Exploring Gender Biases in Information Retrieval Relevance Judgement Datasets

Amin Bigdeli, Negar Arabzadeh, Morteza Zihayat, and Ebrahim Bagheri

Ryerson University, Toronto, Canada
{abigdeli,narabzad,mzihayat,bagheri}@ryerson.ca

Abstract. Recent studies in information retrieval have shown that gender biases have found their way into representational and algorithmic aspects of computational models. In this paper, we focus specifically on gender biases in information retrieval gold standard datasets, often referred to as relevance judgements. While not explored in the past, we submit that it is important to understand and measure the extent to which gender biases may be presented in information retrieval relevance judgements primarily because relevance judgements are not only the primary source for evaluating IR techniques but are also widely used for training end-to-end neural ranking methods. As such, the presence of bias in relevance judgements would immediately find its way into how retrieval methods operate in practice. Based on a fine-tuned `BERT` model, we show how queries can be labelled for gender at scale based on which we label MS MARCO queries. We then show how different psychological characteristics are exhibited within documents associated with gendered queries within the relevance judgement datasets. Our observations show that stereotypical biases are prevalent in relevance judgement documents.

1 Introduction

Extensive research in the psychology and sociology literature has shown that gender stereotypes can affect an individual's life descriptively and prescriptively [1, 2]. These gender stereotypes not only affect the expectations of women and men about their behaviour, qualities, priorities, and personal needs implicitly, but can also influence the way they process information [3, 4]. Besides, gender stereotypes can influence an individual's judgements, leading to unfair treatments and outcomes [5]. While individuals' perception of gender differences might be aligned with reality in certain cases, such perceptions often originate from gender stereotypes [6]. Recently, the impact of various biases has been a topic of interest among researchers in a variety of domains, including Information Retrieval (IR). [7–13]. For instance, given the wide adoption of neural embeddings in IR, various researchers have already begun investigating the impact of implicit biases that are embedded in neural embeddings. In [9], Bolukbasi et al. highlighted the fact that sexism implicit within pre-trained neural embeddings has the potential to pose the risk of introducing different types of biases in practically deployed applications; hence, reflecting gender stereotypes in real time.

Given the impact of biases when seeking information, Rekabsaz et al. [7] have examined the degree of gender bias among several neural retrieval methods. They found that the utilization of already biased pre-trained embeddings considerably amplifies gender biases among the retrieved documents. In another important study [12], Fabris et al. proposed a word genderedness measure to detect and quantify how various types of information retrieval methods respond to gendered queries by retrieving documents that are inclined towards similar gender stereotypes. As a result of their experiments, the authors found that lexical, semantic, and neural models reinforce gender stereotypes in their results.

While biases among different retrieval methods and neural embeddings have been generally studied in IR, to the best of our knowledge, potential biases within gold standard benchmark datasets (often known as relevance judgements, aka *qrels*) have not yet been explored. We believe that it is important to study whether biases may have been introduced in gold standard datasets, which in essence govern how retrieval methods are trained and evaluated. An inclination towards a specific gender or the ascription of implicit biases towards them can result in a biased retrieval method. As such, the objective of this paper is to study potential stereotypical gender biases in information retrieval relevance judgements. The other distinguishing aspect of our work is that unlike earlier work [7, 12], we do not propose a certain computational metric for measuring gender biases, but rather we measure various psychological characteristics of document content associated with gendered queries. This way, we quantify, if and when, systematic differences are exhibited between queries of different genders.

In summary, our work distinguishes itself from the literature by (1) offering an accurate and well-validated query gender classifier that can be used to label queries based on gender at scale; (2) studying potential biases at the level of gold standard relevance judgements through widely adopted psychological characteristics; and (3) revealing systematic biases aligned with perceptual stereotypes within query relevance judgements.

2 Methodology

We follow a three-staged methodological process in this paper: **(1)** In order to be able to determine query gender at scale, we benefit from the dataset of gendered queries provided by [7] to train a contextualized classifier to predict query gender. Subsequently, the trained model is used to label MS MARCO queries [14] (*c.f.* Section 2.2). **(2)** Based on these classified gendered queries from MS MARCO, we identify the associated relevant documents for each query and quantitatively measure various psychological characteristics of each such document using the well-established Linguistic Inquiry and Word Count (LIWC) toolkit (*c.f.* Section 2.3). **(3)** We report on gender stereotypical biases in information retrieval gold standards (query relevance judgements, i.e., qrels), which align with well-documented perceived biases in the psychological literature (*c.f.* Section 3). We note that all of our data, code, and results are made publicly accessible[1].

2.1 Datasets

Dataset for query gender identification. We employed the publicly available gender-annotated dataset released by [7] that consists of queries labeled by one of

[1] `https://github.com/aminbigdeli/gender-bias-in-relevance-judgements`

Table 1. The accuracy and F1 score of each classifier by gender.

Category	Classifier	Accuracy	F1-Score		
			Female	Male	Neutral
Dynamic Embeddings	BERT (base uncased)	**0.856**	**0.816**	**0.872**	**0.862**
	DistilBERT (base uncased)	0.847	0.815	0.861	0.853
	RoBERTa	0.810	0.733	0.820	0.836
	DistilBERT (base cased)	0.800	0.730	0.823	0.833
	BERT (base cased)	0.797	0.710	0.805	0.827
	XLNet (base cased)	0.795	0.710	0.805	0.826
Static Embeddings	Word2Vec	0.757	0.626	0.756	0.809
	fastText	0.750	0.615	0.759	0.792

the following classes: 1) non-gendered (neutral), 2) female, 3) male, and 4) other or multiple genders. The dataset consists of 742 female, 1,202 male and 1,765 neutral queries. We removed the 41 queries related to the 'Other or Multiple Genders' class as there were not sufficient instances to train a classifier. We also benefited from 32 pairs of gendered terms released by the same authors.

Dataset for measuring bias. For the purpose of measuring bias in relevance judgements, we adopted the queries in MS MARCO Dev set [14] that had at least one related human-judged relevance judgement document – equivalent to 51,827 queries. Note that, the queries from [7] were removed from this dataset to avoid unintended leakage.

2.2 Query Gender Identification and Labeling

As the first step and in order to be able to label gendered queries at scale, we employ the dataset released by [7] to train relevant classifiers. We adopt two recent yet widely adopted techniques for this purpose, namely *dynamic embeddings* and *static embeddings*. More specifically, dynamic embeddings include models such as BERT [15], DistilBERT [16], RoBERTa [17], and XLNet [18], which are pre-trained models that have been trained on large corpora. For our work, we used the sequence classification class of BERT, DistilBERT, RoBERTa, and XLNet that have a linear layer over the pooled output, which is used to compute class likelihood scores. We used this fine tuning capability of these models with a batch size of 16 and the Adamw optimizer with learning rate $2e - 5$. We set the number of epochs to 10 for BERT, DistilBERT, RoBERTa and 20 for XLNet.

Unlike dynamic embeddings, static embeddings such as fastText [19] and Word2Vec [20] create a single vector representation per token without regard for context. In order to train a fastText model, we used pre-trained vectors based on the Common Crawl dataset [2] and fine-tuned them based on the pair of gendered terms and the queries from [7]. As another model, we employed the pretrained Google News Word2Vec model and adopted the average of each query's term vectors to represent the query. Based on this, an SVM classifier with a polynomial kernel function was applied to classify the queries.

In order to evaluate the performance of the classifiers, we adopt a 5-fold cross-validation strategy. As shown in Table 1, the uncased fine-tuned BERT model shows the best performance for query gender identification. Now, using the uncased fine-tuned BERT model, we labeled all of the 51,827 queries in the

[2] https://bit.ly/3oBFTJ0

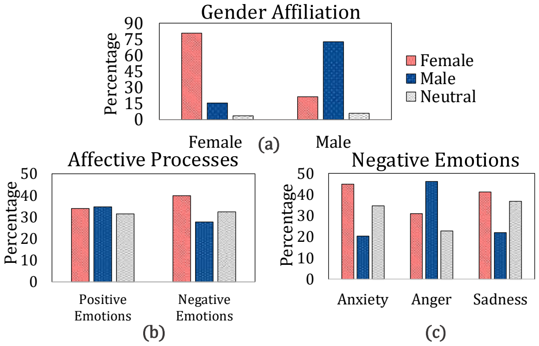

Fig. 1. (a) Percentage of female and male affiliations in relevant documents for each of the female, male and neutral query groups.(b) and (c) Differences between affective processes of relevance judgements for different gendered queries. The y-axis shows the percentage of each characteristic across female, male and neutral query sets.

MS MARCO query set. In total, we ended up with 48,200 neutral queries, 2,222 male queries, and 1,405 female queries. To have a balanced setup, we retained all 1,405 female queries and randomly selected 1,405 queries from each of the other two classes. We utilized 1,405 queries in each class and their associated relevance judgement documents to investigate the presence of stereotypical gender biases.

2.3 Quantifying Psychological Characteristics

Our approach for quantifying bias is based on measuring different psychological characteristics of the relevance judgement documents associated with each query. We expect the measures of psychological characteristics across genders to align with findings from well-founded psychological experiments and not to exhibit behavior consistent with stereotypical biases associated with gender. To investigate this, we employ Linguistic Inquiry and Word Count (LIWC) [21] text analytics toolkit to compute the degree to which different psychological characteristics are observed in relevance judgement documents. We consider stereotypical biases relating to affective processes, cognitive processes, drive, and personal concerns.

Before we present our findings, we benefit from LIWC to validate the performance of our BERT-based gender classifier. LIWC can be used to measure the male or female affiliation of a document. We measure such gender affiliations through LIWC for all relevance judgement documents and report the percentage of gender affiliations related to each query gender type in Figure 1(a). This figure asserts the efficiency of the BERT-based gender classifier as it shows that female queries are primarily associated with female affiliated documents, while male queries are related to male affiliated documents. Furthermore, neutral queries do not show affiliation with either gender. We consider the consistent behavior between LIWC and the gender classifier as a sign of the utility of the gender classifier as well as appropriateness of LIWC to be applied to such documents.

3 Findings

During information processing, individuals might make observations that are compatible with their stereotypical mental presumptions [3]. We are interested

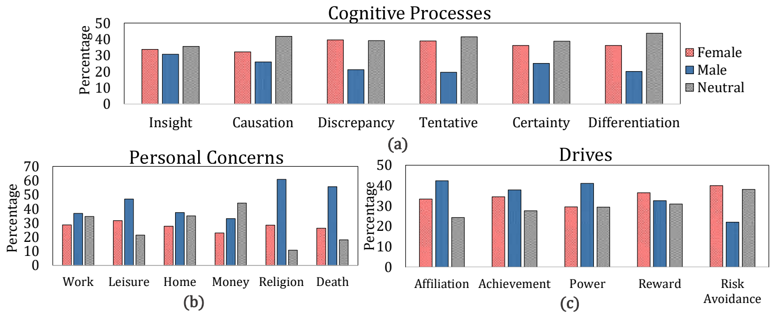

Fig. 2. Differences between the three psychological processes of relevance judgements for different gendered queries. The y-axis is similar to Figure 1.

in exploring if such stereotypical biases are incorporated into gold standard relevance judgments, as they might pass biases onto retrieval methods.

Affective Processes are defined as the expression of positive and negative emotions by an individual. We visualize the degree of positive and negative emotions expressed in relevance judgement documents associated with gendered queries in Figure 1(b). As shown, the documents present a similar degree of positive emotions regardless of the gender type of the query they are associated with. However, when considering negative emotions, documents that are related to female queries exhibit a higher degree of negativity compared to male and neutral queries. To further understand this, we explore the three sub-characteristics of negative emotions as described in [21], namely *anxiety*, *anger*, and *sadness*. We find that relevance documents associated with male queries exhibit higher rates of anger whereas higher degrees of anxiety and sadness are observed for documents associated with female queries. This implies that stereotypical biases can be observed in gold standard relevance judgements with regards to negative emotions of affective processes. Psychological studies [22, 23] have already shown that there are no systematic differences between males and females as it relates to affective processes such as the experience or expression of anger. Gao et al [24] also reported that there were no significant gender differences in average depression and stress levels among female and male students. While the study did find significant gender differences in stress problems, it did report higher levels of anxiety for females consistent with the observations in Figure 1(c).

Cognitive processes are the higher-level functions of the brain and are represented through characteristics such as insight, causation, discrepancy, tentativeness, certainty, and differentiation within the LIWC toolkit. As shown in Figure 2(a), documents associated with female queries show superior degrees of cognitive capacity compared to those related to male queries. However, based on psychological literature, males and females share similar cognitive abilities on most of cognitive functions [25]. Various researchers have argued that potentially observable differences between the sexes relating to intellectual and cognitive functions can be attributed to patterns of abilities as opposed to overall intellectual function of each gender [26–28]. Our observations show that there are implicit biases encoded within the relevance judgment documents associated with gendered queries in terms of psychological expression of cognitive processes.

Personal Concerns such as work, leisure, home, money, religion, and death are investigated and the findings presented in Figure 2(b), reveal that relevance judgement documents associated with male queries have a higher degree of focus on personal concerns compared to female queries. This finding is aligned with the literature when it comes to personal concerns for leisure. The literature [29] reports that distribution of leisure time is significantly impacted by gender, especially for time allocated over the weekend. However, social psychology research has shown that such differences do not exist in other aspects of personal concern such as death, anxiety, and religiosity [30–33]. Furthermore, the literature reports that although the number of females has increased in the workplace and their presence in traditionally male-dominated professions has grown, there are still descriptive gender stereotypes in environments [34–38]. In another study [4], Heilam discussed that prescriptive and descriptive gender stereotypes result in gender bias in the workplace, which are unfounded. We find that such stereotypical biases do exist in relevance judgement documents and reflect biases that have been reported in the literature in the past.

Drives focus on characteristics of individuals that guide them towards achieving goals or accomplishing milestones. They can be defined with five key characteristics including affiliation, achievement, power, reward, and risk avoidance [21]. We find, as shown in Figure 2(c), that relevance judgement documents associated with male queries express higher degrees of affiliation, achievement, and power compared to female queries, while the inverse is observed for reward and risk avoidance. These findings are supported by the literature that males seek for power and achievement more than females, but contradict studies that report higher degrees of affiliation for the female gender [39]. In addition, Byrens et al. have shown that males are more likely to take greater risks compared to females [40], which is compatible with our observations on degrees of risk avoidance.

Similar to other psychological characteristics, we find that differences can be observed regarding different personal drive characteristics between relevance documents associated with female and male queries. However, in this case, the differences are not due to stereotypical differences and have already been shown in the related literature that such personal drive characteristics are observed in practice for reasons such as physiological differences in gender.

4 Concluding Remarks

This paper investigated gender biases in gold standard IR relevance judgement datasets. We found that gender biases are prevalent in relevance judgements across a range of psychological processes. While some of the biases are expected as a result of physiological differences between genders, most gender biases are a result of the stereotypical perception of gender differences. We submit that regardless of the source of gender bias, be it stereotypical or physiological, IR relevance judgement documents should not show significant differences across various psychological processes based on the gender of the submitted query. Unbiased gold standards will ensure that gender biases do not get translated into representation and algorithmic aspects of retrieval methods.

References

1. Burgess, Diana, and Eugene Borgida. “Who women are, who women should be: Descriptive and prescriptive gender stereotyping in sex discrimination.” Psychology, public policy, and law 5, no. 3 (1999): 665.
2. Heilman, Madeline E. “Description and prescription: How gender stereotypes prevent women’s ascent up the organizational ladder.” Journal of social issues 57, no. 4 (2001): 657-674.
3. Ellemers, Naomi. “Gender stereotypes.” Annual review of psychology 69 (2018): 275-298.
4. Heilman, Madeline E. “Gender stereotypes and workplace bias.” Research in organizational Behavior 32 (2012): 113-135.
5. Swim, Janet, Eugene Borgida, Geoffrey Maruyama, and David G. Myers. “Joan McKay versus John McKay: Do gender stereotypes bias evaluations?.” Psychological Bulletin 105, no. 3 (1989): 409.
6. Huddy, Leonie, and Nayda Terkildsen. “Gender stereotypes and the perception of male and female candidates.” American journal of political science (1993): 119-147.
7. Rekabsaz, Navid, and Markus Schedl. “Do Neural Ranking Models Intensify Gender Bias?.” arXiv preprint arXiv:2005.00372 (2020).
8. Sun, Tony, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. “Mitigating gender bias in natural language processing: Literature review.” arXiv preprint arXiv:1906.08976 (2019).
9. Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. “Man is to computer programmer as woman is to homemaker? debiasing word embeddings.” In Advances in neural information processing systems, pp. 4349-4357. 2016.
10. Zhao, Jieyu, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. “Learning gender-neutral word embeddings.” arXiv preprint arXiv:1809.01496 (2018).
11. Rekabsaz, Navid, James Henderson, Robert West, and Allan Hanbury. “Measuring Societal Biases in Text Corpora via First-Order Co-occurrence.” arXiv preprint arXiv:1812.10424 (2018).
12. Fabris, Alessandro, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. “Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms.” Information Processing Management (2020): 102377.
13. Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. “Semantics derived automatically from language corpora contain human-like biases.” Science 356, no. 6334 (2017): 183-186.
14. Nguyen, Tri, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. “Ms marco: A human-generated machine reading comprehension dataset.” (2016).
15. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805 (2018).
16. Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.” arXiv preprint arXiv:1910.01108 (2019).
17. Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. “Roberta: A robustly optimized bert pretraining approach.” arXiv preprint arXiv:1907.11692 (2019).

18. Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. “Xlnet: Generalized autoregressive pretraining for language understanding.” In Advances in neural information processing systems, pp. 5753-5763. 2019.
19. Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. “Enriching word vectors with subword information.” Transactions of the Association for Computational Linguistics 5 (2017): 135-146.
20. Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. “Distributed representations of words and phrases and their compositionality.” In Advances in neural information processing systems, pp. 3111-3119. 2013.
21. Pennebaker, James W., Martha E. Francis, and Roger J. Booth. “Linguistic inquiry and word count: LIWC 2001.” Mahway: Lawrence Erlbaum Associates 71, no. 2001 (2001): 2001.
22. Milovchevich, Darryl, Kevin Howells, Neil Drew, and Andrew Day. “Sex and gender role differences in anger: An Australian community study.” Personality and Individual differences 31, no. 2 (2001): 117-127.
23. Deffenbacher, Jerry L., Eugene R. Oetting, Gregory A. Thwaites, Rebekah S. Lynch, Deborah A. Baker, Robert S. Stark, Stacy Thacker, and Lora Eiswerth-Cox. “State–trait anger theory and the utility of the trait anger scale.” Journal of Counseling Psychology 43, no. 2 (1996): 131.
24. Gao, Wenjuan, Siqing Ping, and Xinqiao Liu. “Gender differences in depression, anxiety, and stress among college students: a longitudinal study from China.” Journal of affective disorders 263 (2020): 292-300.
25. Hyde, Janet S. “Sex and cognition: gender and cognitive functions.” Current opinion in neurobiology 38 (2016): 53-56.
26. Halpern, D. F. 2012. “Sex Differences in Cognitive Abilities.” 4th ed. New York, NY: Psychology Press.
27. Collins, David W., and Doreen Kimura. “A large sex difference on a two-dimensional mental rotation task.” Behavioral neuroscience 111, no. 4 (1997): 845.
28. Mollet, Gina A. “Fundamentals of human neuropsychology.” Journal of Undergraduate Neuroscience Education 6, no. 2 (2008): R3.
29. Shaw, Susan M. “Gender and leisure: Inequality in the distribution of leisure time.” Journal of Leisure research 17, no. 4 (1985): 266-282.
30. Dickstein, Louis S. “Attitudes toward death, anxiety, and social desirability.” OMEGA-Journal of Death and Dying 8, no. 4 (1978): 369-378.
31. McDonald, Rita T., and William A. Hilgendorf. “Death imagery and death anxiety.” Journal of clinical psychology 42, no. 1 (1986): 87-91.
32. Francis, Leslie J. “The personality characteristics of Anglican ordinands: feminine men and masculine women?.” Personality and individual differences 12, no. 11 (1991): 1133-1140.
33. Deconchy, Jean-Pierre. “Boys and girls choices for a religious group.” Psychology and religion. Harmondsworth: Penguin (1973): 284-300.
34. Schein, Virginia E. “A global look at psychological barriers to women’s progress in management.” Journal of Social issues 57, no. 4 (2001): 675-688.
35. Heilman, Madeline E., Caryn J. Block, and Richard F. Martell. “Sex stereotypes: Do they influence perceptions of managers?.” Journal of Social behavior and Personality 10, no. 4 (1995): 237.
36. Heilman, Madeline E., Caryn J. Block, Richard F. Martell, and Michael C. Simon. “Has anything changed? Current characterizations of men, women, and managers.” Journal of applied psychology 74, no. 6 (1989): 935.

37. Brenner, O. C., Joseph Tomkiewicz, and Virginia Ellen Schein. “The relationship between sex role stereotypes and requisite management characteristics revisited.” Academy of management journal 32, no. 3 (1989): 662-669.
38. Dodge, Katherine A., Faith D. Gilroy, and L. Mickey Fenzel. “Requisite management characteristics revisited: Two decades later.” Journal of Social Behavior and Personality 10, no. 4 (1995): 253.
39. Denzinger, Ferdinand, Sabine Backes, Veronika Job, and Veronika Brandstätter. “Age and gender differences in implicit motives.” Journal of Research in Personality 65 (2016): 52-61.
40. Byrnes, James P., David C. Miller, and William D. Schafer. “Gender differences in risk taking: a meta-analysis.” Psychological bulletin 125, no. 3 (1999): 367.