Are Fair Learning To Rank Models Really Fair? An Analysis Using Inferred Gender

Alexander Pietrick
Departments of Computer Science and Data Science
Worcester Polytechnic Institute
Worcester, MA, United States
ajpietrick@wpi.edu

Alyssa Romportl
Departments of Mathematics and Psychology
St. Olaf College
Northfield, MN, United States
rompor1@stolaf.edu

Shailen Smith
Department of Mathematics
Stony Brook University
Stony Brook, NY, United States
shailen.smith@stonybrook.edu

Oluseun Olulana
Department of Data Science
Worcester Polytechnic Institute
Worcester, MA, United States
omolulana@wpi.edu

Kathleen Cachel
Department of Data Science
Worcester Polytechnic Institute
Worcester, MA, United States
kcachel@wpi.edu

Elke Rundensteiner
Department of Data Science
Worcester Polytechnic Institute
Worcester, MA, United States
rundenst@wpi.edu

2022 IEEE MIT Undergraduate Research Technology Conference (URTC) | 978-1-6654-7345-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/URTC56832.2022.10002233

***Abstract*—Fair Learning To Rank (LTR) frameworks require demographic information; however, that information is often unavailable. Inference algorithms may infer the missing demographic information to supply to the fair LTR model. In this study, we analyze the effect of using a trained fair LTR model with uncertain demographic inferences. We show that inferred data results in varying levels of fairness and utility depending on inference accuracy. Specifically, less accurate inferred data adversely affects the rankings' fairness, while more accurate inferred data creates fairer rankings. Therefore, we recommend that a careful evaluation of demographic inference algorithms before use is critical.**

***Keywords*—Fair ranking, learning to rank, demographic inference, fairness, uncertainty**

I. Introduction

Ranking algorithms have become increasingly widespread through their use in job candidate searches, lending, college admissions decisions, and more [1]. As a result, controlling for social biases in these algorithms has become an important area of focus. A considerable amount of work has been done to integrate fairness into automated systems [2]. A Fair Learning To Rank (LTR) framework works to reduce bias against a *protected group*, a group against which it is illegal to discriminate [3]. Reducing this bias may create a ranking that is less relevant to the targeted use of the ranked result list. Thus, fair LTR models affect both fairness and utility of the ranking [2].

Fair LTR frameworks need demographic information about protected groups to control for potential systemic biases in the training data. However, in practice, such demographic information is often unavailable [4]. To solve this problem, AI system developers may use demographic inference methods to infer the missing information. Such information could then be supplied to the fair LTR model, allowing the model to function [5] [6]. Inference methods make use of the available data, which could be first names, last names, zip code, images, and even email addresses, to predict demographic information such as race or gender of individuals, which is needed by the model to ensure fairness.

However, misclassifications by AI inference mechanisms may lead to unintended consequences and inadvertently introducing bias. Santamaría and Mihaljević [5] have compared the efficacy of several gender inference methods, finding that errors are common and accuracy is never guaranteed. This raises the critical research question: Are "fair" AI algorithms, such as fair LTR, actually fair when applied to real-world datasets with potential missing demographic information?

Our Approach. In this research, we investigate how uncertainty in gender inference affects the performance of a trained fair LTR model. We adopt DELTR, a popular fair LTR model created in 2020 by Zehlike and Castillo [2]. Furthermore, we work with three real-world gender inference algorithms: Facebook Generated Name List [7], Genderize.io [8], and Gender-API [9].

We then employ the COMPAS dataset collected by ProPublica [10], which describes Broward County defendants in 2013 and 2014, as well as ground-truth demographic information including "male" and "female" assignments[1] for each defendant–which is essential for assessing our research question. This information is used to train both *fairness-aware* and *fairness-unaware* LTR models. We then analyze the performance of the models in terms of both fairness and utility.

Deploying these two models on both the data with actual ground-truth gender and inferred gender, we sought to answer

[1]While "male" and "female" are typically sex terms, we refer to this information as "gender' since we are applying gender inference algorithms.

978-1-6654-7345-3/22/$31.00 ©2022 IEEE

the following important research questions:

- When using inferred demographic information, are the benefits of a fairness-aware LTR model still present compared to the use of a fairness-unaware LTR model?
- How does the performance of a fairness-aware LTR model change when using inferred demographic information instead of actual demographic information?

We find that less accurate inferred gender adversely affects the rankings' fairness, while more accurate inferred gender creates fairer rankings. We recommend a careful evaluation of demographic inference algorithms before use to ensure no harm is inadvertently done to disadvantaged groups.

II. Background and Related Work

We briefly discuss how fairness has been defined in the literature, introduce fair ranking methods, and present previous work in fairness and demographic inference.

A. Algorithmic Fairness

Discrimination is ever present in our world and as such, often shows up in data-driven modeling. Thus, fairness has now become a critical concern in the context of applying machine learning in the real world [11]. Žliobaitė [12] defines fairness as: (1) people that are similar in terms of non-protected characteristics should receive similar predictions, and (2) differences in predictions across groups of people can only be as large as justified by non-protected characteristics. Additionally, fairness literature has defined group fairness– as opposed to individual fairness–which seeks to achieve demographic parity and equalized odds [13].

B. Fair Ranking Algorithms

Fair ranking algorithms can be defined under three categories: *pre-*, *in-*, and *post-processing* methods [2]. Pre-processing aims to reduce bias at the training stage. In-processing learns a model that can control for bias. Post-processing is instead given an unfairly ranked list, then re-ranks the list to improve the fairness. Zehlike and Castillo [2] show that *pre-processing* and *post-processing* methods can be problematic, while *in-processing* methods tend to be more effective in addressing these problems.

C. Inference Algorithms

Demographic inference algorithms attempt to predict the demographic information, like race or gender, of individuals. They are commonly used in fair LTR models when demographic information is missing. However, there is little work on understanding the effect of demographic inference on fairness. The one exception we are aware of is the recent study by Ghosh et al. [6], which is restricted to *post-processing* fair ranking solutions. Ghosh et al. [6] found that using inference algorithms in the context of post-processing fair ranking tends to result in more harm to vulnerable groups than without any fair re-ranking method. They report that the detriment inferred data has on fairness is often hard to predict, which prompted us to explore this research question in the context of in-processing methods–a currently open research question.

III. Algorithms and Metrics

A. Metrics for Ranking Evaluation

Metrics have been designed to quantify fairness and utility of the results of fair ranking solutions. Two popular metrics are Skew [1], a representation-based fairness metric, and NDCG [14], a utility metric. Representation-based fairness metrics consider the underlying population in the dataset with respect to the protected group G_1 and its proportion in the top entries in the ranking generated by the fair ranking algorithm. On the other hand, utility metrics determine the quality of a ranking, i.e., how relevant the top entries are to the targeted application (e.g., towards meeting their employee hiring objective).

Skew. First introduced in 2019 by Geyik et al. [1], the $Skew$ for a group G_i at position k in a ranking τ is

$$Skew_{G_i}@k(\tau) = \frac{p_{\tau^k,G_i}}{p_{\tau,G_i}} \tag{1}$$

where p_{τ^k,G_i} denotes the proportion of members in group G_i in the top k positions of the ranking τ, p_{τ,G_i} the proportion of members in group G_i of the entire ranking τ, and $i \in \{0,1\}$. $Skew_{G_i}@k(\tau) > 1$ indicates the group G_i is over-represented in the top k elements of τ, whereas $Skew_{G_i}@k(\tau) < 1$ means the group G_i is under-represented.

We also evaluate fairness at a position using the difference in Skew values between the two relative groups, which we call $\Delta Skew$:

$$\Delta Skew@k(\tau) = |Skew_{G_0}@k(\tau) - Skew_{G_1}@k(\tau)| \tag{2}$$

for a ranking τ with non-protected group G_0 and protected group G_1 (e.g., males and females). $\Delta Skew@k(\tau) = 0$ occurs only when both groups G_0 and G_1 have skew values of 1 at position k, i.e., both groups are represented accurately in the top k elements. Thus, values of ΔSkew closer to 0 can be considered more fair.

NDCG. As proposed by Järvelin and Kekäläinen in 2002 [14], Normalized Discounted Cumulative Gain (NDCG) is commonly used to judge overall utility of a ranking τ.

$$NDCG(\tau) = \frac{1}{Z}\sum_{i=1}^{|\tau|}\frac{s_i}{\log_2(i+1)} \tag{3}$$

where s_i represents the utility score of the ith element in the ranking τ, and $Z = \sum_{i=1}^{|\tau|}\frac{1}{\log_2(i+1)}$. Larger NDCG values correspond to rankings with higher overall utility.

B. Fair Learning to Rank Algorithm

For our study, we select the fair LTR framework *DELTR* created by Zehlike and Castillo [2], which builds off of the ListNet algorithm [15]. Like ListNet, DELTR is an in-processing fair ranking algorithm built for supporting fairness relative to a protected and non-protected group. Additionally, DELTR actively considers the average exposure of each group G_i to ensure equal treatment of members within these groups. Exposure asserts that items at the top of a ranked list will receive more attention from the reader than items at the

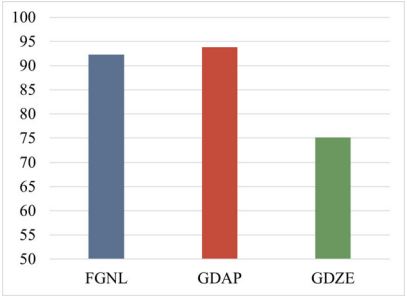

Fig. 1. Inference Percent Accuracy

Fig. 2. Sankey plot for FGNL. "M" represents male and "F" represents female.

bottom. Thus, DELTR is mainly concerned with increasing fairness for the top items of a ranking.

DELTR balances fairness and utility using the parameter γ. Training DELTR with $\gamma = 0$ does not consider exposure of each group, creating a *fairness-unaware* trained model that solely considers utility in its ranking. A non-zero γ value will prioritize a model that provides equal average exposure between the protected and non-protected groups instead of focusing solely on accurate utility scores. For our study, we define a *fairness-aware* model as one where $\gamma = 1$.

C. Gender Inference Methods

In order to accommodate the binary nature of DELTR and the current limitations of inference algorithms, we adopt a simplistic binary system to classify gender.

The gender inference algorithms, given a first name, will infer and return a gender. Some of the algorithms return unknown or ambiguous values, which we handle by assigning them as "male" because the COMPAS dataset is majority male.

We use three popular algorithms for gender inference:

Facebook Generated Name List (FGNL): All inference calls to FGNL use a static database created from observing the gender listed on Facebook users' profiles [7].

Gender-API (GDAP): Trained on publicly available governmental archives and social network information, GDAP is a high-end gender inference algorithm [9]. The algorithm accepts both first name and full names to provide gender inferences with supporting meta data such as confidence percentage and regional information.

Genderize.io (GDZE): GDZE is a widely used gender inference algorithm that bases its predictions off of a person's first name and, optionally, their country [8]. It has been used by sources such as The Washington Post, The Atlantic, and The Guardian.

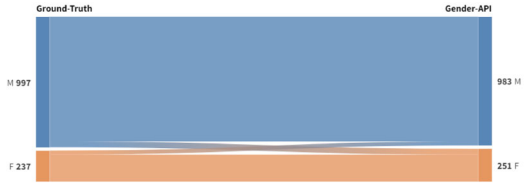

Fig. 3. Sankey plot for Gender-API.

Fig. 4. Sankey plot for Genderize.io.

Figure 1 highlights the percent accuracy–i.e., correctly inferred gender–of our inference process after unknown values were assigned to "male". Figures 2-4 show the number of accurate and inaccurate classifications for each algorithm on the COMPAS test data. We see that FGNL and GDAP have similar accuracies, while GDZE is much less accurate.

IV. Experiments

A. COMPAS Dataset

Commonly used in fairness research [16] [17], the COMPAS dataset was collected by ProPublica for an article assessing discriminatory patterns in the COMPAS Recidivism assessment, a national screening tool for future criminal behavior [10]. The data describes 6,172 Broward County pretrial defendants from 2013 and 2014 with name and sex information for each defendant. The non-demographic features that the trained DELTR model utilizes to predict COMPAS Recidivism scores for each defendant are juvenile felony count, juvenile misdemeanor count, total priors count (including juvenile felonies and misdemeanors), and days in jail.

B. Analysis Method

Our method can be broken into several parts which are described below.

Prepare Dataset. We first produce an 80/20 training/testing split on the dataset, i.e., 80% of the items in our dataset are used for training, and the other 20% are used for testing.

Train DELTR. Using the training split, we apply the publicly available code[2] to train two DELTR models, a fairness-unaware ($\gamma = 0$) and fairness-aware ($\gamma = 1$). This dataset includes ground-truth gender information, with females as the protected group G_1.

Inference Algorithms. We compute the inferred gender for the COMPAS test split separately for our three inference algorithms FGNL, GDZE, and GDAP, generating three copies

[2]https://github.com/fair-search/fairsearch-deltr-python

of the test split with inferred gender information. We keep a fourth testing split with actual gender information.

Test DELTR. We run the four test datasets through the two trained DELTR models, obtaining eight distinct rankings.

Metric Computation. We input each ranking shown in Table 1 through the metrics described in Section 3A.

Ranking	Inference	Fair-
ORIG	none	none
ORCL	none	aware
vLTR	none	unaware
vFGNL	FGNL	unaware
vGDAP	Gender-API	unaware
vGDZE	Genderize.io	unaware
FGNL	FGNL	aware
GDAP	Gender-API	aware
GDZE	Genderize.io	aware

TABLE I
RANKED SETS

V. RESULTS

We present our results corresponding to our two research questions.

A. Fairness-aware vs. Fairness-unaware Models

We first answer the question: When using inferred demographic information, are the benefits of a fairness-aware LTR model still present compared to the use of a fairness-unaware LTR model?

Shown in Figures 5 and 6, the fairness-aware ground-truth ranking has a smaller ΔSkew value and a larger NDCG value than the fairness-unaware ground-truth ranking, as expected. We note that the fairness-unaware ground-truth, FGNL, and GDAP rankings show the same high ΔSkew and NDCG values. This is intuitive, as small differences in gender information should have little effect on a fairness-unaware model. For both the FGNL and GDAP rankings, the fairness-aware models show a notable improvement in ΔSkew compared to the fairness-unaware models, with just a marginal decrease in NDCG. However, GDZE shows no difference in ΔSkew and an increase in NDCG from the fairness-unaware to the fairness-aware model. The differences in the behavior of GDZE is likely due to its lower percent accuracy.

B. Performance Using Inferred Data

We then answer the question: How does the performance of a fairness-aware LTR model change when using inferred demographic information instead of actual demographic information?

By examining Figure 7, it is clear that the original ranking (ORIG) is the least fair, showing a significant under-representation of females in the top 20 positions. The fairness-aware model with actual gender information (ORCL) over-corrects for this and over-represents females. The skews of FGNL, GDAP, and GDZE are closer to 1 than ORCL, showing

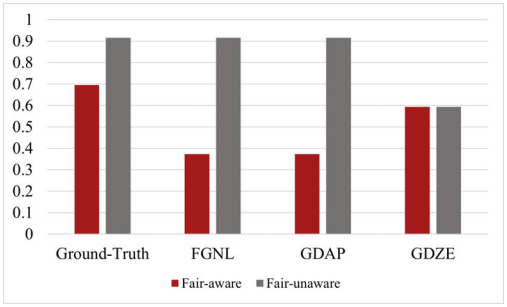

Fig. 5. ΔSkew values at position 20 for the eight rankings described in Section 4B.

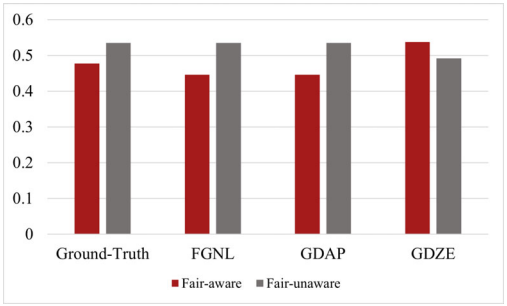

Fig. 6. NDCG utility values at position 20 for the eight rankings described in Section 4B.

that the misclassifications of the inferences actually help with equal representation by diminishing DELTR's over-correction.

However, these results do not hold at all positions. Shown in Figure 8, GDZE performs much worse than ORCL, FGNL, and GDAP at positions 10 and 100 in terms of fairness. In fact, ΔSkew doesn't change depending on position; males are over-represented up to 100 positions. As shown in Figure 4, GDZE often misclassifies females as males, thus DELTR fails to place these females in higher positions, which leads to a persistent over-representation of males. Overall, our most accurate inference algorithm, GDAP, receives the lowest ΔSkew, thus making it the most fair.

When observing Figure 9, we see that ORIG has the highest utility, highlighting the decrease in relevance that often comes from DELTR's consideration of fairness. NDCG remains roughly the same for the other rankings, showing that inferred data may have a negligible effect on utility.

VI. DISCUSSION

We study the effect of inputting inferred gender information to a trained fair LTR model. We report that less accurate inferred gender from GDZE adversely affects the rankings' fairness, while more accurate inferred gender from FGNL and GDAP create fairer rankings. We find that FGNL and GDAP give very similar results, most likely because their accuracies are almost the same.

Our results for NDCG show that there is indeed a trade-off when considering fairness and utility. Our least fair ranking, ORIG, has the highest NDCG, whereas the most fair, ORCL, FGNL, and GDAP, have the lowest NDCG.

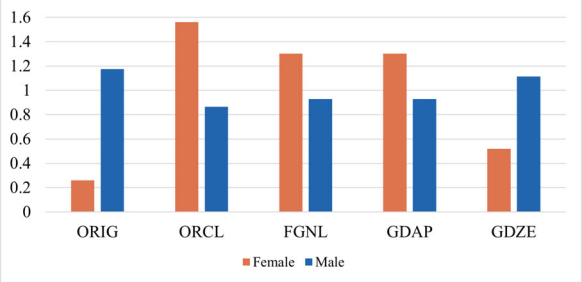

Fig. 7. Skew at position 20 for the original ranking and each fairness-aware ranking.

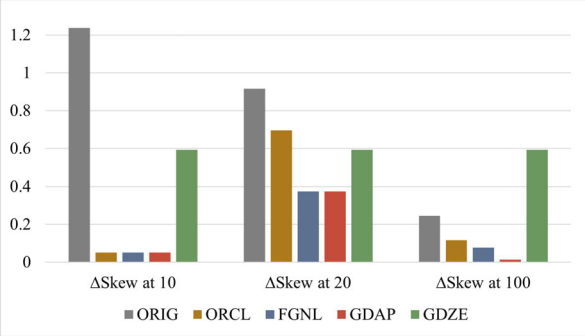

Fig. 8. ΔSkew at position 20 for the original ranking and each fairness-aware ranking.

Limitations and Future Work. Due to time limitations, our study focuses on one dataset. In addition, this dataset included more males than females, which may have influenced our results. Future work should test on additional real datasets to discover the impact inference methods could have on real people. We also suggest more research investigating other inferred demographic information like race.

Additionally, we are limited by the binary nature of DELTR. Thus, we only used the labels "male" and "female" without including other gender identities. Future research should look into fair LTR frameworks and inference methods that can accommodate more than two groups, including non-binary gender identities.

Our research shows that using inference algorithms as input to a fair LTR model can be problematic and unpredictable with varying levels of fairness and utility. We recommend the careful evaluation of demographic inference algorithms before use with fair LTR models to optimize fairness and ensure groups are not inadvertently harmed.

ACKNOWLEDGMENTS

The authors would like to thank the National Science Foundation for funding this research under the NSF REU site grant 1852498 and IIS grant 2007932. We also thank Dr. Kelsey Briggs and other WPI DS REU students for their great assistance.

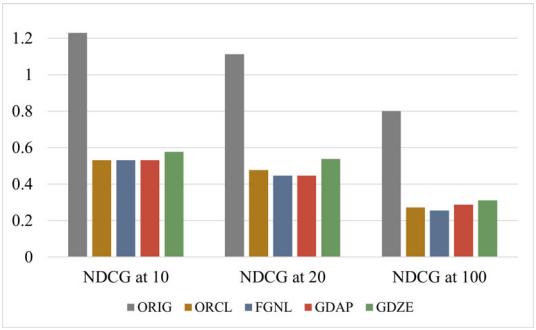

Fig. 9. NDCG at position 20 for the original ranking and each fairness-aware ranking.

REFERENCES

[1] S. C. Geyik, S. Ambler, and K. Kenthapadi, "Fairness-aware ranking in search recommendation systems with application to Linkedin talent search," in *Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and DataMining*, 2019, pp. 2221–2231.

[2] M. Zehlike and C. Castillo, "Reducing disparate exposure in ranking: A learning to rank approach," in *Proceedings of the Web Conference 2020*, 2020, pp. 2849–2855.

[3] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 2018, pp. 1–7.

[4] M. Andrus, E. Spitzer, J. Brown, and A. Xiang, ""what we can't measure, we can't understand": Challenges to demographic data procurement in the pursuit of fairness." arXiv, 2020. [Online]. Available: https://arxiv.org/abs/2011.02282

[5] L. Santamaría and H. Mihaljević, "Comparison and benchmark of name-to-gender inference services," *PeerJ Computer Science*, vol. 4, p. e156, 2018.

[6] A. Ghosh, R. Dutt, and C. Wilson, "When fair ranking meets uncertain inference," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1033–1043.

[7] C. Tang, K. Ross, N. Saxena, and R. Chen, "What's in a name: A study of names, gender inference, and gender behavior in facebook," in *International Conference on Database Systems for Advanced Applications*. Springer, 2011, pp. 344–356.

[8] E. Ehrhardt. (2018). [Online]. Available: https://genderize.io/

[9] M. Perl. (2018). [Online]. Available: https://gender-api.com/

[10] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. (2016) How we analyzed the COMPAS recidivism algorithm. [Online]. Available: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[11] M. Hardt. (2014) How big data is unfair: Understanding sources of unfairness in data driven decision making. [Online]. Available: https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de

[12] I. Žliobaitė, "A survey on measuring indirect discrimination in machine learning," 2015.

[13] A. Singh and T. Joachims, "Fairness of exposure in rankings," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2219–2228.

[14] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Transactions on Information systems*, vol. 20, no. 4, pp. 422–446, 2002.

[15] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 129–136. [Online]. Available: https://doi.org/10.1145/1273496.1273513

[16] K. Yang and J. Stoyanovich, "Measuring fairness in ranked outputs," in *Proceedings of the 29th international conference on scientific and statistical database management*, 2017, pp. 1–6.

[17] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates, "Fa*ir: A fair top-k ranking algorithm," ser. CIKM '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 1569–1578. [Online]. Available: https://doi.org/10.1145/3132847.3132938