s feminism [8]. This recent work provides numerical metrics of the gender biases in word embeddings.

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.
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 and characterizing their combined value 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. Our implicit cultural data includes tweets from 99 countries that we use to build per-country embeddings and assess each country's gender bias across multiple themes. Our explicit cultural data includes 18 international gender gap statistics from the World Economic Forum's 2018 Global Gender Gap (GGG) report.[1] Our survey data includes eight questions about the value of men and women in economic and university settings from the World Value Survey (WVS) [10].

The primary claims of this paper are that (1) implicit gender biases correlate selectively and intuitively with relevant gender gaps and survey data and (2) implicit data (i.e., biases from word embeddings) and explicit data (i.e., gender gap statistics) model cultures' survey values more accurately than either alone. Our empirical results from two experiments support these claims.

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 Gap Statistics and Gender Valuation Surveys

Research in anthropology suggests that the public sphere (e.g., domains of politics and economics) is often associated with the male gender and traits of assertiveness and competitiveness [4]. Conversely, private or domestic spheres (e.g., domains of family and social relationships) are traditionally related to women [5], although social relationships are considered more important by people independent of gender [7]. Surveys such as the World Values Survey (WVS) [10]

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

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 [14, 16, 2]. 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.

This paper's international analyses utilize 18 gender gap statistics from the GGG and 8 survey questions concerning gender valuation from the WVS. We found the GGG to have high validity, reliably measuring well-defined gender-differentiated dimensions across most countries [9]. We used 44 countries' survey responses[2] from the following WVS questions (some abbreviated):

1. Having a job is the best way for a woman to be an independent person.
2. Sex before marriage is justifiable.
3. If a woman earns more than her husband, it will cause problems.
4. When jobs are scarce, men should have more rights to jobs than women.
5. When a mother works for pay, the children suffer.
6. Men make better political leaders than women do.
7. Men make better business executives than women do.
8. University education is more important for boys than girls.

3 Training Data

Our implicit cultural data includes biases derived from word embeddings trained on different countries' Twitter posts. Our training data include public tweets from 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., [12]), we replaced URLs, user names, hashtags, images, and emojis with other tokens.

The dataset contains 99 countries ranging from 98K tweets (Maritius) to 122M tweets (U.K.). We sampled 10 million tweets for all countries that exceeded that number. These corpora are orders of magnitude smaller than other approaches for tweet embeddings (e.g., [12]). We use Word2Vec [13] skip-gram algorithm to construct separate word embeddings for each country in our analyses (i.e., 99 countries resulted in 99 separate word embeddings).

[2] WVS responses pertain to 44 of the 99 countries for which we have word embeddings, so when correlating against WVS data, we use that 44-country subset.

4 Experiments

We describe our empirical results, including a cross-correlation of linguistic gender bias with gender gaps and survey data (Sec. 4.1) and an analysis combining gender gaps and gender bias to correlate with countries' survey results (Sec. 4.2).

4.1 Correlating word embeddings' gender biases with gender gaps

This analysis characterizes the relationship between (1) implicit gender biases and (2) statistical gender gaps and survey data. We compute the gender biases of different countries by projecting word-sets onto the female-male axis in each country's word vector space. By varying the word-set that we project onto the axis, we can assess different *themes* of gender bias. Here we characterize the relationship of thematic gender biases to statistical gender gaps.

Word-sets. Our materials included word-sets based in part on survey-based experiments [17] and recent work on word embeddings [8]. 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 [8] and extended to add additional nouns found in tweets (e.g., girlfriend, boyfriend, wife, husband, mom, dad, mama, papa). Our *intellect* thematic adjectives are from [8], and we generated 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.

Axis projection as gender bias. We compute per-gender vectors $\overrightarrow{female}$ and $\overrightarrow{male}$ by averaging the vectors of each constituent word, following [8]. Within a country's word embedding, we compute the country's gender bias of a thematic word-set W as an *average axis projection* of the W onto the male-female axis as:

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

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.

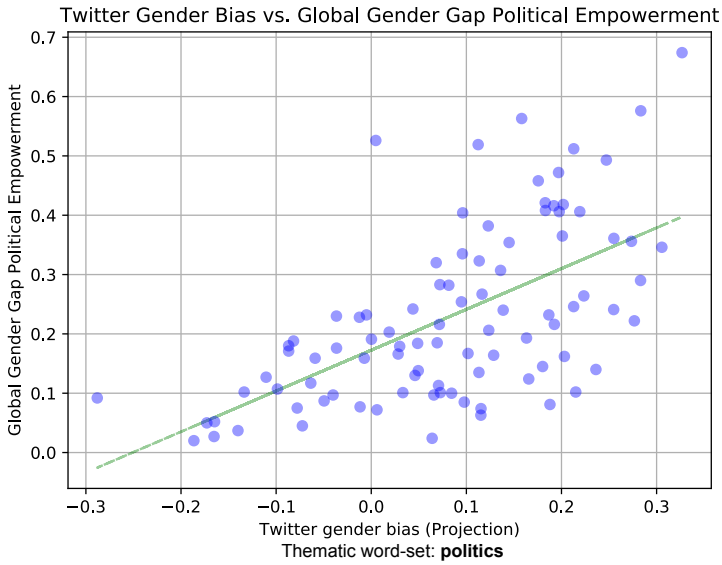
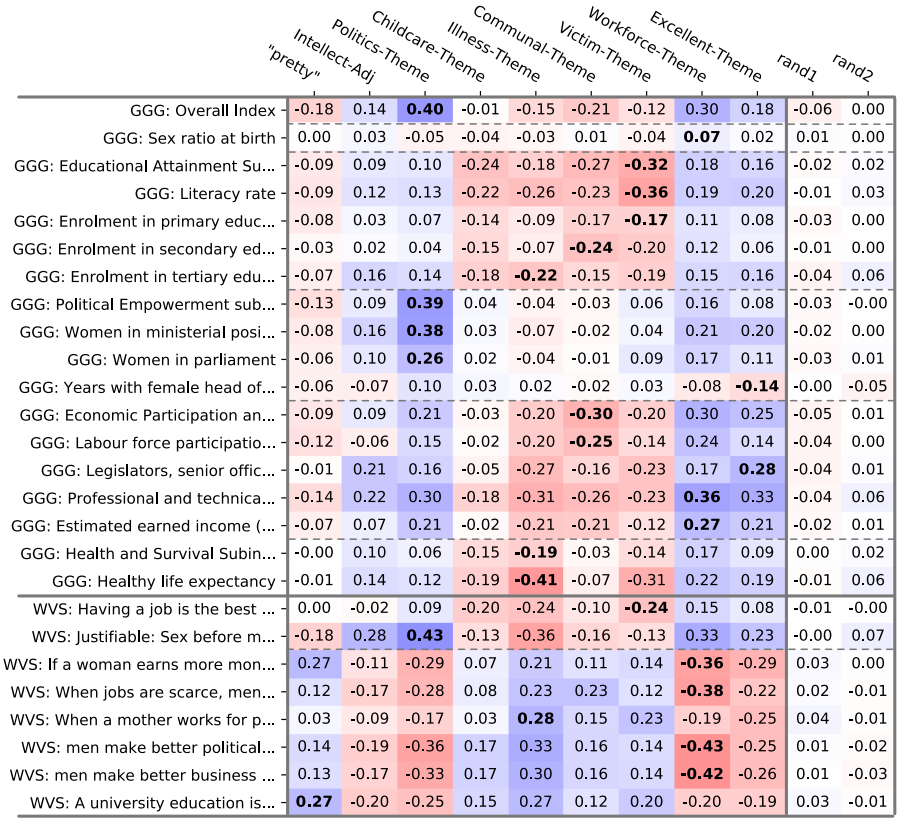


Fig. 1. Correlation of countries' gender bias of political words (x-axis; female association increases in positive direction) against the GGG gender political empowerment index (y-axis; gender equality increases in positive direction).

Correlating gender biases with gender gaps. For any neutral word list (e.g., political terms), we compute the average axis projection for all countries and compute its correlation to international gender gaps. Fig. 1 plots each country's political word-set bias against the GGG political empowerment gender gap subindex (where greater score indicates greater female political empowerment) with $R^2 = 0.40$. Female bias increases along the x-axis, where 0.0 indicates no bias. The Fig. 1 plot is consistent with the hypothesis that— globally, over our set of countries— women's political opportunities and status increase (relative to men) as political language shows a more female bias. This is not a surprising finding, but it supports our claim that implicit cultural data— estimated via word embedding biases— help intuitively model explicit cultural data.

We ran similar analyses of nine themed word-sets and two randomly-generated word-sets against all 18 GGG statistics and all 8 WVS survey results. For each pair of word-set and GGG/WVS value, 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^2 for that pair. We manually negate the R^2 value to indicate indirect correlation of the gender language bias and the statistic. We plot this in Fig. 2, with dotted horizontal lines separating different topics of gender gaps and a bold line separating GGG data from WVS.

Themed word-sets vary in their direction and strength of correlation across different GGG statistic groups: the *politics* theme correlates strongest with the political empowerment indices, and also positively with economic indices, and relatively weaker over health and education indices; *intellect* and *workplace* terms



	"pretty"	Intellect-Adj	Politics-Theme	Childcare-Theme	Illness-Theme	Communal-Theme	Victim-Theme	Workforce-Theme	Excellent-Theme	rand1	rand2
GGG: Overall Index	-0.18	0.14	**0.40**	-0.01	-0.15	-0.21	-0.12	0.30	0.18	-0.06	0.00
GGG: Sex ratio at birth	0.00	0.03	-0.05	-0.04	-0.03	0.01	-0.04	**0.07**	0.02	0.01	0.00
GGG: Educational Attainment Su...	-0.09	0.09	0.10	-0.24	-0.18	-0.27	**-0.32**	0.18	0.16	-0.02	0.02
GGG: Literacy rate	-0.09	0.12	0.13	-0.22	-0.26	-0.23	**-0.36**	0.19	0.20	-0.01	0.03
GGG: Enrolment in primary educ...	-0.08	0.03	0.07	-0.14	-0.09	-0.17	**-0.17**	0.11	0.08	-0.03	0.00
GGG: Enrolment in secondary ed...	-0.03	0.02	0.04	-0.15	-0.07	**-0.24**	-0.20	0.12	0.06	-0.01	0.00
GGG: Enrolment in tertiary edu...	-0.07	0.16	0.14	-0.18	**-0.22**	-0.15	-0.19	0.15	0.16	-0.04	0.06
GGG: Political Empowerment sub...	-0.13	0.09	**0.39**	0.04	-0.04	-0.03	0.06	0.16	0.08	-0.03	-0.00
GGG: Women in ministerial posi...	-0.08	0.16	**0.38**	0.03	-0.07	-0.02	0.04	0.21	0.20	-0.02	0.00
GGG: Women in parliament	-0.06	0.10	**0.26**	0.02	-0.04	-0.01	0.09	0.17	0.11	-0.03	0.01
GGG: Years with female head of...	-0.06	-0.07	0.10	0.03	0.02	-0.02	0.03	-0.08	**-0.14**	-0.00	-0.05
GGG: Economic Participation an...	-0.09	0.09	0.21	-0.03	-0.20	**-0.30**	-0.20	0.30	0.25	-0.05	0.01
GGG: Labour force participatio...	-0.12	-0.06	0.15	-0.02	-0.20	**-0.25**	-0.14	0.24	0.14	-0.04	0.00
GGG: Legislators, senior offic...	-0.01	0.21	0.16	-0.05	-0.27	-0.16	-0.23	0.17	**0.28**	-0.04	0.01
GGG: Professional and technica...	-0.14	0.22	0.30	-0.18	-0.31	-0.26	-0.23	**0.36**	0.33	-0.04	0.06
GGG: Estimated earned income (...	-0.07	0.07	0.21	-0.02	-0.21	-0.21	-0.12	**0.27**	0.21	-0.02	0.01
GGG: Health and Survival Subin...	-0.00	0.10	0.06	-0.15	**-0.19**	-0.03	-0.14	0.17	0.09	0.00	0.02
GGG: Healthy life expectancy	-0.01	0.14	0.12	-0.19	**-0.41**	-0.07	-0.31	0.22	0.19	-0.01	0.06
WVS: Having a job is the best ...	0.00	-0.02	0.09	-0.20	-0.24	-0.10	**-0.24**	0.15	0.08	-0.01	-0.00
WVS: Justifiable: Sex before m...	-0.18	0.28	**0.43**	-0.13	-0.36	-0.16	-0.13	0.33	0.23	-0.00	0.07
WVS: If a woman earns more mon...	0.27	-0.11	-0.29	0.07	0.21	0.11	0.14	**-0.36**	-0.29	0.03	0.00
WVS: When jobs are scarce, men...	0.12	-0.17	-0.28	0.08	0.23	0.23	0.12	**-0.38**	-0.22	0.02	-0.01
WVS: When a mother works for p...	0.03	-0.09	-0.17	0.03	**0.28**	0.15	0.23	-0.19	-0.25	0.04	-0.01
WVS: men make better political...	0.14	-0.19	-0.36	0.17	0.33	0.16	0.14	**-0.43**	-0.25	0.01	-0.02
WVS: men make better business ...	0.13	-0.17	-0.33	0.17	0.30	0.16	0.14	**-0.42**	-0.26	0.01	-0.03
WVS: A university education is...	**0.27**	-0.20	-0.25	0.15	0.27	0.12	0.20	-0.20	-0.19	0.03	-0.01

Fig. 2. Correlation of themed word sets' gender bias (columns) against GGG gender gap statistics and WVS survey responses about gender (rows). Values are R^2 coefficient of determination, where negation is added to indicate inverse correlation. The two word sets *rand1* and *rand2* were randomly sampled from the embeddings for comparison.

positively correlate with economic statistics but are weak correlates otherwise; *illness* terms indirectly correlated strongest with health and survival statistics, and weaker elsewhere. The word *"pretty"* (Fig. 2, left column), was the single word with the strongest determination to the overall gender gap and other sub-indices. The random word-sets in Fig. 2 do not exceed $R^2 = 0.07$ in any row.

Themed word-sets also varied in correlation with WVS data. Specifically, agreement with questions "Having a job is the best way for a women to be an independent person" and "Justifiable: Sex before marriage" was correlated with female association with *politics*, *workforce*, and *excellent* themes (neutral or positive valence), and inversely correlated with *victim*, *childcare*, *communal*, and *illness* themes (neutral or negative valence). Those correlations were generally reversed for the other six WVS, which asserted men had greater value— or should receive more opportunities— in economic, political, or university settings. Intuitively, the *workforce* word-set was the largest determiner of this group of

WVS questions: as countries' female bias of the *workforce* theme increased in their word embeddings, survey respondents in those countries were less likely to agree to reduced female value and opportunity in politics and economics.

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.

None of our themed word-sets strongly correlated with (1) sex ratio at birth, which was 1.0 for the vast majority of countries or (2) percentage of last 50 years with female head of state. These may correlate with other themes, or they may have a weaker or more idiosyncratic relationship to implicit gender bias.

	Explicit								**Explicit + Implicit**	
	Econ		*Econ/Health*		*Econ/Edu*		*Econ/Poli*		*Econ/Implicit*	
International Survey Question	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**
If a woman earns more money than her husband, its almost certain to cause problems.	0.36	19.09	0.39	18.62	0.6	15.67	0.5	16.92	**0.6**	**15.08**
University education is more important for boys than girls.	0.49	16.7	0.57	15.41	0.63	14.37	**0.71**	**12.65**	0.69	13.01
Justifiable: Sex before marriage.	0.53	1.51	0.58	1.43	0.62	1.36	0.72	1.17	**0.78**	**1.04**
When a mother works for pay, the children suffer.	0.59	23.04	0.6	22.54	0.64	21.64	0.62	22.15	**0.7**	**19.45**
Men make better political leaders than women do.	0.55	29.82	0.56	29.57	0.65	26	**0.76**	**21.74**	0.75	22.03
Men make better business executives than women do.	0.58	25.5	0.62	24.21	0.69	22.35	0.73	20.42	**0.76**	**19.38**
When jobs are scarce, men should have more rights to jobs than women.	0.54	29.83	0.64	26.48	0.64	27.01	**0.69**	**24.39**	0.68	24.82
Having a job is the best way for a woman to be an independent person.	0.23	15.22	0.41	13.35	0.44	12.28	0.39	13.51	**0.61**	**10.76**

Table 1. Correlation strength of WVS international survey questions using explicit GGG statistics alone versus also utilizing implicit language bias data from countries' word embeddings.

4.2 Correlating Survey Results with Gender Gaps and Gender Bias

Here we regress survey responses (WVS data) against both implicit cultural data (i.e., gender biases derived from word embeddings) and explicit cultural data (GGG statistics). Our objective is to broadly assess the additional value of combining these types of cultural data— rather than using them in isolation— to determine the joint utility of the features for understanding and predicting group biases. We use multiple linear regression and compare the error and coefficient of determination of two different settings:

- **Explicit**: We regress WVS responses against a subset of GGG data (e.g., economic, education, health, or political indices listed in Fig. 2).
- **Explicit+Implicit**: We regress WVS responses against a combination of GGG data and five gender bias themes (*politics*, *intelligent*, *victim*, *workforce*, and *excellent*).

Table 1 lists the R^2 and RMSE for four Explicit conditions (Econ, Econ/Health, Econ/Edu, and Econ/Poli) and one Explicit+Implicit condition. We included the economics GGG group in all conditions because it is the strongest correlate

of all GGG statistic groups. As shown in Table 1, the Explicit+Implicit condition scored highest R^2 correlations for 5/8 survey questions, followed by the economic/political explicit case with 3/5. This suggests that the implicit data added discriminatory power that was not captured by the GGG explicit data.

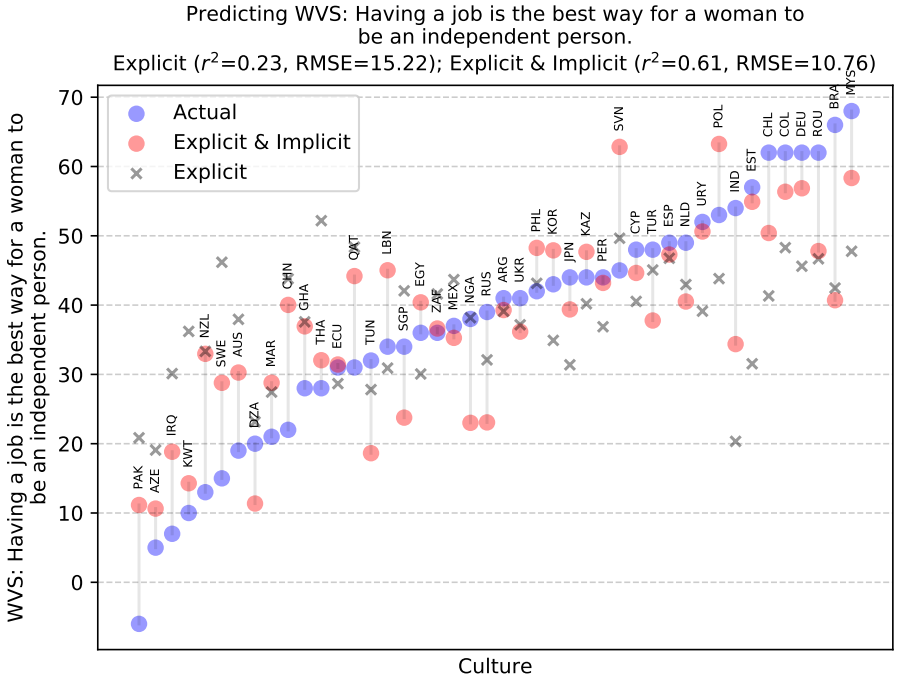


Fig. 3. Multiple linear regression of countries' survey responses using explicit data alone (i.e., GGG statistics) versus also integrating implicit data (i.e., thematic gender bias from countries' word embeddings). Countries are sorted by ascending (actual) y-value, where agreement with the survey question increases in the y-direction.

For the last survey question, adding implicit data increased correlation strength by 165% and decreased RMSE by 29%. We plot the regression of this survey question against both Explicit (Econ) and Explicit+Implicit (Econ/Implicit) conditions in Fig. 3. Countries are ordered left-to-right by their observed agreement with the survey question (blue dots). The linear model improves its accuracy for most countries with the addition of implicit data.

The results in Table 1 and Fig. 3 support our claim that implicit and explicit data jointly predict cultures' survey values more accurately than either alone.

5 Conclusions

This paper characterized gender biases in Twitter-derived word embeddings from 99 countries against (1) 18 statistical gender gaps and (2) eight survey questions from a 44-country subset.

Our first analysis (Sec. 4.1) demonstrated 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 (from Sec. 1) that implicit gender biases correlate selectively and intuitively with relevant explicit data and survey data.

Our second analysis (Sec. 4.2) performed multiple linear regression with explicit and implicit cultural data. Our results show that combining these data substantially strengthen the correlation to survey data. Furthermore, adding implicit gender bias data to economic gender gap statistics outperformed conditions that utilized economic stats with other categories of gender gap statistics (e.g., political, education, and health). This analysis supports our second claim (from Sec. 1) that implicit data and explicit data jointly correlate with survey responses more accurately than either alone.

All of our empirical results are consistent with the social theory that differences in implicit gender bias (e.g., linguistic gender bias) are associated with gender valuations (assessed via survey responses) and differences in gender opportunities and status (i.e., gender gaps) [1, 15].

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 [6]. Consequently, 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 [12]. 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. For instance, our second experiment (Sec. 4.2) utilized the same gender gaps and gender biases to predict eight different survey responses— spanning female employment, female education, and sex before marriage— but a more interpretable model would operate only on the subset of these variable-to-variable relationships that have explanatory power and theoretical merit in the social science domain. Integrating our existing methods with additional data and causal models (e.g., Dirichlet mixture models and Bayesian networks) will jointly improve interpretation and accuracy.

References

1. Berger, J., Cohen, B.P., Zelditch Jr, M.: Status characteristics and social interaction. American Sociological Review pp. 241–255 (1972)
2. Bishu, S.G., Alkadry, M.G.: A systematic review of the gender pay gap and factors that predict it. Administration & Society **49**(1), 65–104 (2017)
3. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Advances in neural information processing systems. pp. 4349–4357 (2016)
4. Butler, J.: Gender trouble: Feminism and the subversion of identity. routledge (2011)
5. De Beauvoir, S., Parshley, H.M.: The second sex. Vintage books New York (1953)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Friedman, S.D., Greenhaus, J.H.: Work and family–allies or enemies?: what happens when business professionals confront life choices. Oxford University Press, USA (2000)
8. Garg, N., Schiebinger, L., Jurafsky, D., Zou, J.: Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences **115**(16), E3635–E3644 (2018)
9. Hawken, A., Munck, G.L.: Cross-national indices with gender-differentiated data: what do they measure? how valid are they? Social indicators research **111**(3), 801–838 (2013)
10. Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., Puranen, B., et al.: World values survey: Round six-country-pooled datafile 2010-2014. JD Systems Institute, Madrid (2014)
11. Kozlowski, A.C., Taddy, M., Evans, J.A.: The geometry of culture: Analyzing meaning through word embeddings. arXiv preprint arXiv:1803.09288 (2018)
12. Li, Q., Shah, S., Liu, X., Nourbakhsh, A.: Data sets: Word embeddings learned from tweets and general data. In: Eleventh International AAAI Conference on Web and Social Media (2017)
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26, pp. 3111–3119 (2013)
14. Mitra, A.: Establishment size, employment, and the gender wage gap. The Journal of Socio-Economics **32**(3), 317–330 (2003)
15. Rashotte, L.S., Webster Jr, M.: Gender status beliefs. Social Science Research **34**(3), 618–633 (2005)
16. Vincent, C.: Why do women earn less than men. CRDCN Research Highlight/RCCDR en évidence **1**(5), 1 (2013)
17. Williams, J.E., Best, D.L.: Sex and psyche: Gender and self viewed cross-culturally. Sage Publications, Inc (1990)
18. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340. ACM (2018)
19. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: 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 (2018)