Gender Stereotypes Differ between Male and Female Writings

Yusu Qian
Tandon School of Engineering
New York University
6 MetroTech Center
Brooklyn, NY 11201
yq729@nyu.edu

Abstract

Written language often contains gender stereotypes, typically conveyed unintentionally by the author. Existing methods used to evaluate gender stereotypes in a text compute the difference in the co-occurrence of gender-neutral words with female and male words. To study the difference in how female and male authors portray people of different genders, we quantitatively evaluate and analyze the gender stereotypes in their writings on two different datasets and from multiple aspects, including the overall gender stereotype score, the occupation-gender stereotype score, the emotion-gender stereotype score, and the ratio of male words used to female words. We show that writings by females on average have lower gender stereotype scores. We also find that emotion words in writings by males have much lower stereotype scores than the average score of all words, while in writings by females the scores are similar. We study and interpret the distributions of gender stereotype scores of individual words, and how they differ between male and female writings.

1 Introduction

Gender stereotypes in language have been receiving more and more attention from researchers across different fields. In the past, these studies have been carried out mainly by conducting surveys with humans (Williams and Best, 1977), requiring a large amount of human labor. Garg et al. (2018) quantified gender stereotypes by analyzing word embeddings trained on US Census over the past 100 years. Word embeddings capture gender stereotypes in the training data and transfer them to downstream applications (Bolukbasi et al., 2016). For example, if *programmer* appears more frequently with *he* than *she* in the training corpus, in the word embedding it will have a closer distance to *he* compared with *she*.

In this study, we analyze gender stereotypes directly from writings under different metrics. Specifically, we compare the writings by males and females to see how gender stereotypes differ between writings by the gender of authors. Our results show that writings by female authors contain much fewer gender stereotypes than writings by male authors. We recognize that there are more than two types of gender, but for the sake of simplicity, in this study we consider just female and male.

To the best of our knowledge, this study is the first quantitative analysis of how gender stereotypes differ between writings by authors of different genders. Our contributions are as follows: 1) we show that writings by females contain fewer gender stereotypes; 2) we find that over the past few decades, gender stereotypes in writings by males have decreased.

2 Related Work

Quantifying Gender Stereotypes It has been noticed that stereotypes might be implicitly introduced to image corpora and text corpora in procedures such as data collection (Misra et al., 2016; Gordon and Durme, 2013). Particularly in gender stereotypes, Garg et al. (2018) bridged social science with machine learning when they quantified gender and ethnic stereotypes in word embeddings. Park et al. (2018) measured gender stereotypes on various abusive language models, while analyzing the effect of different pre-trained word embeddings and model architectures. Zhao et al. (2018) showed the effectiveness of measuring and correcting gender stereotypes in co-reference resolution tasks.

Categorizing Text by Author Gender Shimoni et al. (2002) proposed techniques to categorize text by author gender. They selected multiple fea-

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 48–53
Florence, Italy, July 28 - August 2, 2019.

tures, for example, determiners and prepositions, and calculated their frequency means and standard errors in texts. They showed that the distributions of some of these features differ between writings by female and male. Mukherjee and Liu (2010) used POS sequence patterns to capture stylistic regularities in male and female writings. To reduce the number of features, they also proposed a selection method. They showed that author gender can be revealed by multiple features of their writings. Cheng et al. (2011) based on psycholinguistics and gender-preferential cues to build a feature space and trained machine learning models to identify author gender. They pointed out that function words, word-based features and structural features can act as gender discriminators. All these three studies achieved accuracy above 80% for identifying author gender.

3 Methodology

3.1 Dataset

In the first experiment, we use a dataset by Lahiri (2013), which consists of 3,036 English books written by 142 authors. Among these, 189 books were written by 14 female authors, others were produced by male authors.

In the second experiment, we use a dataset by Schler et al. (2006), which consists of 681,288 posts from 19,320 bloggers; approximately 35 posts and 7250 words from each blogger. The blogs are divided into 40 categories, for example, agriculture, arts and science, etc. Female bloggers and male bloggers are of equal number.

3.2 Evaluation Methods

Overall Gender Stereotypes We define the gender stereotype score of a word as:

$$b(w) = \left| \log \frac{c(w, m)}{c(w, f)} \right|,$$

where f is a set of female words, for example, *she*, *girl*, and *woman*. m is a set of male words, for example, *he*, *actor*, and *father*. $c(w, g)$ is the number of times a gender-neutral word w co-occurs with gendered words. The gendered word lists are by Zhao et al. (2018).We use a window size of 10 when calculating co-occurrence.

A word is used in a neutral way if the stereotype score is 0, which means it occurs equally frequently with male words and females word in the text. The overall stereotype score of a text, T_b, is the sum of stereotype scores of all the gender-neutral by definition words that have more than 10 co-occurrences with gendered words in the text, divided by the total count of words calculated, N.

$$T_b = \frac{1}{N} \sum_{w \in N} b(w)$$

Ratio of Male Words to Female Words To compare the frequency of male words with that of female words in a text, we calculate the ratio of male word count to female word count and denote it by R.

Occupation-Gender Stereotypes Occupation stereotypes are the most common stereotypes in studies on gender stereotypes (Lu et al., 2018). A few decades ago, females normally worked as dairy maids, housemaids and nurses, etc, while males worked as doctors, smiths, and butchers, etc. Nowadays both genders have more choices when looking for a job and for most occupations, there isnt a restriction on gender. Therefore, it is interesting to study how occupation stereotypes change over the years in female and male writings.

Occupation stereotypes score, O_b, in a text is the average stereotype score of a list of 200 gender-neutral occupations, O, in the text.

$$O_b = \frac{1}{|O|} \sum_{w \in O} b(w)$$

Emotion-Gender Stereotypes Emotion stereotypes are another kind of common gender stereotypes. In writings, especially novels, different genders are associated closely with different emotions, resulting in emotion stereotypes.

Emotion stereotypes score, E_b, in a text is the average stereotype score of a list of 200 emotion words, E, in the text.

$$E_b = \frac{1}{|E|} \sum_{w \in E} b(w)$$

Distribution of Stereotype Scores We compare the distributions of stereotype scores to analyze differences in writings by females and writings by males. We consider the following aspects of distributions: mean, variance, skewness, and kurtosis. We use S_v, S_s, and S_k to denote the average of variance, skewness, and kurtosis respectively of the distributions of stereotype scores. We plan to also add directions to individual scores by removing the absolute value function when calculating

	Gutenberg Novels							Blogs						
	T_b	R	O_b	E_b	S_v	S_s	S_k	T_b	R	O_b	E_b	S_v	S_S	S_k
f	0.54	1.14	0.70	0.62	0.25	1.58	3.48	0.56	1.46	0.72	0.56	0.21	1.74	4.81
m	1.41	3.40	1.62	1.04	0.43	0.60	0.92	0.74	2.79	0.82	0.52	0.29	1.26	2.35

Table 1: Statistics of gender stereotypes in female and male writings

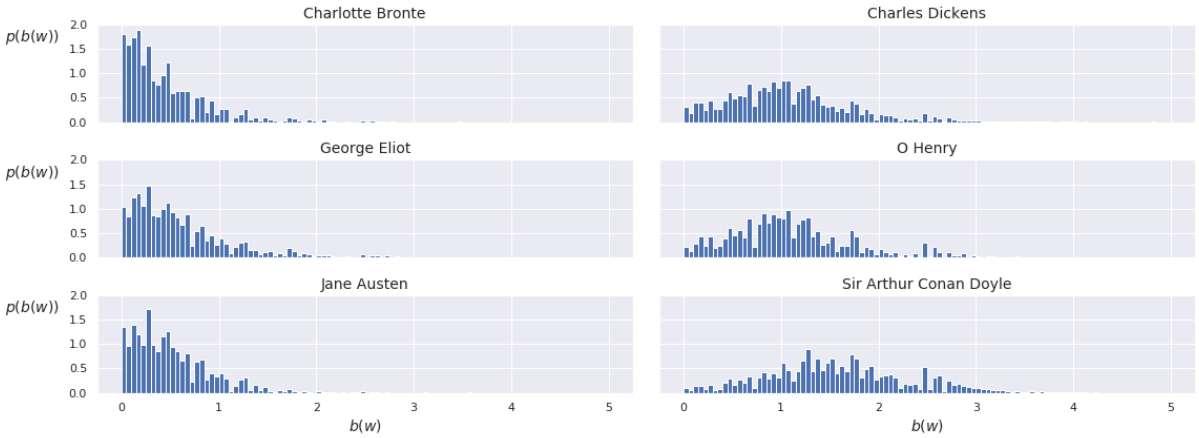

Figure 1: Distribution of stereotype scores in novels written by female(left) and male(right) authors

the scores, and analyze the distribution. We use the absolute function for most of the experiments because positive and negative values will cancel off each other when they are summed up.

Words Most Biased We alter the equation used for evaluating individual stereotype scores by removing the absolute value function, so that words occurring more with female words have negative values and words occurring more with male words have positive values. By sorting individual stereotype scores, we collect lists of words most biased towards the female gender or the male gender.

4 Results

4.1 Gender Stereotypes in Novels

We categorize 3036 books written in English and analyze the overall gender stereotypes in writings by each author. When sorted by overall stereotype scores from low to high, 12 female authors out of 14 are ranked among the top 20, or in another word, top 13.8%.

The average ratio of the total number of male words to female words in novels by female authors is close to 1, indicating that female authors mention the two genders in their novels almost equally frequently. Male authors, on the other hand, tend to write three times more frequently about their own gender.

Figure 1 shows the distribution of stereotype scores in example novels written by female and male authors. Inspection shows that the individual scores of female writings tend to cluster around score value 0 or other small values close to 0, and the percentage of words among all words calculated constantly decreases when stereotype score increases, while the individual scores of male writings tend to cluster around score values between 0.5 and 1.5, and the percentage of words among all words calculated first increases and then decreases.

Statistical analysis on the distributions confirms our observation. Table 1 shows that the average variance of stereotype scores in male writings is much larger than that of female writings, indicating that stereotype scores in female writings tend to gather near the mean while those in male writings spread out more broadly. The distribution of stereotype scores in female writings has both larger average skewness and larger kurtosis, in accordance with our observation that the distribution is skewed right with a sharp peak at a small stereotype score. In contrast, the distribution of stereotype scores in male writings has much smaller average skewness and kurtosis, in accordance with our observation that the distribution has tails on both left and right sides and has a less distinct peak.

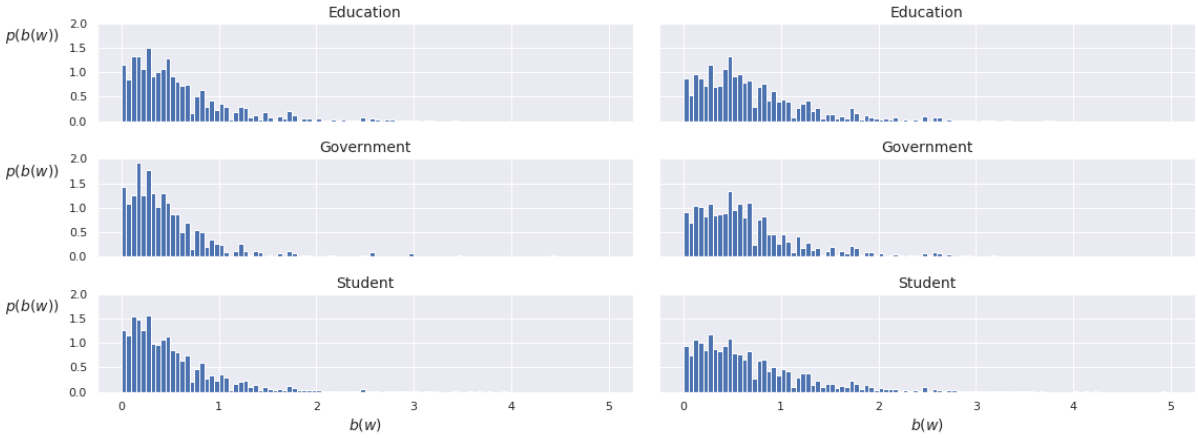

Figure 2: Distribution of stereotype scores in blogs written by female(left) and male(right) authors

Category	Author	Bias Direction	Top 20 Words in the Most Biased Wordlist
novel	male	male	judge, us, speech, friends, much, ask, created, made, never, life, framed, yet, knows, also, like, declared, each, great, believe, political
novel	male	female	necessarily, married, constitution, struck, need, short, votes, before, want, consent, taught, due, but, portion, course, alone, bread, engage, equal, five
novel	female	male	pocket, russian, hands, few, probably, said, round, that, admitted, out, way, caught, read, sure, stared, coming, gravely, began, followed, face
novel	female	female	suppose, bed, set, new, suddenly, door, right, morning, meant, remembered, given, well, up, lay, possible, realized, smiled, kind, lips, eyes
blog	male	male	sure, over, three, saw, got, if, now, did, things, as, two, before, really, this, gets, our, back, being, left, feels
blog	male	female	bring, and, issue, friends, so, said, what, wet, take, telling, wanted, call, going, much, me, always, something, same, little, met
blog	female	male	mail, does, stories, report, lucky, online, beat, imagine, surprised, reply, tonight, reporting, cut, blue, radio, reports, jeans, story, thank, forget
blog	female	female	talk, body, baby, age, death, won, pain, weight, together, later, beautiful, ears, walk, head, large, sees, sexy, dress, passed, family

Table 2: A sample of most biased words in female and male writings from experiments on novels and blogs

4.2 Gender Stereotypes in Blogs

After analyzing blogs on 40 categories written by equal numbers of male and female bloggers, we find out that for 35 categories, writings by males contains more gender stereotypes by 41.39% on average. Only in 5 categories including accounting, agriculture, biotech, construction and military, writings by female contains more gender stereotypes than male writings by 16.29% on average.

The average ratio of the total number of male words to female words in blogs by female authors is around 1.5, while in blogs by male authors, the ratio is around 2.8. Similar to the findings in the first experiment, male authors write more about the male gender.

Figure 2 shows a similar pattern in blogs with the pattern in novels. Individual stereotype scores also cluster closer around 0 or a relatively small value in female writings, while those of male writings cluster around a larger value. This pattern, however, is weaker than that found in experiments on novels. Both Figure 2 and statistics in Table 1 show that difference in blogs written by female and male authors in terms of gender stereotypes is smaller than the difference in novels. It is also worth mentioning that while T_b of blogs written by females is almost the same as that of novels written by females, T_b of blogs written by males is much lower than that of novels written by males. The trend in the ratio of male word count to female word count is similar. One possible interpretation of this is that while the blogs were written in 2004, the novels in the Gutenberg subsample were written decades ago, when the society had more constraints on female and gender equality was not

paid as much attention to as it is today.

4.3 Gender Stereotypes Categories

For both two datasets, O_b is larger than T_b, indicating that occupation words in both female and male writings contain more gender stereotypes than most other words. E_b is almost the same as T_b in female writings, while it is much lower than T_b in male writings, indicating that gender stereotypes in emotion words are not the main contributors to the overall gender stereotypes in male writings.

5 Conclusion and Discussion

In this study, we perform experiments on two datasets to analyze how gender stereotypes differ between male and female writings. From our preliminary results we observe that writings by female authors contain fewer gender stereotypes than writings by male authors. This difference appears to have narrowed over time, mainly by the reduction of gender stereotypes in writings by male authors. We plan to: 1)further analyze the typical types of gender stereotypes in writings by authors of different genders and how they resemble with or differ from each other, by studying the most biased words and the average stereotype scores of different categories of words, for example, verbs, adjectives, etc.; 2) perform experiments on more writings from the past century to inspect more closely if there exists a trend in the transformation of gender stereotypes; 3) existing stereotype evaluation methods evaluate every word not in the excluded word lists, in our case, the male and female word lists. Some frequently used words, such as *the*, *one*, and *an*, are not considered to be able to contain stereotypes, unlike words such as *strong*, *doctor*, and *jealous*, which are more closely associated with one gender in writings. We plan to seek a way to filter gender-neutral words and only keep those capable of carrying stereotypes for stereotype quantification.

6 Acknowledgments

Many thanks to Gang Qian, Cuiying Yang, Pedro L. Rodriguez, Hanchen Wang, and Tian Liu for helpful discussion, and reviewers for detailed comments and advice.

References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS'16 Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4356–4364.

Na Cheng, Rajarathnam Chandramouli, and K. P. Subbalakshmi. 2011. Author gender identification from text. *Digital Investigation*, 8(1):78–88.

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.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction, AKBC@CIKM 13, San Francisco, California, USA, October 27-28, 2013*, pages 25–30.

Shibamouli Lahiri. 2013. Complexity of word collocation networks: A preliminary structural analysis. *CoRR*, abs/1310.5111.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing. ArXiv:1807.11714v1.

Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross B. Girshick. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2930–2939.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 207–217, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2799–2804.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. 2006. Effects of age and gender on blogging. In *Computational Approaches to Analyzing Weblogs - Papers from the AAAI Spring Symposium, Technical Report*, volume SS-06-03, pages 191–197.

Anat Rachel Shimoni, Moshe Koppel, and Shlomo Argamon. 2002. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4):401–412.

John E. Williams and Deborah L. Best. 1977. Sex stereotypes and trait favorability on the adjective check list. *Educational and Psychological Measurement*, 37(1):101–110.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Chang Kaiwei. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 48474853. Association for Computational Linguistics.