

Investigating gender and racial-ethnic biases in sentiment analysis of language

Steven Zhou & Arushi Srivastava

To cite this article: Steven Zhou & Arushi Srivastava (2024) Investigating gender and racial-ethnic biases in sentiment analysis of language, Cogent Psychology, 11:1, 2396695, DOI: [10.1080/23311908.2024.2396695](https://doi.org/10.1080/23311908.2024.2396695)

To link to this article: <https://doi.org/10.1080/23311908.2024.2396695>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



[View supplementary material](#)



Published online: 29 Aug 2024.



[Submit your article to this journal](#)



Article views: 52



[View related articles](#)



[View Crossmark data](#)



This article has been awarded the Centre for Open Science 'Open Data' badge.



This article has been awarded the Centre for Open Science 'Open Materials' badge.

Investigating gender and racial-ethnic biases in sentiment analysis of language

Steven Zhou^a  and Arushi Srivastava^b

^aDepartment of Psychology, George Mason University, Fairfax, Virginia, USA; ^bSchool of Human Ecology, Tata Institute of Social Sciences, Mumbai, India

ABSTRACT

Recently, there has been an increase in text analysis and natural language processing for both research and applied practice, especially to quantify emotions in language (i.e. sentiment analysis). Building on different theories of how language and emotions interact and how these interactions differ by gender and race/ethnicity, our study assesses for bias in the use of common sentiment analysis tools (e.g. AFINN, NRC). Specifically, we focus on measurement bias and predictive bias between genders and races/ethnicities using a novel real-world dataset of participant interviews in a simulated multi-day team-based competition. There was no evidence of measurement bias by race/ethnicity, but there were some biases by gender; specifically, females tended to express higher mean levels and more variance in emotion. There was no evidence of predictive bias by gender or race/ethnicity, though the latter was marginally significant. We hope this study paves the way towards more inclusive and accurate analytical tools to help researchers reduce demographic biases in their research. These findings also hold importance for organizations in employing equitable tools to better understand the needs of their diverse customers and employees.

ARTICLE HISTORY

Received 5 June 2024
Revised 16 August 2024
Accepted 19 August 2024

KEYWORDS

Sentiment analysis; text analysis; emotion; language; bias

SUBJECTS


Quantitative Methods;
Cross-Cultural/
Multicultural Testing and
Assessment; General
Psychology; Social
Psychology

Recent years have witnessed an increase in using text data in psychological science, with sentiment analysis being one of the most popular methods, offering ample applications to fields such as psychology, business, consumer science, and communications (Yadav & Vishwakarma, 2020). Sentiment analysis is a form of text analysis that focuses on identifying the underlying sentiment or emotion for a given text (Medhat et al., 2014). One particular method involves a bag-of-words approach such that text is stemmed into individual words, then counted up according to a lexicon (i.e. dictionary) that assigns sentiment scores to each word. This approach to sentiment analysis is easy to apply and often used in applied settings such as analysis of consumer feedback, company culture surveys, and social media usage (Agarwal et al., 2011; Yang et al., 2010).

However, the use of sentiment analysis might pose risks of measurement bias when applied to human behavior, defined as the accuracy of an algorithm in reflecting genuine individual differences or similarities as opposed to 'systematic error that magnifies or diminishes such differences or similarities' (Tay et al.,

2022, p. 7). In other words, the use of a computer algorithm – in this case, sentiment analysis – should accurately reflect different emotions between individuals and groups and should not include systematic errors that systematically elevate or diminish the level of emotion captured by the algorithm. For example, sentiment analysis would be biased if it systematically and artificially scores men as having higher scores in a certain emotion compared to women. Some of these biases may be inappropriate, and their consequences could be grave, especially when they concern high-stakes decision-making. For example, Amazon's use of machine learning algorithms for employee selection was found to be biased against women engineers (Dastin, 2018). The application of sentiment analysis without consideration for gender or racial-ethnic biases could lead to major problems, especially if sentiment scores are used in situations such as predicting consumer feedback (e.g. the most negative consumer feedback gets flagged, and the consumer gets a discount code as an apology).

There has yet to be an investigation into gender and racial-ethnic biases of these popular sentiment analysis

CONTACT Steven Zhou  szhou9@gmu.edu

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

tools. In the present study, we begin by reviewing the literature on how language is used to express emotion, and how this might differ between demographic variables such as gender and race/ethnicity. Next, we discuss in detail the concept of statistical bias using sentiment analysis scores, reviewing the importance of measurement invariance and differential prediction and how it could lead to downstream effects of bias against underrepresented groups. We then present novel empirical data using a unique archival dataset consisting of reality TV competitions that were televised publicly over the course of several years. The text data comes from transcripts of each team member's verbal statements made before and immediately after each competition that they participated in. This novel dataset allows us to uniquely test for effects in a real-world environment where people are providing their honest expressions of emotion before and after competing in a real-world game with real outcomes. Put together, this study contributes important advances in the theory of how language is used to express emotion, the practical use (and misuse) of sentiment analysis, and the unique large dataset drawn from real-world experiences to maximize external validity of findings.

Background

There is a vast history of sociolinguistic research that has long since established that language use for emotion differs between gender and racial-ethnic groups. Here, we focus on the verbal expression of emotion, defined as 'psychological states that are experienced as coordinated patterns of physiology, behavior, and thoughts that occur within certain types of situations, and which are described with certain emotion category words (e.g. in English, "anger," "disgust," "fear," "happiness," "sadness")' (Lindquist et al., 2016, p. 580). For example, women tend to show greater emotional expressivity, specifically for positive and prosocial emotions (e.g. empathy, joy, enthusiasm) and other emotions that express powerlessness, like fear, sadness, and shame (Brody & Hall, 1993; Kring & Gordon, 1998). Moreover, differences between the sexes are larger for expressions than experiences; women are especially likely to show their emotions to a greater extent than men (Fischer & Manstead, 2000; LaFrance & Banaji, 1992). At the same time, there is evidence suggesting that these gender differences are not as apparent as originally thought. For example, in a meta-analysis, reports of pride showed no gender differences despite the stereotypes that men display more pride compared to women (Else-Quest et al., 2012).

There is likewise a substantial body of research on the nature and extent to which cultural and racial-ethnic differences exist. For example, Hwang and Matsumoto (2012) examined how people from different ethnic groups express and modify their perceived emotions in relationships; they found that Asian Americans and European Americans differed in how they perceive and moderate their emotions. Similarly, Hatfield et al. (2009) found ethnic differences in emotion expression in people of Chinese, Filipino, and Japanese ancestry, as they had different ideologies regarding how people should deal with strong emotions regarding expression in intimate relationships. In another study by Consedine and Magai (2002), four minority ethnic groups within the US, namely, African Americans, West Indiana (Jamaican) and Eastern Slavs (Russians and Ukrainians) from the former Soviet Republic and US-born European Americans showed significant variations in 9 of 10 basic emotions measured by differential emotions scale which favored a cultural-developmental interpretation.

However, these gender and racial-ethnic differences in how individuals use language to express emotions have essentially been ignored when applying sentiment analysis tools. Here, we focus on the bag-of-words approach, a simple text analysis method that counts up words within a block of text according to a lexicon (i.e. dictionary) that assigns sentiment scores to each word. One popular lexicon is the AFINN lexicon, which assigns a sentiment score ranging from -5 (negative) to +5 (positive) for each word; this can then be aggregated across the block of text to estimate the sentiment of the person who wrote or spoke that text (Nielsen, 2011). Another popular lexicon is the NRC lexicon, which assigns a specific emotion to each word; this can then be counted up across a block of text (e.g. 10 out of the 100 words expressed the 'angry' emotion) to estimate the emotions displayed by the person who wrote or spoke that text (Mohammad & Turney, 2010). Thus, both the AFINN and NRC methods of sentiment analysis produce a numerical score for that block of text that indicates how positive or negative that block of text is, and how many words that block of text uses that belong to one of eight different distinct emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust).

When these numerical sentiment scores are used in any sort of multivariate analysis with other variables of interest (e.g. team performance), there may be statistical biases at play (Tay et al., 2022). The present study focuses on two key areas of bias: measurement invariance and differential prediction. Measurement

invariance refers to when a psychological measure produces scores that *mean the same thing* between different demographic groups (Vandenberg & Lance, 2000). In a simple example, a measure is invariant if a score of 3.5 out of 5 on anxiety represents the same amount of anxiety for males compared to females. If a measure is non-invariant, then the scores are biased; two individuals of different groups who have the same amount of anxiety would get different scores. On the other hand, differential prediction refers to differences in the slopes and intercepts of regression equations when the psychological measure is used to predict a desired outcome (Berry, 2015). In a simple example, a measure shows differential prediction if, for one group of people, the slope is larger (i.e. stronger criterion-related validity) and/or the intercept is larger (i.e. higher base rate scores on the measure). In both cases, these forms of statistical bias result in inaccurate scoring, whether it is inaccuracy in the score's representation of the target construct or in the use of the score to predict an outcome. This can lead to further downstream effects. For example, if a measure is non-invariant (i.e. biased), then underrepresented groups might systematically receive higher scores on a negative construct than they would in reality on a non-biased measure. Similarly, if a measure shows differential prediction (i.e. biased), it might have poorer criterion-related validity among an underrepresented group than it would among other groups. Subsequently, any decisions made based on these scores could lead to systematic bias against the underrepresented group.

Thus, our study is among the first to directly investigate the extent to which gender and racial-ethnic biases exist in the use of popular sentiment analysis tools, empirically tested on a novel real-world archival dataset of reality TV interviews in a competitive game setting. Here, our focus is on identifying evidence of measurement non-invariance and differential prediction using sentiment scores based on the following research questions:

1. Is gender bias present in the use of sentiment analysis in terms of measurement invariance and differential prediction?
2. Is racial-ethnic bias present in the use of sentiment analysis in terms of measurement invariance and differential prediction?

Methods

This is an archival dataset consisting of reality TV competitions that were televised publicly over several years, and the data was obtained through the

recordings. The advantage to using this archival, real-world dataset is that it was collected independently of the research question at hand and features real-world individuals engaging in team-based win vs. lose challenges, offering their raw, unfiltered (apart from TV editing) thoughts as they experienced the team-based game and subsequent consequences. In other words, it is a high-fidelity situation for studying how people use language to express emotions in a team-based environment with high stakes (i.e. up to \$1,000,000 for the first-place winner), which suggests much stronger external validity and generalizability to real-life behavior compared to the traditional lab experiment used in most psychological research (Debouwere & Rosseel, 2022). The reality TV show features two or three teams competing head-to-head in a series of 6 to 8 games with team-level outcomes (i.e. one team loses and the other teams win). Each team starts with between 6 to 10 players; players have never met prior to their competition, and there is no pre-designated leader. The primary variable of interest is text data: transcripts of each team member's verbal statements made before and immediately after each game they participate in. In total, we have a sample size of 1253 individual participant observations nested in 17 teams.

Measures

First, we conducted sentiment analysis on each of the 1253 text observations. AFINN gives us an average sentiment score (ranging from -5 for negative to +5 for positive) for each observation (Nielsen, 2011), while NRC gives us a count of the total words found for each of eight different emotional categories (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust; Mohammad & Turney, 2010). Next, we used team-level outcome (dichotomous win/lose) as the outcome variable to be predicted by sentiment scores. Finally, we used the following demographic variables: presenting race-ethnicity and presenting gender. Because our dataset did not include exact gender identity and race-ethnicity, we used dichotomous coding (i.e. majority vs. minority race, male vs. female gender) based on gender and racial-ethnic presentation on the TV show.

Analysis

We examined potential gender and racial biases in sentiment analysis by conducting tests of measurement invariance and differential prediction. To test measurement invariance between NRC emotions

among race and gender groups, we fit a two-factor CFA model with the eight emotions loading onto latent factors of positive and negative affect based on the structure defined by the NRC lexicon (Mohammad & Turney, 2010). Specifically, anticipation, joy, trust, and surprise were coded as positive emotions (Mohammad & Turney, 2010), while anger, disgust, fear, and sadness were coded as negative emotions (Mohammad & Turney, 2010); surprise could be coded as both positive and negative (see Figure 1 for a visual explanation of the factor structure). We subsequently tested for measurement invariance using the moderated nonlinear factor analysis (MNLFA) method developed by Bauer (2017). This method has the advantage of testing specific differential item functioning (DIF) of each item (in this case, each emotion) as opposed to just testing the overall fit of the model; it is capable of testing more than one grouping variable at a time (in this case, both race and gender simultaneously). Briefly, the procedure involves establishing a baseline model to test for the impact of gender and race on the mean and variance parameters of the two factors, then testing the impact of gender and race on each specific factor loading parameter within the model (i.e. each of the eight emotions loading onto positive and negative affect). We refer interested readers to Bauer (2017) and Bauer et al. (2020) for details on the procedure. To test for differential prediction of AFINN lexicon predicting team outcome, we fit a multilevel logistic regression with the Level 1 predictor (AFINN) and the Level 2 outcome (team win or

loss), then adding gender and race as moderators. A significant moderation effect indicates differential prediction; that is, the ability of AFINN to predict team-level outcomes depends on the gender and/or race of the individual.

Results

Measurement invariance (RQ1a and RQ2a)

Following the MNLFA method (Bauer, 2017; Bauer et al., 2020), we first fit a two-factor CFA of the eight emotions loading onto positive and negative affected, which produced excellent fit: $\chi^2(18) = 70.147$, $p < 0.001$, robust CFI = 0.976, robust TLI = 0.963, robust RMSEA = 0.060, and SRMR = 0.035. Figure 1 depicts the baseline factor structure.

Table 1 depicts the impact of the two demographic variables on the mean positive and negative affect scores and the variance; significant results indicate that there are differences in mean scores and variance based on that demographic variable. The only significant demographic variable was gender. In other words, the structure of NRC emotions (i.e. loadings onto positive/negative affect and variance of emotion) was invariant between race-ethnicity but not gender. Thus, subsequent analyses focused on gender.

Next, the MNLFA procedure tested the impact of gender on each of the eight specific NRC emotions in separate models. There were significant differences in anger ($\chi^2(2) = 70.477$, $p < 0.001$) and fear ($\chi^2(2) = 20.015$, $p < 0.001$). In other words, anger and fear

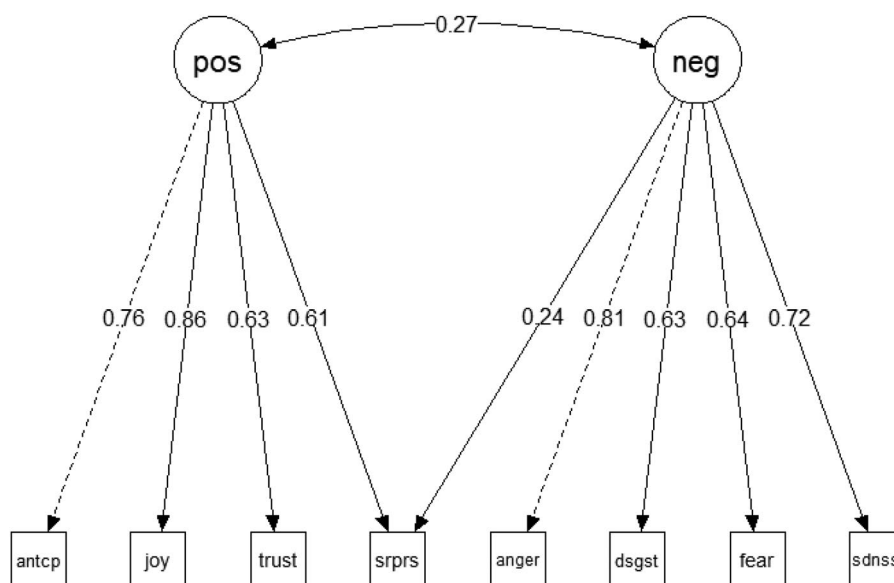


Figure 1. Baseline two-factor structure for measurement invariance testing.

Notes: standardized parameter values are shown. pos=positive affect, neg=negative affect, antcp=anticipation, joy=joy, trust=trust, srprs=surprise, anger=anger, dsgst=disgust, fear=fear, and sdnss=sadness.

Table 1. Mean and variance impact of gender and race on NRC emotions.

Mean impact of	Parameter	Est	SE	<i>p</i>	Interpretation
gender	positive	0.202	0.088	0.022	Females score higher on positive emotions
	negative	0.193	0.085	0.023	Females score higher on negative emotions
race	positive	0.089	0.109	0.414	
	negative	-0.127	0.087	0.145	
Variance impact of					
gender	variance	0.657	0.222	0.003	Females show more variance in emotions
race	variance	-0.283	0.174	0.103	

Table 2. Final model results from MNLFA.

	Est	SE	<i>p</i>	Interpretation
positive BY				
anticipation	0.035	0.005	0.000	invariant item, interpret normally
joy	0.037	0.005	0.000	invariant item, interpret normally
surprise	0.019	0.003	0.000	invariant item, interpret normally
trust	0.033	0.004	0.000	invariant item, interpret normally
negative BY				
anger	NA			DIF shown
disgust	0.013	0.002	0.000	invariant item, interpret normally
fear	NA			DIF shown
sadness	0.013	0.003	0.000	invariant item, interpret normally
surprise	0.009	0.005	0.058	invariant item, interpret normally
positive ON				
gender	0.182	0.083	0.027	females score higher on positive emotions
negative ON				
gender	0.354	0.161	0.028	females score higher on negative emotions
positive WITH				
negative	0.491	0.174	0.005	significant correlation between positive and negative emotion
variance impact				
gender_positive	0.358	0.155	0.020	females show more variance in positive emotions
gender_negative	1.405	0.447	0.002	females show more variance in negative emotions
factor loading DIF				
anger base	0.028	0.006	0.000	NA
anger by gender	-0.008	0.002	0.000	anger matters less for females in explaining negative emotions
fear base	0.020	0.005	0.000	NA
fear by gender	-0.006	0.002	0.003	fear matters less for females in explaining negative emotions

showed evidence of differential item functioning based on gender. The final model results are shown in Table 2. The factor loadings for anger and fear were weaker for females compared to males, meaning that anger and fear matter *less* among females in explaining negative emotions.

Differential prediction (RQ1b and RQ2b)

The baseline multilevel model of AFINN (Level 1) predicting team win/loss (Level 2) was significant: $\beta = -0.134$, $p=0.015$. In other words, as AFINN increased (i.e. more positive), the likelihood of team loss slightly decreased. When gender was added as a moderator, the model comparison between the baseline model and the model with gender was not significant: $\chi^2(2) = 0.416$, $p=0.812$. Thus, there is no evidence for differential prediction by gender. When race was added as a moderator, the model comparison between the baseline model and the model with race was marginally significant: $\chi^2(2) = 5.884$, $p=0.053$. Thus, there is potentially some evidence for differential prediction by race/ethnicity.

Discussion

The results show mixed evidence for gender and racial-ethnic bias in the use of language to express emotion. In terms of measurement invariance, there was no evidence of differences by race-ethnicity. However, there were some differences by gender: specifically, females tended to express higher mean levels of both positive and negative emotion and more variance in both positive and negative emotion. Moreover, the factor loadings for anger and fear (onto negative emotion) were weaker for females compared to males. In terms of differential prediction, there was no evidence of bias based on gender. However, the model including race-ethnicity as a moderator was marginally significant, which suggests some potential effect such that the ability of sentiment to predict team outcomes varies based on race/ethnicity.

The results from the measurement invariance tests highlighting gender differences imply that women express higher levels of positive and negative emotions and a broader range of emotions than men. Additionally, anger and fear matter less among females in explaining

negative emotions. This aligns with decades of previous research that has shown gender variations in emotion expression (e.g. Durik et al., 2006; Hatfield et al., 2009). However, our finding on the factor loadings of anger and fear opens the door to further investigation; prior research has focused on how females express *more amount* of emotions like anger and fear (Else-Quest et al., 2006). The finding that anger and fear ‘matter less’ in explaining negative emotions points to further research to investigate if *more amount* of these emotions, as prior studies suggest, lead to more downstream experiences of negative emotions. Alternatively, even if females express more of these emotions, they may not reflect the same amount of negativity compared to males.

In terms of differential prediction, having no evidence for significant variation in gender implies that the sentiment score (AFINN) predicts team performance outcomes similarly for men and women. This suggests that there is no bias in predicting outcomes based on gender. However, we find potential differences in team performance outcomes for race-ethnicity, implying that the ability of sentiment scores to predict outcomes may vary across different ethnic and racial groups. This finding aligns with prior literature highlighting ethnic differences in emotion expression (Hatfield et al., 2009; Hwang & Matsumoto, 2012). Although this evidence is not strong, it raises the possibility that sentiment analysis might not be equally effective for predicting team outcomes across different racial-ethnic groups. This paves the way for future research to delve more thoroughly into the efficacy and fairness of using sentiment analysis as a predictor of various outcomes, such as performance.

Recent research in language, gender and emotion studies echo similar findings to our analyses. In a study on emotion expression, gender, and voter reactions during German televised debates, Boussalis et al. (2021) found that the former female Vice Chancellor of Germany expressed less anger than her male counterparts. Voters tended to punish the Vice Chancellor for expressing anger and reward her for expressing happiness, with the opposite effects for the male counterparts. Voters also responded positively when she expressed more emotions, with a similar trend for other women candidates in the minor party debates. This finding somewhat aligns with our findings and previous research on how women are expected to express more cordial emotions and stray away from authoritative emotions like anger. Another study examining the gender differences in the use of language using a keyword-based approach found that women used more positive emotions (more words related to

joy), and men used more words in the category of anger and swear words. The researchers also highlighted that women did not necessarily use *more* words related to emotions than men. Building on to our finding of the differences between racial-ethnic groups, a relevant study by Jackson et al. (2019) underscores that emotion concepts have different patterns of associations and ways of expression across different languages and cultures and highlights the possibility that emotion experiences vary across cultures.

Limitations and future directions

First and foremost, our data and methods were limited by the use of the *Survivor* TV dataset and the application of simpler bag-of-words sentiment analysis methods. While the reality TV setting offered unique opportunities to measure real-world human behavior outside of a lab setting and in high-stakes competition, future studies can utilize different settings such as everyday communications, transcripts from court hearings or business meetings, or public speeches to investigate if differences exist based on the setting of communication. Additionally, more recent developments in sentiment analysis use sophisticated models such as BERT (Hoang et al., 2019) and large language models (e.g., GPT). However, these can be difficult to implement without advanced training in natural language processing and coding, especially if one wishes to fine-tune the GPT model to use it for sentiment, and thus, they are still less accessible to the general population. As such methods become more popular, future studies can test for biases in these modern algorithms.

Second, while our study offers an initial investigation into measurement bias from sentiment analysis, it does not take the next step of disaggregating measurement bias from true subgroup differences. In other words, our study suggests that women score higher in positive and negative emotions and score in a broader range of emotions than men. It does not determine if these differences are due to measurement bias in the algorithm – and thus should be statistically eliminated – or due to true subgroup differences by gender. Future research can explore the reasons behind the variations and potential biases found in our analyses to identify specific sources or causes, specifically to separate measurement bias from true subgroup differences.

Conclusion

Building on a robust literature of emotions, language, and demographic differences in the expression of

emotion, our study offers a unique investigation into how popular sentiment analysis of text could perpetuate or exacerbate existing demographic biases. We use a novel real-world dataset of reality TV transcriptions and apply measurement invariance and differential prediction analyses to demonstrate some evidence of gender and racial-ethnic differences; that is, the use of language to express emotion, as measured by AFINN and NRC, differs somewhat between gender groups and racial-ethnic groups.

This study makes several key contributions by highlighting the potential flaws in widely used methodologies in emotion detection, including studies as recent as the past five years, thus suggesting the need for adjustments. It also sheds light on how different gender and ethnic-racial groups differ in their use of language to express emotions, allowing for more tailored and fair applications of sentiment analysis across diverse populations. These findings may benefit researchers and academicians by helping them consider demographic biases in their analysis by rigorously testing and adjusting their tools for accuracy and fairness. In terms of organizational setting, it may be helpful in equitable decision-making, enhanced customer feedback analysis, and better market research and consumer insights. We hope this study inspires future researchers to more thoroughly examine bias in new and exciting text analysis technologies and utilize novel and interesting data sources beyond the traditional student or online panel samples.

Open Scholarship



This article has earned the [Center for Open Science](#) badges for Open Data and Open Materials through Open Practices Disclosure. The data and materials are openly accessible at <https://osf.io/ynvq9/>.

Disclosure statement

No potential conflict of interest was reported by the author(s).

About the authors

Steven Zhou completed his PhD in industrial-organizational psychology from George Mason University in 2024, where his research focused on leadership, quantitative methods, and careers. He also regularly teaches quantitative methods at the undergrad and grad level, consults with various non-profits, and serves in academic administration.

Arushi Srivastava is a first year PhD student in Management at Rady School of Management, University of California, San Diego. Her research focuses on exploring how emotions and emotion regulation shapes personal and relational wellbeing, with an emphasis on the cultural differences that may shape our behavior.

ORCID

Steven Zhou  <http://orcid.org/0000-0003-0710-7065>

Data availability statement

All data and code are available on OSF for future researchers to explore and replicate our findings: <https://osf.io/ynvq9>

References

- Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011). Sentiment analysis of twitter data. *Proceedings of the Workshop on Language in Social Media (LSM 2011)* (pp. 30–38). June. <https://aclanthology.org/W11-0705.pdf>
- Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, 22(3), 507–526. <https://doi.org/10.1037/met0000077>
- Bauer, D. J., Belzak, W. C., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: a Multidisciplinary Journal*, 27(1), 43–55. <https://doi.org/10.1080/10705511.2019.1642754>
- Berry, C. M. (2015). Differential validity and differential prediction of cognitive ability tests: Understanding test bias in the employment context. *Annual Review of Organizational Psychology and Organizational Behavior*, 2(1), 435–463. <https://doi.org/10.1146/annurev-orgpsych-032414-111256>
- Boussalis, C., Coan, T. G., Holman, M. R., & Müller, S. (2021). Gender, candidate emotional expression, and voter reactions during televised debates. *American Political Science Review*, 115(4), 1242–1257. <https://doi.org/10.1017/S0003055421000666>
- Brody, L. R., & Hall, J. (1993). Gender and emotion. In M. Lewis & J. Haviland (Eds.), *Handbook of emotions* (pp. 447–461). Guilford Press.
- Consedine, N. S., & Magai, C. (2002). The uncharted waters of emotion: Ethnicity, trait emotion and emotion expression in older adults. *Journal of Cross-Cultural Gerontology*, 17(1), 71–100. <https://doi.org/10.1023/A:1014838920556>
- Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. Retrieved from <https://www.reuters.com/article/us-amazon-com-jobsautomation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-womenidUSKCN-1MK08G>
- Debrouwere, S., & Rosseel, Y. (2022). The conceptual, cuning, and conclusive experiment in psychology. *Perspectives on Psychological Science: a Journal of the*

- Association for Psychological Science*, 17(3), 852–862. <https://doi.org/10.1177/17456916211026947>
- Durik, A. M., Hyde, J. S., Marks, A. C., Roy, A. L., Anaya, D., & Schultz, G. (2006). Ethnicity and gender stereotypes of emotion. *Sex Roles*, 54(7–8), 429–445. <https://doi.org/10.1007/s11199-006-9020-4>
- Else-Quest, N. M., Higgins, A., Allison, C., & Morton, L. C. (2012). Gender differences in self-conscious emotional experience: A meta-analysis. *Psychological Bulletin*, 138(5), 947–981. <https://doi.org/10.1037/a0027930>
- Else-Quest, N. M., Hyde, J. S., Goldsmith, H. H., & Van Hulle, C. A. (2006). Gender differences in temperament: A meta-analysis. *Psychological Bulletin*, 132(1), 33–72. <https://doi.org/10.1037/0033-2909.132.1.33>
- Fischer, A. H., & Manstead, A. S. (2000). The relation between gender and emotions in different cultures. In A. H. Fischer (Ed.), *Gender and emotion: Social psychological perspectives* (pp. 71–94). Cambridge University Press.
- Hatfield, E. C., Rapson, R. L., & Le, Y. C. L. (2009). Ethnic and gender differences in emotional ideology, experience, and expression. *Interpersona: An International Journal on Personal Relationships*, 3(1), 30–57. Retrieved from <https://dialnet.unirioja.es/servlet/articulo?codigo=5216156> <https://doi.org/10.5964/ijpr.v3i1.31>
- Hoang, M., Bihorac, O. A., & Rouces, J. (2019). Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics* (pp. 187–196). <https://aclanthology.org/W19-6120>
- Hwang, H. S., & Matsumoto, D. (2012). Ethnic differences in display rules are mediated by perceived relationship commitment. *Asian American Journal of Psychology*, 3(4), 254–262. <https://doi.org/10.1037/a0026627>
- Jackson, J. C., Watts, J., Henry, T. R., List, J. M., Forkel, R., Mucha, P. J., Greenhill, S. J., Gray, R. D., & Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science (New York, N.Y.)*, 366(6472), 1517–1522. <https://doi.org/10.1126/science.aaw8160>
- Kring, A. M., & Gordon, A. H. (1998). Sex differences in emotion: Expression, experience, and physiology. *Journal of Personality and Social Psychology*, 74(3), 686–703. <https://doi.org/10.1037/0022-3514.74.3.686>
- LaFrance, M., & Banaji, M. (1992). Toward a reconsideration of the gender emotion relationship. In M. S. Clark (Ed.), *Emotions and social behavior* (pp. 178–202). Sage.
- Lindquist, K. A., Gendron, M., Satpute, A. B., Lindquist, K., et al. (2016). Language and emotion. In L. F. Barrett (Eds.), *Handbook of emotions* (pp. 579–594). Guilford Press.
- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113. <https://doi.org/10.1016/j.asej.2014.04.011>
- Mohammad, S., & Turney, P. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. June. <https://saifmohammad.com/WebDocs/Mohammad-Turney-NAACL10-EmotionWorkshop.pdf>
- Nielsen, F. A. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages*. <http://arxiv.org/abs/1103.2903>
- Tay, L., Woo, S. E., Hickman, L., Booth, B. M., & D'Mello, S. (2022). A conceptual framework for investigating and mitigating machine-learning measurement bias (MLMB) in psychological assessment. *Advances in Methods and Practices in Psychological Science*, 5(1), 251524592110613. <https://doi.org/10.1177/25152459211061337>
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4–70. <https://doi.org/10.1177/109442810031002>
- Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review*, 53(6), 4335–4385. <https://doi.org/10.1007/s10462-019-09794-5>
- Yang, C. C., Tang, X., Wong, Y. C., & Wei, C. P. Drexel University. (2010). Understanding online consumer review opinions with sentiment analysis using machine learning. *Pacific Asia Journal of the Association for Information Systems*, 2(3), 73–89. <https://doi.org/10.17705/1pais.02305>