Understanding and Mitigating Gender Bias in Information Retrieval Systems

Amin Bigdeli[1(✉)], Negar Arabzadeh[2], Shirin Seyedsalehi[1], Morteza Zihayat[1], and Ebrahim Bagheri[1]

[1] Toronto Metropolitan University, Toronto, Canada
{abigdeli,shirin.seyedsalehi,bagheri}@torontomu.ca
[2] University of Waterloo, Waterloo, Canada
narabzad@uwaterloo.ca

Abstract. Recent studies have shown that information retrieval systems may exhibit stereotypical gender biases in outcomes which may lead to discrimination against minority groups, such as different genders, and impact users' decision making and judgements. In this tutorial, we inform the audience of studies that have systematically reported the presence of stereotypical gender biases in Information Retrieval (IR) systems and different pre-trained Natural Language Processing (NLP) models. We further classify existing work on gender biases in IR systems and NLP models as being related to (1) relevance judgement datasets, (2) structure of retrieval methods, (3) representations learnt for queries and documents, (4) and pre-trained embedding models. Based on the aforementioned categories, we present a host of methods from the literature that can be leveraged to measure, control, or mitigate the existence of stereotypical biases within IR systems and different NLP models that are used for down-stream tasks. Besides, we introduce available datasets and collections that are widely used for studying the existence of gender biases in IR systems and NLP models, the evaluation metrics that can be used for measuring the level of bias and utility of the models, and de-biasing methods that can be leveraged to mitigate gender biases within those models.

1 Motivation and Overview

There have been both qualitative and quantitative studies that have effectively shown that societal biases have become prevalent in various Natural Language Processing (NLP) and Information Retrieval (IR) techniques, models, and datasets [2,3,10,11,13,15,17,18,24,28,35,37]. Given these tools are often deployed at scale, such biases have the potential to directly impact the lives of many people. More specifically within the context of IR systems, biased methods can exacerbate biases by exposing users to a set of biased documents in response to user queries. Such biases can have a potentially harmful impact on the users' judgments when exposed to unfair and biased search results. This is concerning since not only do a large number of search engine users heavily rely on retrieval systems on a daily basis but also the search results often constitute a major component of important practical systems such as recommendation systems, question answering systems, intelligent assistants, to name a few.

J. Kamps et al. (Eds.): ECIR 2023, LNCS 13982, pp. 315–323, 2023.
https://doi.org/10.1007/978-3-031-28241-6_32

Researchers such as Draws et al. [14] have most recently shown that when search results are biased, the users who are exposed to the biases results will tend to favor the biased viewpoint. This aligns very well with several forms of cognitive bias identified by Azzopardi [1] including *Availability bias*, which points to user biases towards content that are more easily accessible, and *Anchoring Bias* that reports that users are more likely to focus on the first piece of information that they receive. Thus, it is important to systematically control the degree of biases that are exhibited by such retrieval systems to avoid their detrimental effects on the users' beliefs and decisions. To systematically address such biases, various researchers have proposed methods that can help control and/or mitigate biases, such as gender biases, in information retrieval systems. In this tutorial, we will provide a classification of existing work in the literature [2,5,7–10,12,16,19–21,23,30,32,33,35,38,39] and introduce the state-of-the-art methods that are available for managing gender biases within IR systems. The structure of this tutorial can be summarized as follows:

1. The tutorial will present concrete evidence, using real-world examples of cases where gender biases are introduced and intensified in natural language processing and IR systems;
2. The tutorial will draw inspiration from and provide adequate contextual information from experience reports and methodological work in natural language processing that have already explored gender biases [34,35];
3. A systematic classification of possible sources for gender biases will be presented and details of how biases can be transferred from these sources will be provided. These sources include relevance judgement collections, ranker characteristics, objective functions, and neural embeddings, to name a few.
4. The tutorial will review existing methods that have attempted to control or mitigate gender biases and will also provide an in-depth treatment of the retrieval effectiveness-bias tradeoff. This tradeoff is concerned with the right balance between maximizing retrieval effectiveness and minimizing gender bias, which may not always be synergistic;
5. A clear description of evaluation methodology, datasets, and metrics that have been used in the literature for investigating gender bias will be provided.

Our tutorial will build on our recent tutorial at SIGIR 2022 [6] (https://bit.ly/3TfDMss) as well as four invited talks that we have delivered at Microsoft Research, Center for Intelligent Information Retrieval at UMass Amherst, the keynote talk at the BIAS workshop at ECIR 2022, and Radboud University. The central focus of the tutorial and these talks have been on methods for controlling and mitigating gender biases, which can broadly be classified as follows:

1. *Relevance Judgement Collections*: Relevance judgment documents are often considered as gold standard benchmark datasets used for training and evaluating ranking models. Researchers have already introduced methodological processes for studying possible traces of gender bias in relevance judgment collections [9,27], and show that stereotypical gender biases can be observable in these collections, which are capable of making their way into the algorithmic aspects of ranking models that are trained and evaluated based on them.

We will also introduce those approaches that have been introduced in the literature for de-biasing relevance judgment collections. We will report on the findings from these studies that when neural ranking models are trained based on de-biased relevance judgments, the level of gender biases may be reduced while retrieval effectiveness is maintained.

2. *Neural Representations*: Neural embeddings have been widely adopted in IR systems for different tasks such as document retrieval [4]. Since neural representations have often been pre-trained on large corpora, they may have picked up existing gender stereotypes and biases. Many research works [5,10,12,35,39] have investigated gender biases within these neural representations, and have proposed methods for mitigating the levels of bias using different approaches such as data augmentation and embeddings de-biasing techniques [10,17,25,29]. We will cover how such techniques can be adopted in practice to manage gender biases.
3. *Query Representation*: The query submitted by the user can itself be highly influential on the retrieved list of documents. For instance, Kulshrestha et al. [22] examine the impact of such biases in the context of political search queries. Therefore, we will report on studies that explore the gendered nature of search queries [36], as well as those that present query reformulation mechanisms that attempt to revise an initial query in a way that will lead to a less biased list of documents while maintaining retrieval effectiveness [8].
4. *Retrieval Methods*: Recent studies show that neural-based retrieval methods can intensify the level of gender biases within the retrieved list of documents [16,31]. Therefore, it is important to manage the level of gender bias at the ranker level. Researchers have already looked into how rankers can be made less biased (or in other terms more fair) through approaches such as introducing bias-aware loss functions or bias-aware negative sampling strategies. In the tutorial, we will cover various existing work in this space. For instance, we will introduce methods such as AdvBert [30] that leverages adversarial components within the `BERT` reranker loss function for decreasing the level of bias in neural ranking models. We will also introduce the bias-aware neural ranker [32], which explicitly incorporates a notion of gender bias and hence control how bias is expressed in documents that are retrieved. We will also cover bias-aware negative sampling strategies that consider the degrees of gender bias when sampling documents to be used for training neural rankers [7].

We highlight that this tutorial will build on but significantly expand the scope of our talks by providing comprehensive information about evaluation strategies, available datasets, and bias measurement techniques. Most important, we will discuss the limitations of existing work from both technical as well as conceptual perspectives. For instance, we will highlight the following two limitations: (1) existing work in the literature have focused primarily on the notion of sex as a binary construct and assumed that search queries and results can be analyzed from their association with the male or female gender. This is a major limitation that needs to be addressed in future work; and (2) Most existing work assume that gender bias can be measured based on the frequency of gendered terms. This overlooks the complexity associated with the stance and position of documents with regards to different gender identities in favor of simplifying computation.

2 Objectives

The objectives of this tutorial can be enumerated as follows: 1) Show the presence of gender biases in IR systems and large scale corpora relevance judgments; 2) Introduce bias measurement metrics used for calculating the level of gender biases within the retrieved list of documents; 3) Present datasets used for investigating gender biases in IR results; 4) Introduce de-biasing methods for reducing the level of bias in relevance judgment datasets; 5) Describe existing methods for mitigating the level of bias within neural ranking models; 6) Present existing methods for the exploration and mitigation of bias in neural embeddings; 7) Highlight important theoretical and conceptual limitations of existing work when dealing with the concept of gender.

The aforementioned topics will give participants a thorough understanding of existing datasets and bias measurement metrics used for investigating gender biases within information retrieval results. Besides, they become familiar with methods used for reducing gender biases within IR systems. As a result, they can take advantage of these techniques to release models that are aware of gender biases and expose users to a less biased list of documents without being worried about the retrieval effectiveness of their model. In addition, these topics can be beneficial for researchers who are conducting research in a similar area in terms of applying introduced de-biasing methods for other *types of societal biases* and can serve as a useful starting point.

3 Format and Schedule

The length of this tutorials will be half day, i.e., 3 h plus breaks and will be delivered in-person by the presenters. This tutorial covers the following sections:

Introduction to the Topic of Gender Biases in IR [30 min]. The tutorial will begin by covering the foundations of IR methods as well as the datasets, which will be referred to throughout the tutorial. We will provide evidence to show the footprints of various forms of gender bias in IR systems and will introduce bias measurement methods that will be used for measuring the level of bias in retrieval results.

Exploration of Gender Biases in IR Relevance Judgments and Retrieval Methods [30 min]. We discuss the presence of stereotypical gender biases within various IR methods and compare the level of gender bias among their retrieved results. Subsequently, we explore the possibility of gender biases within relevance judgement datasets, also known as gold standard datasets. Through a methodological approach, we discuss that such datasets could be a potential source of bias.

Mitigation of Gender Biases in IR Methods [60 min]. In this session, we review existing methods for reducing gender bias through different classes of retrieval methods, namely, term-frequency-based methods as well as neural ranking models. These de-biasing methods can be classified based on four different strategies, namely (1) Adversarial Training, (2) Regularizing the Loss Function,

(3) Data Augmentation, and (4) Query Representation. Additionally, we show the effectiveness of each of the proposed methods in reducing the level of bias within the retrieved results and their utility. We will demonstrate that leveraging these methods allows for maintaining utility and at the same time mitigating the level of bias. Finally, we demonstrate how each of the proposed methods can be applied for other societal attributes other than gender.

Exploration and Mitigation of Gender Biases in Neural Embeddings [40 min]. In this session, we will explore the existence of gender biases within the representation and algorithmic aspects of different classes of neural embeddings, namely (1) static word embeddings and (2) dynamic word embeddings. In addition, we will cover the proposed methods used for de-basing neural embeddings and will show their impact on both reducing gender biases and performance of down-stream tasks.

Limitations and Future Work [20 min]. This session will discuss major theoretical and conceptual limitations of existing work and will present avenues for future work.

4 Audience and Relevance

Fairness and ethical issues surrounding the practice of IR has become a major topic of concern among IR researchers [8–10,30,31,39]. The existence of gender stereotypes in IR systems can influence an individual's judgments, leading to unfair treatments and outcomes. In an ideal world, the expectation from IR systems is to be fair towards different gender identities and avoid reflecting unfair prejudices that may exist within society. We hope that our work contributes to the growing body of knowledge in this area, and helps the IR community to become familiar with the datasets, metrics, and methods that can be used for reducing the level of such biases in retrieval methods.

It is worth mentioning that there have been many attempts by industrial entities to address biases from a practical sense. For instance, we can point to the investigation of fairness in neural-based models by Microsoft, the responsible machine learning initiative at Twitter, which tackles gender and racial biases, or the PAIR group at Google Brain that explores responsible AI in Google systems.

We note that while there have been similar tutorials related to investigating fairness issues in IR systems in other venues, the topics proposed in this tutorial distinguish themselves by focusing on proposing systematic and well-validated methods for reducing gender biases in retrieval results. The following tutorials can be considered complementary and synergistic to the theme of ours:

1. *Addressing Bias and Fairness in Search Systems* by Ruoyuan Gao and Chirag Shah at SIGIR 2021. Similar to our topic, this tutorial focuses on introducing the issue of bias in data, algorithms, and search process.
2. *Fairness of Machine Learning in Recommender Systems* by Yunqi Li, Yingqiang Ge, Yongfeng Zhang at CIKM 2021. This tutorial introduces and describes fairness definitions as well as evaluation metrics in recommender systems.

3. *Fair Graph Mining* by Jian Kang, Hanghang Tong at CIKM 2021. The purpose of this tutorial is to introduce state-of-the-art techniques for increasing fairness on graph mining and describe challenges as well as future directions.

There are many other similar tutorials presented at major venues similar to the above. Our goal in this tutorial is to provide comprehensive knowledge about the methods and techniques that can be used for reducing gender bias in IR systems, while past tutorials are not related to retrieval tasks.

The target audience for this tutorial will be those who have interest in IR methods especially neural ranking models and well-known datasets. The tutorial will provide an overview of some of the IR concepts and components for those who are new to the field of IR. As such, sufficient details will be provided as appropriate so that the content will be accessible and understandable to those who only have a basic understanding of IR principles. This tutorial will only assume that the audiences is familiar with different topics included in an undergraduate IR course such as those covered in [26].

5 Presenters

Amin Bigdeli is a Data Scientist at Warranty Life and a Research Associate at Toronto Metropolitan University. His research work focuses on issues of fairness in information retrieval systems. Amin has published multiple research papers in this area in top IR venues such as SIGIR, CIKM, EDBT, and ECIR.
Negar Arabzadeh is a Research intern at Google brain working on fairness evaluation of text to image generation models. She is also completing her Ph.D. at the University of Waterloo. Her research is aligned with Ad hoc Retrieval and Conversational search in IR and NLP. Negar has published relevant papers in prestigious conferences and journals such as SIGIR, CIKM, ECIR, and IP&M. She previously interned at Microsoft Research and Spotify research and is one of the lead organizers of NeurIPS 2022 IGLU competition on NLP task.
Shirin Seyedsalehi is a Ph.D. student at Toronto Metropolitan University. Her research so far is focused on fairness in Information Retrieval and Neural Rankers. She has published papers in well known conferences such as SIGIR, CIKM and EDBT. She previously interned at Microsoft Research.
Morteza Zihayat is an Associate Professor and co-founder of the centre for Digital Enterprise Analytics & Leadership (DEAL) at Toronto Metropolitan University. His research concerns user modeling, applied machine learning and bias, debiasing, and fairness in NLP and IR. He has published in various well-respected journals and conferences in Information Retrieval, Machine Learning, and Information Systems such as IEEE TKDE, Information Processing and Management, ACM SIGKDD, SIGIR, ECIR, PKDD, SIAM, and SDM.
Ebrahim Bagheri is a Professor and the Director for the Laboratory for Systems, Software and Semantics (LS^3) at Toronto Metropolitan University. He holds a Canada Research Chair (Tier II) in Social Information Retrieval as well as an NSERC Industrial Research Chair in Social Media Analytics. He currently leads the NSERC Program on Responsible AI (http://responsible-ai.ca). He is

an Associate Editor for ACM Transactions on Intelligent Systems and Technology (TIST) and Wiley's Computational Intelligence.

6 Type of Support Materials

As for the supporting materials, we will publicly share a Github repository several weeks prior to the conference so the participants of the tutorial can familiarize themselves with the content. The repository will include a comprehensive slide deck, links to code, models, datasets, and run files.

References

1. Azzopardi, L.: Cognitive biases in search: a review and reflection of cognitive biases in information retrieval. In: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval, pp. 27–37 (2021)
2. Baeza-Yates, R.: Bias on the web. Commun. ACM **61**, 54–61 (2018)
3. Baeza-Yates, R.: Bias in search and recommender systems. In: Fourteenth ACM Conference on Recommender Systems, p. 2 (2020)
4. Bagheri, E., Ensan, F., Al-Obeidat, F.: Neural word and entity embeddings for ad hoc retrieval. Inf. Proc. Manag. **54**(4), 657–673 (2018)
5. Basta, C., Costa-Jussà, M.R., Casas, N.: Evaluating the underlying gender bias in contextualized word embeddings. arXiv preprint arXiv:1904.08783 (2019)
6. Bigdeli, A., Arabzadeh, N., SeyedSalehi, S., Zihayat, M., Bagheri, E.: Gender fairness in information retrieval systems. In: Proceedings of the 45th International ACM SIGIR Conference (2022)
7. Bigdeli, A., Arabzadeh, N., Seyedsalehi, S., Zihayat, M., Bagheri, E.: A light-weight strategy for restraining gender biases in neural rankers. In: Hagen, M., et al. (eds.) ECIR 2022. LNCS, vol. 13186, pp. 47–55. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99739-7_6
8. Bigdeli, A., Arabzadeh, N., Seyersalehi, S., Zihayat, M., Bagheri, E.: On the orthogonality of bias and utility in ad hoc retrieval. In: Proceedings of the 44rd International ACM SIGIR Conference (2021)
9. Bigdeli, A., Arabzadeh, N., Zihayat, M., Bagheri, E.: Exploring gender biases in information retrieval relevance judgement datasets. In: Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) ECIR 2021. LNCS, vol. 12657, pp. 216–224. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72240-1_18
10. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
11. Bordia, S., Bowman, S.R.: Identifying and reducing gender bias in word-level language models (2019)
12. Brunet, M.E., Alkalay-Houlihan, C., Anderson, A., Zemel, R.: Understanding the origins of bias in word embeddings. In: International Conference on Machine Learning, pp. 803–811. PMLR (2019)
13. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)

14. Draws, T., Tintarev, N., Gadiraju, U., Bozzon, A., Timmermans, B.: This is not what we ordered: exploring why biased search result rankings affect user attitudes on debated topics (2021)
15. Ekstrand, M.D., Das, A., Burke, R., Diaz, F.: Fairness in information access systems. arXiv preprint arXiv:2105.05779 (2021)
16. Fabris, A., Purpura, A., Silvello, G., Susto, G.A.: Gender stereotype reinforcement: measuring the gender bias conveyed by ranking algorithms. Inf. Proc. Manag. **57**(6), 102377 (2020)
17. Font, J.E., Costa-Jussa, M.R.: Equalizing gender biases in neural machine translation with word embeddings techniques. arXiv preprint arXiv:1901.03116 (2019)
18. Gerritse, E.J., Hasibi, F., de Vries, A.P.: Bias in conversational search: the double-edged sword of the personalized knowledge graph. In: Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval (2020)
19. Klasnja, A., Arabzadeh, N., Mehrvarz, M., Bagheri, E.: On the characteristics of ranking-based gender bias measures. In: 14th ACM Web Science Conference 2022, pp. 245–249 (2022)
20. Krieg, K., Parada-Cabaleiro, E., Medicus, G., Lesota, O., Schedl, M., Rekabsaz, N.: Grep-BiasIR: a dataset for investigating gender representation-bias in information retrieval results. arXiv preprint arXiv:2201.07754 (2022)
21. Krieg, K., Parada-Cabaleiro, E., Schedl, M., Rekabsaz, N.: Do perceived gender biases in retrieval results affect relevance judgements. In: Boratto, L., Faralli, S., Marras, M., Stilo, G. (eds.) BIAS 2022. Communications in Computer and Information Science, vol. 1610, pp. 104–116. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-09316-6_10
22. Kulshrestha, J., et al.: Quantifying search bias: investigating sources of bias for political searches in social media. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 417–432 (2017)
23. Liu, H., Dacon, J., Fan, W., Liu, H., Liu, Z., Tang, J.: Does gender matter? Towards fairness in dialogue systems. arXiv preprint arXiv:1910.10486 (2019)
24. Liu, H., Wang, W., Wang, Y., Liu, H., Liu, Z., Tang, J.: Mitigating gender bias for neural dialogue generation with adversarial learning (2020)
25. Lu, K., Mardziel, P., Wu, F., Amancharla, P., Datta, A.: Gender bias in neural natural language processing. In: Nigam, V., et al. (eds.) Logic, Language, and Security. LNCS, vol. 12300, pp. 189–202. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62077-6_14
26. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
27. Nguyen, T., et al.: MS MARCO: a human generated machine reading comprehension dataset. In: CoCo@ NIPS (2016)
28. Olteanu, A., et al.: FACTS-IR: fairness, accountability, confidentiality, transparency, and safety in information retrieval. In: ACM SIGIR Forum, vol. 53, pp. 20–43. ACM New York, NY, USA (2021)
29. Prost, F., Thain, N., Bolukbasi, T.: Debiasing embeddings for reduced gender bias in text classification. arXiv preprint arXiv:1908.02810 (2019)
30. Rekabsaz, N., Kopeinik, S., Schedl, M.: Societal biases in retrieved contents: measurement framework and adversarial mitigation for BERT rankers (2021)
31. Rekabsaz, N., Schedl, M.: Do neural ranking models intensify gender bias?. In: Proceedings of the 43rd International ACM SIGIR Conference (2020)

32. SeyedSalehi, S., Bigdeli, A., Arabzadeh, N., Mitra, B., Zihayat, M., Bagheri, E.: Bias-aware fair neural ranking for addressing stereotypical gender biases. In: EDBT, pp. 2–435 (2022)
33. Seyedsalehi, S., Bigdeli, A., Arabzadeh, N., Zihayat, M., Bagheri, E.: Addressing gender-related performance disparities in neural rankers. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2484–2488 (2022)
34. Stanczak, K., Augenstein, I.: A survey on gender bias in natural language processing. arXiv preprint arXiv:2112.14168 (2021)
35. Sun, T., et al.: Mitigating gender bias in natural language processing: literature review. arXiv preprint arXiv:1906.08976 (2019)
36. Wang, J., Liu, Y., Wang, X.E.: Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. arXiv preprint arXiv:2109.05433 (2021)
37. Yang, Z., Feng, J.: A causal inference method for reducing gender bias in word embedding relations. In: Proceedings of the AAAI Conference (2020)
38. Zhao, J., Mukherjee, S., Hosseini, S., Chang, K.W., Awadallah, A.H.: Gender bias in multilingual embeddings and cross-lingual transfer (2020)
39. Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., Chang, K.W.: Gender bias in contextualized word embeddings. arXiv preprint arXiv:1904.03310 (2019)