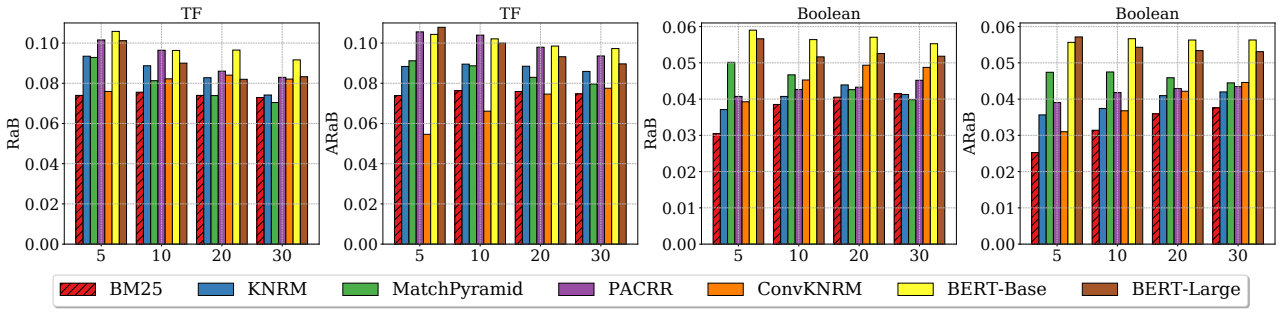

Figure 1: Results of retrieval gender bias metrics. Higher (positive) values indicate higher bias towards male in retrieval results.

both document gender magnitude measures `TF` and `Boolean`, are shown in Table 3. In the results of the neural ranking models (except the BERT-based ones), the reported values belong to the models with pre-trained GloVe embeddings; the values in parentheses indicate the changes in the retrieval gender bias values in comparison to the ones of the corresponding *RND* models (when the word embeddings are initialized randomly). An arrow down/up in the parentheses indicates a decrease/increase in the bias when using a *RND* model. For easier visual comparison, Figure 1 depicts the results in plots.

The results show the inclination of all the IR models towards the male concepts (despite using non-gendered queries). The neural models consistently increase retrieval gender bias in comparison with BM25 in almost all variations (4 exceptions out of 96 variations). The BERT models, and especially BERT-Base, show the overall highest degrees of gender bias. These confirm that the neural models, despite better retrieval performance, on the whole intensify gender bias in retrieval results toward male when compared with BM25.

Finally, we look at the effect of using pre-trained word embeddings on the retrieval gender bias of neural ranking models. Based on the results, transfer learning either increases (cases with down arrows) or does not affect (0 values) gender bias in the majority of the cases, namely 53 out of 64 cases. This therefore shows that transfer learning tends to increase gender bias in retrieval results.

5 CONCLUSION AND FUTURE WORK

This work takes a first step in measuring the degree of gender bias in retrieval models and studying it in neural IR models. We propose a novel framework to measure gender bias in retrieval results and provide a set of human-annotated non-gendered queries. By submitting these queries to various IR models, we show that the studied neural ranking models intensify gender bias towards male concepts in comparison with BM25. The fine-tuned BERT models show the highest degrees of bias. We also observe that the neural ranking models (excluding the BERT ones) generally increase gender bias when they use transfer learning.

Future research following this work further investigates the relation between bias and relevance in retrieval. Based on the results of this study, the gender bias values of the neural IR models do not fully correlate with their performance. This encourages a deeper analysis of the neural models. Another direction of research is the study of methods to eliminate gender bias in neural IR models while preserving their effectiveness, especially in the light of literature on fairness in ranking and embedding debiasing. Finally, exploring other metrics of bias measurement, as well as studying the relations between the metrics and the human perception of bias in retrieval results, are other future avenues of this work.

ACKNOWLEDGMENTS

Many thanks to Sophia Freynhofer for her help and advice on designing crowd sourcing experiments.

REFERENCES

[1] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* (2018).
[2] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *Proc. of SIGIR*.
[3] 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. In *Proc. of NeurIPS*.
[4] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* (2017).
[5] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the impact of gender on rank in resume search engines. In *Proc. of CHI*.
[6] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proc. of WSDM*.
[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL-HTL*.
[8] Yanai Elazar and Yoav Goldberg. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.
[9] Sebastian Hofstätter, Navid Rekabsaz, Carsten Eickhoff, and Allan Hanbury. 2019. On the Effect of Low-Frequency Terms on Neural-IR Models. In *Proc. of SIGIR*.
[10] Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2017. PACRR: A Position-Aware Neural IR Model for Relevance Matching. In *Proc. of EMNLP*.
[11] Matthew Kay, Cynthia Matuszek, and Sean Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proc. of CHI*.
[12] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, M. Bilal Zafar, S. Ghosh, K. P Gummadi, and K. Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proc. of CSCW*.
[13] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268* (2016).
[14] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).
[15] Jahna Otterbacher, Jo Bates, and Paul Clough. 2017. Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proc. of CHI*.
[16] Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. 2018. Investigating user perception of gender bias in image search: the role of sexism. In *Proc. of SIGIR*.
[17] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text matching as image recognition. In *Proc. of AAAI*.
[18] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proc. of EMNLP*.
[19] Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in IR* (2009).
[20] K Sparck Jones and C Van Rijsbergen. 1975. Report on the Need for and Provision of an 'ideal' information retrieval test collection. *British Library Research and Development Report 5266* (1975).
[21] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-end neural ad-hoc ranking with kernel pooling. In *Proc. of SIGIR*.
[22] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proc. of Conference on Scientific and Statistical Database Management*.
[23] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, V. Ordonez, and K. Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *Proc. of NAACL-HTL*.