Analysis of Gender Bias in Social Perception and Judgement Using Chinese Word Embeddings

Jiali Li[1]†, **Shucheng Zhu**[2]†, **Ying Liu**[2]‡, **Pengyuan Liu**[1,3]‡
[1]School of Information Science, Beijing Language and Culture University, Beijing, China
[2]School of Humanities, Tsinghua University, Beijing, China
[3]National print Media Language Resources Monitoring & Research Center,
Beijing Language and Culture University, Beijing, China
lijiali9925@163.com, zhu_shucheng@126.com,
yingliu@tsinghua.edu.cn, liupengyuan@pku.edu.cn

Abstract

Gender is a construction in line with social perception and judgment. An important means of this construction is through languages. When natural language processing tools, such as word embeddings, associate gender with the relevant categories of social perception and judgment, it is likely to cause bias and harm to those groups that do not conform to the mainstream social perception and judgment. Using 12,251 Chinese word embeddings as intermedium, this paper studies the relationship between social perception and judgment categories and gender. The results reveal that these grammatical gender-neutral Chinese word embeddings show a certain gender bias, which is consistent with the mainstream society's perception and judgment of gender. Men are judged by their actions and perceived as bad, easily-disgusted, bad-tempered and rational roles while women are judged by their appearances and perceived as perfect, either happy or sad, and emotional roles.

1 Introduction

One of the main ways to construct gender in society is through languages. People's languages towards infants of different genders can well illustrate the gender construction of languages as a medium. When people believe that infants are female, they talk to them more gently. When people believe that infants are male, they handle infants more playfully. Through these differential treatments, boys and girls finally learn to be different (Eckert and McConnell-Ginet, 2013). As the boys and girls grow up, they start to perform the "correct" gender manners to be consistent with the gender judgment and perception of mainstream society. In other words, gender possesses performativity (Butler, 2002). As a result, in the process of construction repetition reinforcement, gender gradually solidifies the differences that should not be caused by gender and may cause unexpected biases and harms. The process is always through languages which represent the mainstream social judgment and perception.

As an analytic language, Chinese does have referential gender and lexical gender, such as "她" means "she" in referential gender and "爸爸" means "father" in lexical gender. However, Chinese lacks grammatical gender, comparing to French, Spanish and some of the fusional languages (Cao and Daumé III, 2020). As a result, it is difficult to find explicit and quantitative clues between gender and categories in social perception and judgement in Chinese. Word embedding is powerful and efficient in Natural Language Processing (NLP). Therefore, using word embeddings to find the implicit gender bias in Chinese can be an appropriate tool to analyze the associations between gender and categories in social perception and judgement. To make it clear, we define four categories of social perception and judgment and the linguistic features that can measure their gender bias, as shown in Table 1.

In this paper, we first gave our definition of gender bias. Then, by using semantic similarity, the implicit gender bias was measured in 12,251 Chinese word embeddings. Examples articulate that this measurement can capture the gendered word embeddings in a language without grammatical gender. Then, part-of-speech, sentiment polarity, emotion category, and semantic category were labeled to each word. We analyzed the relationships between gendered word embeddings and linguistic features to find the associations between gender and different categories in social perception and judgement. Results showed that we perceive and judge men and women with different social categories. Men are judged by their actions and perceived as bad, easily-disgusted, bad-tempered

†Equal contribution.
‡Corresponding authors.

Proceedings of the The 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 8 - 16

Category	Definition	Linguistic Metrics
Activity	To what extent do social perception or description of a person relate one's gender to appearance or action.	Part-of-speech
Sentiment Polarity	To what extent do social perception or judgment of a person relate one's gender to positive or negative sentiment.	Sentiment Polarity
Emotion Category	To what extent do social perception or judgment of a person relate one's gender to specific emotion categories, such as anger, happiness and sadness.	Emotion Category
Content	To what extent do social perception or judgment of a person relate one's gender to specific topics, such as psychology, state and abstraction.	Semantic Category

Table 1: Definitions and linguistic features of 4 categories of social perception and judgement

and rational roles while women are judged by their appearances and perceived as perfect, either happy or sad, and emotional roles. This method is neat, while it offers a quantitative view to study the relationship between gender and different categories in perception and judgement in Chinese society and culture.

2 Bias Statement

In this paper, we study stereotypical associations between gender and different categories in social perception and judgment through Chinese word embeddings. Most of the Chinese words are grammatical gender-neutral. However, if the Chinese word embeddings show gender differences in different categories of part-of-speech, sentiment polarity, emotion category and semantic category, it may show that these gender-neutral word embeddings represent our stereotypes towards different genders. For example, we always judge a woman by her appearance but judge a man by his action. Although these stereotypical generalizations may not be negative, once these stereotypical representations are used in downstream NLP applications, the system may ignore, or even do harms to those people who are not consistent with the mainstream social perception and judgement of gender. Hence, this stereotypical association can be regarded as bias which may cause representational harms (Blodgett et al., 2020). In other words, the uniqueness between person and person is erased, and the system only retains gender differences. The ideal state is that people will not be treated unfairly because of their genders, especially to those are not consistent with the mainstream social perception and judgement of gender, and the system should not emphasize certain characteristics of a person according to one's gender.

3 Dataset

The Chinese word embeddings[1] we selected were pre-trained with Baidu Encyclopedia Corpus, using word2vec model and the method of Skip-Gram with Negative Sampling (SGNS). The size of Baidu Encyclopedia corpus is 4.1GB and the corpus contains 745M tokens (Li et al., 2018). Baidu Encyclopedia is an open online encyclopedia like Wikipedia, with entries covering almost all areas of Chinese knowledge. The encyclopedia characteristic of Baidu Encyclopedia determines that the language it uses is more objective and gender-neutral. The total amount of the word embeddings is 636,013 and each word embedding contains 300 dimensions. After labelling part-of-speech, sentiment polarity, emotion category, and semantic category, only 12,376 words contain all the information we need. Then, we calculated Odds Ratio (OR) values of each word and only selected those within three standard deviations from the mean. At last, we kept 12,251 word embeddings as our dataset. Almost all the words are gender-neutral as Chinese is a language without grammatical gender. Different token numbers of Chinese word embeddings in part-of-speech, sentiment polarity, emotion category, and semantic category are shown in Table 2.

Part-of-speech. The part-of-speech labels were selected from Affective Lexicon Ontology[2] (Xu et al., 2008). As we all know, the part-of-speech of many Chinese words may change in different contexts. However, the Chinese word embedding we chose is not contextualized. Among the 12,251 words in our dataset, only 37 words are multi-category words. We thought that the number is small and would not affect the results and analysis. Therefore, we chose one of the tags in Affective

[1] https://github.com/Embedding/Chinese-Word-Vectors

[2] http://ir.dlut.edu.cn/info/1013/1142.htm

Part-of-speech	**Adjective**	**Adverb**	**Idiom**	**Noun**	**Prep**	**Verb**	**Net-words**	**Total**
Tokens	3586	39	3417	2618	63	2514	14	12251
Sentiment Polarity	**Positive**	**Negative**	**Neutral**	**Both**				**Total**
Tokens	4675	4628	2908	40				12251
Emotion Category	**Disgust**	**Good**	**Sadness**	**Fear**	**Anger**	**Happiness**	**Astonishment**	**Total**
Tokens	4694	4858	917	538	169	99	976	12251
Semantic Category	**Activity**	**Action**	**Object**	**Association**	**Aid language**	**Characteristic**	**Honorific language**	**Total**
Tokens	2037	177	367	279	167	4418	22	
	Person	**State**	**Time and space**	**Abstraction**	**Psychology**			12251
	829	1009	1561	1252	133			

Table 2: Word embedding tokens labeled in different linguistic features in our dataset

Lexicon Ontology as its part-of-speech label for analysis. There are 7 labels of the part-of-speech. To balance the amount for analysis, we only chose the words labeled "noun", "verb" and "adjective" to compute and analyze. Here, we assume that nouns and adjectives are related to the appearance of what people perceive and judge, while verbs are related to action.

Sentiment Polarity. Affective Lexicon Ontology also offers 4 labels of the sentiment polarity, and we chose the words labeled "positive" and "negative" to analyze.

Emotion Category. According to Ekman's six basic emotions (Ekman, 1999) and the characteristic of Chinese, the Affective Lexicon Ontology offers 7 labels for the sentiment category: "good" (including "respect", "praise", "believe", "love" and "wish" to make a more detailed division of commendatory emotion), "anger", "disgust", "fear", "happiness", "sadness", and "astonishment".

Semantic Category. Our semantic category labels are from HIT IR-Lab Tongyici Cilin (Extended)[3]. It organized all the entries in a tree-like hierarchy, and divided the words into 12 semantic categories. We only chose the top 5 categories related to human and with the largest number of tokens to analyze: "abstraction", "activity", "characteristic", "state" and "psychology".

4 Experiments

In this section, we will illustrate the methodology to analyze the gendered word embeddings and how they are associated to different categories in our social perception and judgement. We first used semantic similarity and odds ratio to evaluate each word embedding. Then, independent-samples t test, one-factor Analysis of Variance (ANOVA) and Kruskal-Wallis test were used respectively to analyze the relationships between gender and categories in social perception and judgement.

Masculine Words	Meaning	Feminine Words	Meaning
爸爸	dad	妈妈	mom
父亲	father	母亲	mother
姥爷	mother's father	姥姥	mother's mother[4]
外公	mother's father	外婆	mother's mother[5]
爷爷	father's father	奶奶	father's mother
哥哥	elder brother	姐姐	elder sister
弟弟	younger brother	妹妹	younger sister
儿子	son	女儿	daughter
男友	boyfriend	女友	girlfriend
叔叔	uncle	阿姨	aunt
他	he	她	she
男	male	女	female
男人	men	女人	women
男子	man	女子	woman
男士	Mr.	女士	Ms.
先生	Sir	小姐	Miss
男孩	boy	女孩	girl
男性	males	女性	females

Table 3: Gendered Words

Semantic Similarity. We first selected and translated 14 masculine words and corresponding 14 feminine words as Gendered Words G into Chinese from related study in English (Nadeem et al., 2020), showed in Table 3. These words are lexical gender words or referential gender words in Chinese. Then, we calculated the cosine similarity as the semantic similarity S between each word embedding in our dataset W and the word embeddings of Gendered Words G according to equation 1. Here, n means the total dimension of each word embedding. We took the mean cosine similarity between one W and the total Feminine word embeddings as the Feminine Similarity S_f. Masculine Similarity S_m of one W is as the same. The closer to 1 the value of S is, the word W is more masculine or feminine.

$$S = \frac{\sum_{i=1}^{n} W_i \times G_i}{\sqrt{\sum_{i=1}^{n}(W_i)^2} \times \sqrt{\sum_{i=1}^{n}(G_i)^2}} \quad (1)$$

[3] https://github.com/Xls1994/Cilin

[4] "姥爷" and "姥姥" are usually used in northern China.

[5] "外公" and "外婆" are usually used in southern China.

Odds Ratio. OR (Szumilas, 2010) was used to calculate the Gendered value OR of each word embedding W in our dataset as equation 2 shows. Here, N is the total number of word embeddings in our dataset. To facilitate the test, we selected OR values within three standard deviations from the mean and normalized all data to $OR_G \in [-1, 1]$.

$$OR(w) = \frac{S_m(W)}{\sum_{j=1}^{N} S_m(W_j)} / \frac{S_f(W)}{\sum_{j=1}^{N} S_f(W_j)} \quad (2)$$

The closer the OR_G is to 1, the more masculine the word is. The closer the OR_G is to -1, the more feminine the word is.

Independent-samples T Test. On sentiment polarity, we conducted an independent-sample t test of OR_G value to explore the relationship between gender and sentiment polarity in social perception and judgement as the variances are homogeneous.

One-factor ANOVA. On part-of-speech, we conducted one-factor ANOVA of OR_G value to explore the relationship between gender and activity in social perception and judgement as the different token numbers in part-of-speech are sufficient and approximate.

Kruskal-Wallis test. On the categories of emotion category and semantic category, we conducted Kruskal-Wallis test of OR_G value respectively to explore the relationships between gender and emotion category and content in social perception and judgement as the variances in these two categories are different and the token numbers vary widely.

5 Results

Gendered Word Embeddings. We selected the top 5 masculine and feminine word embeddings of grammatical gender-neutral words according to the OR_G value showed in Table 4. It is clear to see that the masculine words are related to "war" and "power" and the feminine words are related to "flower" and "beauty" which conforms to our stereotypes of gender. It indicates our measurement can detect the implicit gender bias in word embeddings of the language without grammatical gender.

Gender and Activity. We define activity as the extent to which we perceive or describe a person's gender in relation to one's appearance or action. Here, we think that verbs can represent perceiving and describing a person's action, and nouns and adjectives can represent perceiving and describing a person's appearance. Figure 1(a) shows that verbs (M=0.022) are more masculine than nouns (M=0.003) and adjectives (M=-0.064) and they have significant differences (p<0.001). It means that in social perception and judgment, we associate actions with men, appearances with women. It may indicate that we always perceive a woman with her appearance and judge a man by his action (Caldas-Coulthard and Moon, 2010). Sociolinguistic clues support this conjecture. Appearance is seen as applicable to the female gender category as there are subcategories elaborated specifically for women far more than men (Eckert and McConnell-Ginet, 2013). This supports that our society emphasizes appearance on women rather than men. Other studies also show that we use positive adjectives to describe a woman's body rather than a man (Hoyle et al., 2019). The most representative example is in mate selection. Men care much about women's appearance and women care much about men's power, status and wealth (Baker, 2014). Once man-action and woman-appearance associations are established, it may cause gender bias. The systems emphasize a woman's appearance over her other strengths, which may hurt women who are less attractive.

Word	Meaning	Part-of-speech	OR_G
所向披靡	invincible	idiom	1
戎马	army horse	noun	0.9985
让位	abdicate	verb	0.9968
广开言路	open communication	idiom	0.9918
死守	defend to death	verb	0.9906
盛开	bloom	verb	-1
婵娟	moon	adjective	-0.9933
火树银花	Hottest Silver	idiom	-0.9927
并蒂莲	Twin flowers	idiom	-0.9879
天仙	fairy	noun	-0.9811

Table 4: The top 5 masculine and feminine word embeddings of grammatical gender-neutral words according to the OR_G value

Gender and Sentiment Polarity. Figure 1(b) shows that positive words (M=-0.017) are more feminine than negative words (M=0.034) and they have significant difference (p<0.001). This associates men with negative sentiments and women with positive ones. This may imply that in our society, we perceive women in a positive way and we can perceive men in a negative way. It can be reflected fully in children's literature which al-

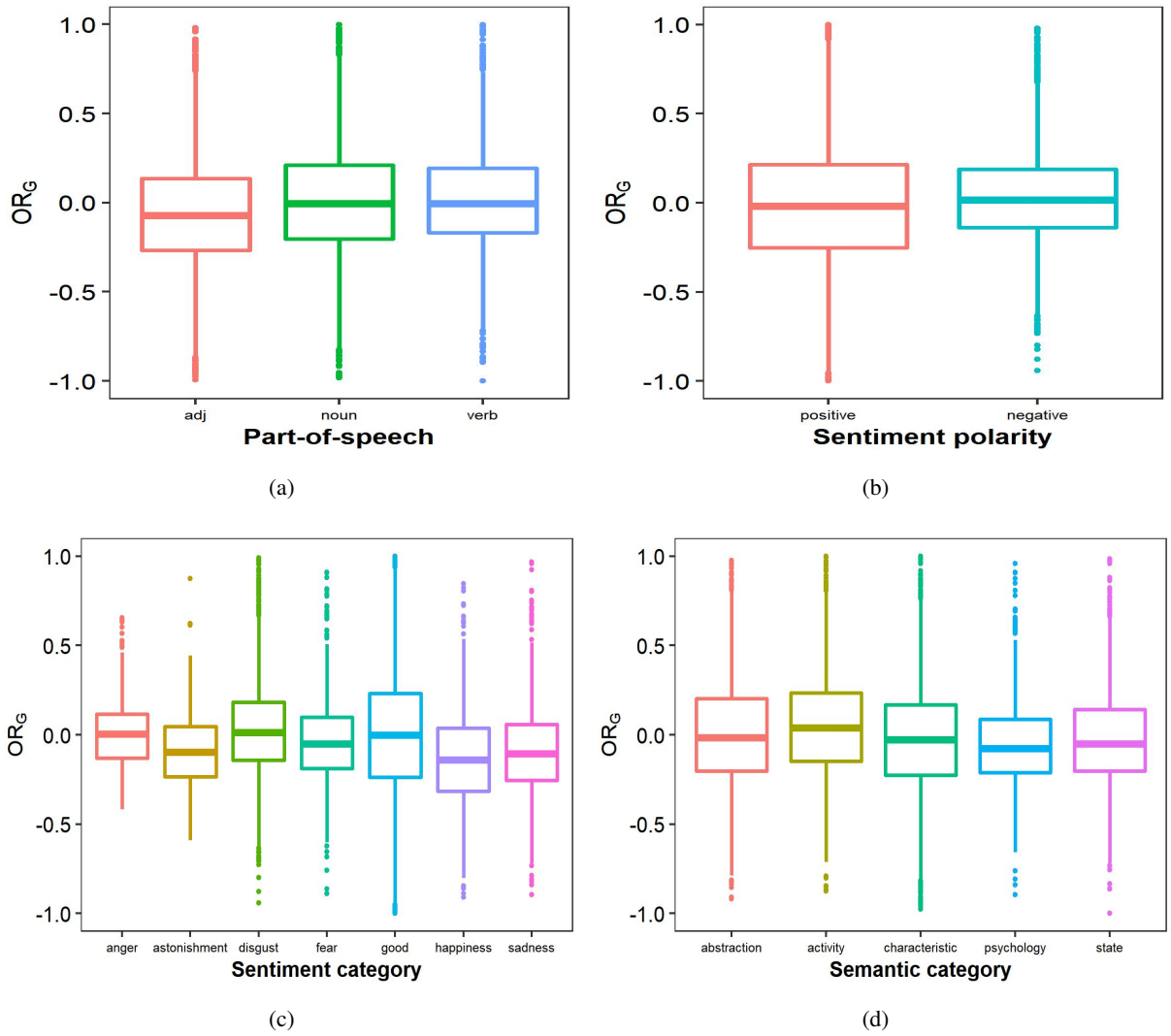

Figure 1: The distribution of OR_G in part-of-speech, sentiment polarity, emotion category, and semantic category

ways portrays "a good girl" and "a bad boy" (Peterson and Lach, 1990; Stevinson Hillman, 1974; Kortenhaus and Demarest, 1993). This point can be explained by the different gender views on compliments. Women are more likely to compliment and be complimented than men, because for women, compliments strengthen their solidarity with others in the communities of practice. However, complimenting men can challenge a men's authority and power because complimenting a man implies that he is being judged (Tannen, 1991; Holmes, 2013). Over time, women tend to develop a steady bond with positive sentiments. This seems to be a protection for women, but it is actually a benevolent sexism (Glick and Fiske, 2001). The negative man image indicates that we have a certain tolerance to man, while the positive woman image is more like a bondage to women. We expect women to be gentle and submissive all the time, while men can be negative and aggressive.

Gender and Emotion Category. Figure 1(c) shows that from the most masculine to the most feminine, the emotion categories are disgust (M=0.030), anger (M=0.025), good (M=-0.003), fear (M=-0.025), astonishment (M=-0.083), sadness (M=-0.089), and happiness (M=-0.130). Disgust and anger emotions have significant differences with other emotions (p<0.05). It indicates that we associate disgust and anger emotions with men rather than women. Sadness and happiness emotions have significant difference with other emotions (p<0.05). It indicates that we associate happiness and sadness emotions with women rather than men. Thus, in our social perception and judgment, men may be viewed with negative emotions,

such as anger and disgust, while women are either happy or sad. In movies and books, whether women are sad and happy depending highly on men, and most of men in books and movies do not show intense emotions of happiness or sadness (Xu et al., 2019). When annotators annotated the author's gender for tweets with unknown gender of authors, the tweets contained anger emotion will be regarded as the most confident male clues, while happy emotion as the most confident female clues (Flekova et al., 2016). These stereotypes associating emotions with genders can lead to bias. Anger and disgust are active emotions, meaning men are free to express their negative emotions. While happy and sad emotions related to women are often passive, meaning that women are dominated. The system may learn such bias when generating text. It may place women in a subordinate position to men.

Gender and Content. Here, Content refers to the specific topics we associate with a gender role. Figure 1(d) shows that activity words (M=0.057) are the most masculine while psychology words (M=-0.050) are the most feminine. Activity words have significant difference with other words ($p<0.001$). So are the psychology words ($p<0.05$). This links men to activity and women to psychology. If we regard activity as a concrete rational action and psychology as an emotional cognition, then in society, man may be a rational role and woman may be an emotional role. In study of different languages used by men and women, it is found that women prefer to use more emotional words than men (Savoy, 2018). Our society has a strong normative view that women are interested in connecting with others and promoting warmth around them. Men are generally not interested in other people and relationships. Men should focus on their goals and achievements and what they can do. As a result, women have a strong motivation to show attachment, a desire to promote the emotional feelings and downplay their personal goals and aspirations. Men, by contrast, have powerful motivations to appear strong and rational, to mask emotions, and to hide a desire to be intimate with others (Eckert and McConnell-Ginet, 2013). Such stereotypes suppress man's emotional needs and ignore woman's rational power.

6 Related Works

It was studied that word embeddings contain all kinds of biases in human society, including gender bias. These biases come from the biased data in the corpus which reflect the biased languages we use daily and from the bias of the annotators when they annotate the datasets (Van Durme, 2009). NLP algorithms may amplify the biases contained in the datasets (Sun et al., 2019). Some word embeddings of neutral words such as "nurse", "social" were proved to have closer similarities with gender words (e.g. "male", "boy", "female", and "girl") (Friedman et al., 2019; Garg et al., 2018; Brunet et al., 2019; Wevers, 2019; Santana et al., 2018; Mishra et al., 2019; Zhao et al., 2018). The latest contextualized word embeddings also have gender bias but the degree of the bias may not as much as that of traditional word embeddings (Zhao et al., 2019; Basta et al., 2019; Kurita et al., 2019; Swinger et al., 2019). In addition, multilingual embeddings contain gender bias (Lewis and Lupyan, 2020) and the bias is related to the types of different languages (Zhao et al., 2020). Word Embedding Association Test (WEAT) can be used to measure gender bias in word embeddings (Caliskan et al., 2017; Tan and Celis, 2019; Chaloner and Maldonado, 2019) and this method can also be expanded to sentence level as Sentence Encoder Association Test (SEAT) (May et al., 2019). Another method to detect and measure the gender bias in word embeddings is to analyze gender subspace in embeddings (Bolukbasi et al., 2016; Manzini et al., 2019). But this method may not show the whole gender bias in word embeddings. Some of the implicit gender bias cannot be measured and caught (Gonen and Goldberg, 2019).

7 Conclusion

In this paper, we used word embeddings to detect and measure the implicit gender bias in a language without grammatical gender. Relationships between gender and four categories in social perception and judgement are also shown according to our measurement values. Word embeddings show that we judge a woman by her appearance and perceive her as a "perfect", either happy or sad, and emotional role while we judge a man by his action and perceive him as a "bad", easily-disgusted, bad-tempered, and rational role. It may cause gender bias. This systematic bias intensifies gender differences, solidifies stereotypes about men and women, erases the uniqueness of differences between person and person, and harms those do not conform to mainstream social perception and judg-

ment and those who do not fit in the gender dichotomy. In the future, we can choose more dimensions rather than man/woman for investigation, such as in-group/inter-group, animate/inanimate, collectivism/individualism, etc.

Acknowledgements

This research project is supported by Fundamental Research Funds for the Central Universities, the Research Funds of Beijing Language and Culture University (22YCX029), and Science Foundation of Beijing Language and Culture University (supported by "the Fundamental Research Funds for the Central Universities") (20YBT07), and 2018 National Major Program of Philosophy and Social Science Fund "Analyses and Researches of Classic Texts of Classical Literature Based on Big Data Technology" (18ZDA238). We thank the reviewers for their useful feedback from both 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP2022) and ACL Rolling Review.

References

Paul Baker. 2014. *Using Corpora to Analyze Gender*. Bloomsbury.

Christine Basta, Marta R Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29:4349–4357.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *International Conference on Machine Learning*, pages 803–811. PMLR.

Judith Butler. 2002. *Gender Trouble*. Routledge.

Carmen Rosa Caldas-Coulthard and Rosamund Moon. 2010. 'curvy, hunky, kinky': Using corpora as tools for critical analysis. *Discourse & Society*, 21(2):99–133.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595.

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32.

Penelope Eckert and Sally McConnell-Ginet. 2013. *Language and Gender*. Cambridge University Press.

Paul Ekman. 1999. Basic emotions. In *Handbook of Cognition and Emotion*, pages 45–60.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 843–854.

Scott Friedman, Sonja Schmer-Galunder, Anthony Chen, and Jeffrey Rye. 2019. Relating word embedding gender biases to gender gaps: A cross-cultural analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 18–24.

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

Peter Glick and Susan T Fiske. 2001. An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American psychologist*, 56(2):109.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614.

Janet Holmes. 2013. *Women, Men and Politeness*. Routledge.

Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. Unsupervised discovery of gendered language through latent-variable modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1706–1716.

Carole M Kortenhaus and Jack Demarest. 1993. Gender role stereotyping in children's literature: An update. *Sex Roles*, 28(3):219–232.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172.

Molly Lewis and Gary Lupyan. 2020. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature human behaviour*, 4(10):1021–1028.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628.

Arul Mishra, Himanshu Mishra, and Shelly Rathee. 2019. Examining the presence of gender bias in customer reviews using word embedding. *arXiv preprint arXiv:1902.00496*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Sharyl Bender Peterson and Mary Alyce Lach. 1990. Gender stereotypes in children's books: Their prevalence and influence on cognitive and affective development. *Gender and Education*, 2(2):185–197.

Brenda Salenave Santana, Vinicius Woloszyn, and Leandro Krug Wives. 2018. Is there gender bias and stereotype in portuguese word embeddings? In *Proceedings of the 13th Edition of the International Conference on the Computational Processing of Portuguese*.

Jacques Savoy. 2018. Trump's and clinton's style and rhetoric during the 2016 presidential election. *Journal of Quantitative Linguistics*, 25(2):168–189.

Judith Stevinson Hillman. 1974. An analysis of male and female roles in two periods of children's literature. *The Journal of Educational Research*, 68(2):84–88.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640.

Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 305–311.

Magdalena Szumilas. 2010. Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent psychiatry*, 19(3):227.

Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Proceedings of Advances in Neural Information Processing Systems*, pages 13209–13220.

Deborah Tannen. 1991. *You Just Don't Understand: Women and Men in Conversation*. Virago London.

Benjamin D Van Durme. 2009. *Extracting Implicit Knowledge from Text*. University of Rochester.

Melvin Wevers. 2019. Using word embeddings to examine gender bias in dutch newspapers, 1950-1990. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 92–97.

Huimin Xu, Zhang Zhang, Lingfei Wu, and Cheng-Jun Wang. 2019. The cinderella complex: Word embeddings reveal gender stereotypes in movies and books. *PloS one*, 14(11):e0225385.

Linhong Xu, Hongfei Lin, Yu Pan, Hui Ren, and Jianmei Chen. 2008. Constructing the affective lexicon ontology. *Journal of the China Society for Scientific and Technical Information*, 27:180–185.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853.