


# Toward Automatic Group Membership Annotation for Group Fairness Evaluation <sup>\*</sup>

Fumian Chen <sup>1</sup>[0009-0001-2391-6578], Dayu Yang<sup>2</sup>[0009-0006-7360-1837], and Hui Fang<sup>3</sup>[0009-0003-1904-787X]

University of Delaware, Newark DE 19702, USA<sup>1,2,3</sup>  
{fmchen, dayu, hfang}@udel.edu

**Abstract.** With the increasing research attention on fairness in information retrieval systems, more and more fairness-aware algorithms have been proposed to ensure fairness for a sustainable and healthy retrieval ecosystem. However, as the most adopted measurement of fairness-aware algorithms, group fairness evaluation metrics, require group membership information that needs massive human annotations and is barely available for general information retrieval datasets. This data sparsity significantly impedes the development of fairness-aware information retrieval studies. Hence, a practical, scalable, low-cost group membership annotation method is needed to assist or replace human annotations. This study explored how to leverage language models to automatically annotate group membership for group fairness evaluations, focusing on annotation accuracy and its impact. Our experimental results show that BERT-based models outperformed state-of-the-art large language models, including GPT and Mistral, achieving promising annotation accuracy with minimal supervision in recent fair-ranking datasets. Our impact-oriented evaluations reveal that minimal annotation error will not degrade the effectiveness and robustness of group fairness evaluation. The proposed annotation method reduces tremendous human efforts and expands the frontier of fairness-aware studies to more datasets.

**Keywords:** Information Retrieval · Fairness Evaluation · Annotation.

## 1 Introduction

From social media to open web searches, information retrieval (IR) systems are ubiquitous and can fundamentally impact how people receive and seek information. As people started to notice the issue of the echo chamber, the polarized online community, and the importance of covering diverse results [9], fairness-aware IR and its evaluation metrics became emerging needs to combat unfairness and biased representation for long-term sustainability [33]. Group fairness evaluation metrics are the most adopted metrics, measuring the disparity between a situation to be evaluated and its ideal situation. When applying them, one of the necessities is the group membership (GM) annotations, which define whether an

---

<sup>\*</sup> Supported by Institute for Financial Services Analytics at the University of Delaware

item is from underrepresented groups. Without GM annotation, applying group fairness evaluation metrics on retrieval results is infeasible, and it is also impossible to apply supervised learning-based fair ranking algorithms that rely on GM annotation for training [4,32]. Therefore, before evaluating group fairness or allocating exposure to the documents, we must know their group membership.

Annotations are usually obtained through costly human annotators, such as crowd annotators and domain experts. High-quality annotations involving annotators’ training, and cross-validation are even more expensive [18]. The annotation process requires annotators to interpret documents’ context and then assign pre-defined labels to the documents based on their contextual information. Since it is very similar to a text classification process, various attempts have been proposed to assist or replace human annotations, especially with the emergence of advanced NLP techniques that can accurately capture contextual features from text [16,18]. However, most of these attempts to replace human annotations focus on accuracy compared with human annotation but ignore the impact when enforcing this replacement on different tasks. It remains unclear how annotation errors would impact the final metrics with machine-learned annotations, especially when previous studies have shown that document-level error might be eliminated when aggregating to higher levels [1]. Since group fairness evaluations are also aggregated metrics, the annotation error might not hurt the ability to evaluate fairness for IR systems. The relation between annotation accuracy and the final evaluation metrics deserves our attention. Moreover, even though generative large language models (LLMs) are not designed for discriminative tasks like text classification, the increasing trend of using generative large language models (LLMs) such as OpenAI GPT on downstream NLP tasks is pushing more and more researchers to scramble for their applications [13]. However, given its economical and computational cost, are generative models with billions of parameters better than discriminative models for fairness-related annotation tasks?

Therefore, to explore how to replace human GM annotation effectively and economically and solve the issue of data sparsity, we compared the performance of four representative language models in predicting group membership for group fairness evaluation. Then, we comprehensively studied the impact of replacing human GM annotation for group fairness evaluations in recent fair-ranking datasets. Confirming the effectiveness of the new GM annotation method with minimal supervision, we believe our work opened a new direction to reduce human efforts on GM annotation and augment traditional IR datasets for future fairness-aware studies. Our implementation code will be available at <https://github.com/fm-chen/nldb-experiments>.

## 2 Related Work

With the rapid development of NLP, especially with the emergence of masked language models such as BERT [6] and generative large language models like OpenAI GPT [26], more and more NLP-related work has been proposed to save or even replace human efforts. Text classification, which assigns one of the pre-defined labels to a given text sequence, is one of the classic NLP tasks. As

one of the most powerful language models, BERT provides various pre-trained models that accurately capture linguistic and semantic information out of text [19]. With proper fine-tuning, previous studies have used BERT for multiple annotation tasks, such as image labeling and dataset annotation, and shown promising results even with fewer training samples and imbalanced class distributions [22,20]. Recently, as generative LLMs have become a hot topic, OpenAI GPT has also attracted increasing research interest, including the use of GPT to assist annotation and labeling. Generative models like GPT have shown to be a valuable tool for predicting searcher preference, validating and assisting human annotations and labelings [31,24,14,8]. Compared with BERT, which has a parameter size from 30 million to about 350 million, LLMs usually involves billions to over hundred billions of parameters, making fine-tuning and using LLMs costly [8]. Even with the open-sourced LLM, Mistral with seven billion parameters [17], deploying the model locally is computationally costly. Since generative models are not designed for discriminative tasks like text classification, previous studies revealed that using LLMs effectively requires meticulously prompt design. Their performance varies dramatically under different contexts [3,34]. Therefore, instead of scrambling for LLMs, we would like to explore the accuracy and impact of using different language models to replace human GM annotations for fairness evaluation tasks.

This work is also closely related to fairness-aware IR and its evaluation metrics. Well-adopted fairness evaluations [7,12,27,29,25] were based on exposure, and their fairness metrics either measure the deviation between system-produced and target exposure distribution or measure the inequality of exposure across groups. Another group of fairness evaluation is based on pair-wise metrics measuring the difference between pairs [2,23]. They are all aggregated measures that treat groups instead of individual documents as the basic unit, and the impact of replacing the costly human GM annotation with NLP techniques is unclear and has never been studied before. Thus, to save human efforts in obtaining GM annotations and solve the issue of data sparsity in fairness evaluation, this study tested four language models to obtain GM annotations and explored the impact of replacing human annotations with different annotators.

### 3 Automate GM Annotation for Fairness Evaluation

#### 3.1 GM in Fairness Evaluations

Group membership (GM) is one of the most essential components in group fairness evaluation. Depending on fairness evaluation goals, group membership can involve one or more fairness categories, such as gender and geographic location. As shown in Fig. 1, to make sure that a search engine result page (SERP) contains items from different geographic locations, we have to know each item’s geographic location information (geographic GM annotation) first. With the GM annotation, merits or exposure distributions across groups can be formulated to construct fairness evaluation. For example, the TREC fair ranking track 2021 [10] and 2022 [11] <sup>1</sup> use the attention-weighted rank fairness (AWRF), a widely

<sup>1</sup> <https://fair-trec.github.io/>

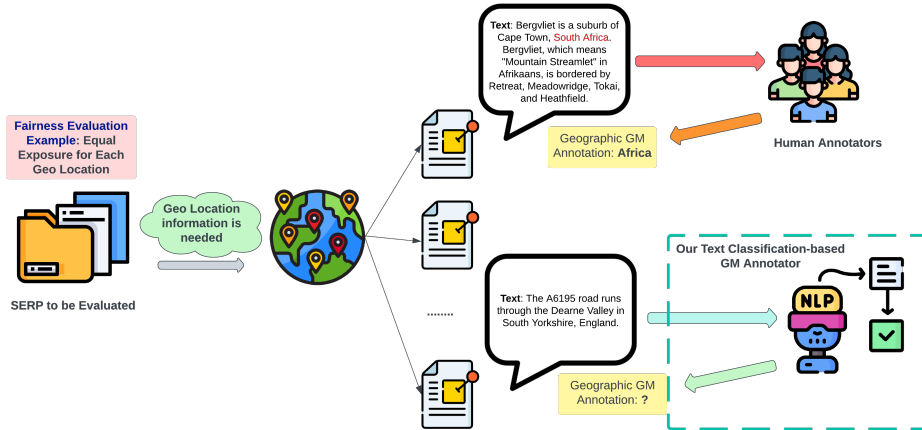


Fig. 1: The necessity of GM annotation in group fairness evaluation

used exposure-based fairness evaluation measuring the difference between ranking  $L$ 's cumulative exposure  $\epsilon_L$  and population estimator  $\hat{\epsilon}$ :

$$\text{AWRF}(L) = \Delta(\epsilon_L, \hat{\epsilon})$$

where  $\Delta$  is a divergence function (e.g., Kullback–Leibler divergence or Jensen–Shannon divergence). The ranking  $L$ 's cumulative exposure  $\epsilon_L$  is computed by  $\sum_{d \in L} w(L) * GM_d$  where  $w(L)$  is an attention decay function and  $GM_d$  is the group membership matrix of document  $d$ . For instance,  $GM_d(\text{Gender}) = (1, 0, 0)$  if the document  $d$  is annotated as group “male” for fairness category gender with three subgroups: “male”, “female”, and “non-binary”. The population estimator  $\hat{\epsilon}$  reflects the target exposure distribution that a fair system should produce, which could also rely on GM annotation. TREC estimates  $\hat{\epsilon}$  by averaging the group membership of all relevant documents to ensure that each group of items receives the same amount of expected exposure as their relevance grade. Moreover, the target exposure distribution can also be given. For example, the NTCIR fairweb1 task [30] assumes a uniform distribution across groups as their target.

### 3.2 Challenges with GM annotation

Obtaining GM annotation can be challenging and requires significant human effort. We investigate three recent fair-ranking tasks: (1) TREC fair ranking track 2021 [10], (2) TREC fair ranking track 2022 [11], and (3) the NTCIR Fair Web task [30]. Details about these tasks are reported in Table 1. TREC fair ranking tasks are based on a Wikipedia corpus containing more than six million English articles and 50/50 training and evaluating queries from various domains, whereas NCTIR fairweb1 is based on an English document collection, Chuweb-21D, containing more than 40 million documents, including research papers, movies, and YouTube Content. Unlike many previous fair-ranking studies based on outdated datasets that only contain numeric features, all three tasks provide full-text fields and enable us to apply NLP techniques. They also offer page meta information consisting of human annotations. Fig. 2 shows the subgroups'

Table 1: Task description and fairness categories: Internal fairness categories are internal attributes which do not require human annotation. We focus on contextual fairness categories.

Dataset	Task Description	Fairness Categories	
		Contextual	Internal
TREC fair ranking track 2021	A Wikipedia article fair ranking task (corpus containing more than 6 million articles): provide fair exposure for each group of documents regarding different fairness categories.	(1) Gender of article’s subject (Gender, 4 subgroups) (2) Geographical location associated with the article (Geo, 8 subgroups)	N/A
TREC fair ranking track 2022		(1) Gender of article’s subject (Gender, 4 subgroups) (2) Geographical location associated with the article (Geo, 21 subgroups)	(1) Age of the article (2) Occupation (3) Alphabetical orders (4) Popularity (5) Replication in other languages
NTCIR fairweb1	A fair ranking tasks (corpus Chuwweb21D containing about 50 million documents): provide group-fair results for research, movie, and YouTube content.	(1) Movie’s country of origin (Movie-Origin, 8 subgroups) (2) Gender of researcher (Research-Gender, 3 subgroups)	(1) Research-Hindex (2) Movie-Ratings (3) YouTube-Subscription

frequency of human annotation by page geographic locations in the TREC 2022 datasets. As can be seen, the documents’ geographic information was annotated into 21 subgroups, and a huge imbalance exists across groups. Almost half of the documents were marked as “unknown” because they either lacked annotation or were non-applicable. Ensuring a high-quality annotation is challenging, given inevitable human error and costly knowledge training for human annotators, let alone annotating GM into large numbers of subgroups. Given these challenges, few datasets with GM annotation are available for fairness-aware studies. Therefore, we aim to automate GM annotation with minimal human efforts to break the data sparsity using NLP techniques.

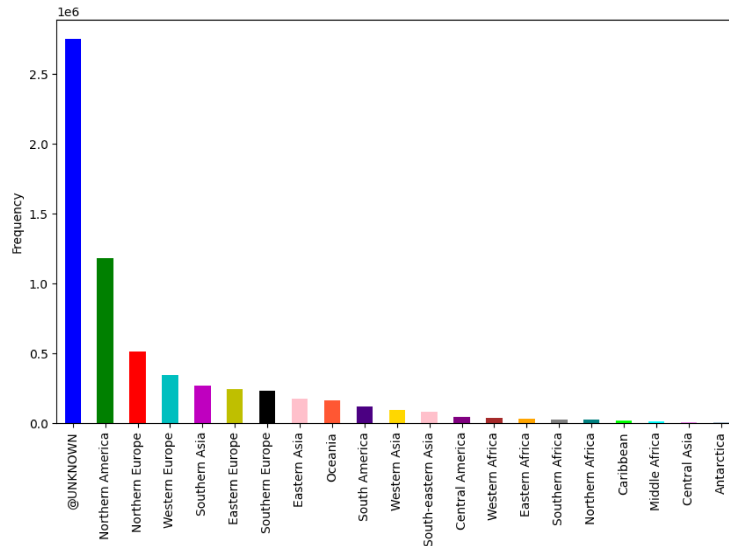


Fig. 2: Geo subgroup frequency of human GM annotation (TREC 2022).

### 3.3 Annotating GM by Text Classification with Language Models

The quality of human annotation heavily depends on the annotators’ knowledge and interpretation of the raw text. The annotation process is similar to text classification algorithms that capture contextual patterns (interpretation of raw text) and categorize raw text based on training data (knowledge). Accordingly, we assume that replacing human annotation with text classification models is possible by adequately utilizing text information, especially with sophisticated language models that can precisely capture linguistic and semantic information and even outperform humans in some studies. In this work, we explored the following text classification models for GM annotation:

- **Linear BoW Model:** a linear bag-of-words model [35] followed by a neural network classifier implemented by *spaCy TextCategorizer*<sup>2</sup>.
- **BERT-based Model:** a fine-tuned BERT sentence classification model [6] “bert-large-uncased”<sup>3</sup> implemented by *PyTorch*<sup>4</sup>.
- **GPT Models:** a generative large language model, with GPT-3.5-turbo and GPT-4, implemented by *spaCy-LLM*<sup>5</sup>.
- **Mistral 7B Models [17]:** a generative large language model, Mistral-7B-Instruct-V0.2<sup>6</sup>, implemented by *PyTorch*.

Linear bag-of-words (BoW) model [35] is one of the simplest statistical language models (SLM) that convert words to numeric representations based on vocabulary set and word count. It is flexible and performs well for simple document classification tasks, but it cannot understand context. Bidirectional Encoder Representations from Transformers (BERT) [6], introduced in 2018, is one of the masked language models (MLM) that can successfully capture semantic and linguistic information from text sequences, which has dominated classification tasks since being introduced. In contrast, generative large language models, such as OpenAI GPT, are not designed for classification tasks but have shown potential in assisting human annotations in recent studies [13]. Unlike GPT, which is fully commercialized and expensive, Mistral 7B [17] is a state-of-the-art, open-sourced LLM that achieved promising performance across various benchmark tasks. The performance of generative LLMs varies task by task and is heavily dependent on their pre-trained data and prompt design. Given the advantages and limitations of these language models, we would like to explore their capability of replacing GM annotation for group fairness evaluations.

We build classification models trained or fine-tuned by small-size human-annotated samples using these language models for GM annotation. For each subgroup within a fairness category, we equally sampled 500 training documents and 100 testing documents from each group. We follow the standard data cleaning process for the text field, including special character removal, stop word removal, and lemmatization. Given the average length of Wikipedia articles,

<sup>2</sup> <https://spacy.io/api/textcategorizer>

<sup>3</sup> <https://huggingface.co/bert-large-uncased>

<sup>4</sup> <https://pytorch.org/>

<sup>5</sup> <https://spacy.io/usage/large-language-models>

<sup>6</sup> <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2>

Table 2: LLM prompts. Shaded text is optional for one-shot or fine-tuning.

Model	Prompt
<b>GPT-3.5-turbo/GPT-4</b>	<p>You are an expert Text Classification system.</p> <p>Your task is to accept text as input and provide a category for the text based on the pre-defined labels. Classify the text below to any of the following labels: [GM Labels] Below are some examples (only use these as a guide): [Example Text], [Answer].</p> <p>Here is the text that needs classification: [Text]</p>
<b>Mistral</b>	<p>[INST]Analyze the [Fairness Category] of the Wikipedia article enclosed in square brackets, determine if it is [GM Labels], and return the answer as the corresponding labels [/INST]</p> <p>[Example Text] = [Answer]</p>

670 tokens, we truncated the full-text field to 512 tokens without losing much information. The optimal model weights are trained or fine-tuned on training samples and obtained by minimizing a KL-divergence classification loss for the linear-bag-of-words and BERT-based models. For the GPT models, we use the prompt shown in Table 2 provided by *spaCy* and set the template to 0.3 for one-shot text classification. To use the Mistral models (*Mistral-7B-Instruct-V0.2*), we utilized low-rank adaption [15] so that we can computationally run the model with our best GPU. The prompt used for Mistral is also reported in 2. Given the high cost of fine-tuning GPT models, we only fine-tuned the Mistral 7B in this study. Finally, we use these text classifiers as annotators to predict the group membership information of new documents. Once fairness annotations are obtained, ideally, we can fit them into any group fairness evaluation metrics or augment other IR datasets for fairness-aware studies.

## 4 Evaluation and Analysis

### 4.1 Prediction Accuracy of GM Annotation Models

We first examine the annotation accuracy between these annotation models when predicting gender GM annotation. The performance of each classifier is reported in Table 3. As can be seen, generative models (LLMs) failed to outperform the discriminative models (BERT and BoW models), especially for the gender subgroup “non-binary.” This might be because LLMs were pre-trained on biased data where the subgroup “non-binary” was under-represented, which is currently a known issue [21]. If we do not want to amplify this pre-existing bias, fine-tuning LLM-based models is required. Given the long text length and large corpus size (e.g., about 6 million for the TREC fair ranking track) to annotate, fine-tuning GPT models and generating GM annotations for new documents would be extremely expensive. The total price of annotating datasets with a size similar to the TREC corpus is over \$2000 even with the cheapest GPT-3.5-turbo model, as shown in Fig. 3. With the open-sourced LLM, Mistral, its fine-tuned models still cannot correctly predict the label of “non-binary”, including using different fine-tuning strategies to improve its performance as shown in the last four rows

Table 3: Classification performance (accuracy and f-1 scores) by different models when predicting “Gender” group membership: “male”, “female”, “non-binary” and “unknown” (TREC 2022). The Mistral model is fine-tuned with a full-, partial-, and proportional- set of training examples. \* indicates the best-performed model.

Models	Overall Accuracy	Overall F-1	Male F-1	Female F-1	NB F-1	Unknown F-1
Linear-BoW	0.905	0.9073	0.8475	0.9800	0.8889	0.9130
BERT*	0.985*	0.9850*	0.9804*	0.9899*	0.9697*	1*
GPT-3.5-turbo	0.820	0.7947	0.8727	0.9009	0.5507	0.8545
GPT-4	0.865	0.8549	0.8403	0.9259	0.6842	0.9691
Mistral (zero-)	0.655	0.6446	0.6076	0.6565	0.5915	0.7227
Mistral (full-)	0.705	0.6921	0.7912	0.7458	0.5135	0.7179
Mistral (part-)	0.425	0.3564	0.6619	0.1515	0.4754	0.1370
Mistral (prop-)	0.655	0.6624	0.8211	0.5686	0.5487	0.7111

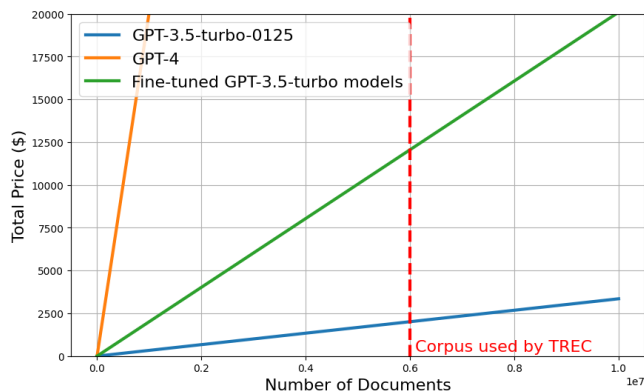


Fig. 3: Total price of annotation using trained GPT models (GPT-4, GPT-3.5-turbo, and fine-tuned GPT-3.5) by number of documents.

in Table 3. Since Mistral has difficulty to predict subgroup “Male” and “non-binary” correctly, we first fine-tuned Mistral with “Male” and “non-binary” only but as shown in the Table 3, we seem to have over-corrected the model. It is also the case when we fine-tuned the model with more “Male” and “non-binary” than the other two groups. In either case, we damage the performance of Mistral compared with the equally and fully sampled fine-tuning. The performance of GPT and Mistral shows the disadvantage of LLMs for classification tasks. Therefore, in terms of using text classification for fairness GM annotation, BERT-based models outperformed LLMs, both economically and computationally.

The fine-tuned BERT-based models demonstrated a promising annotation capability and achieved the highest accuracy and f-1 scores among all models when predicting the GM annotations (It is also true for all contextual fairness categories; we only show the result for gender here to save space). This shows the advantages of the BERT sentence classification model in terms of text under-



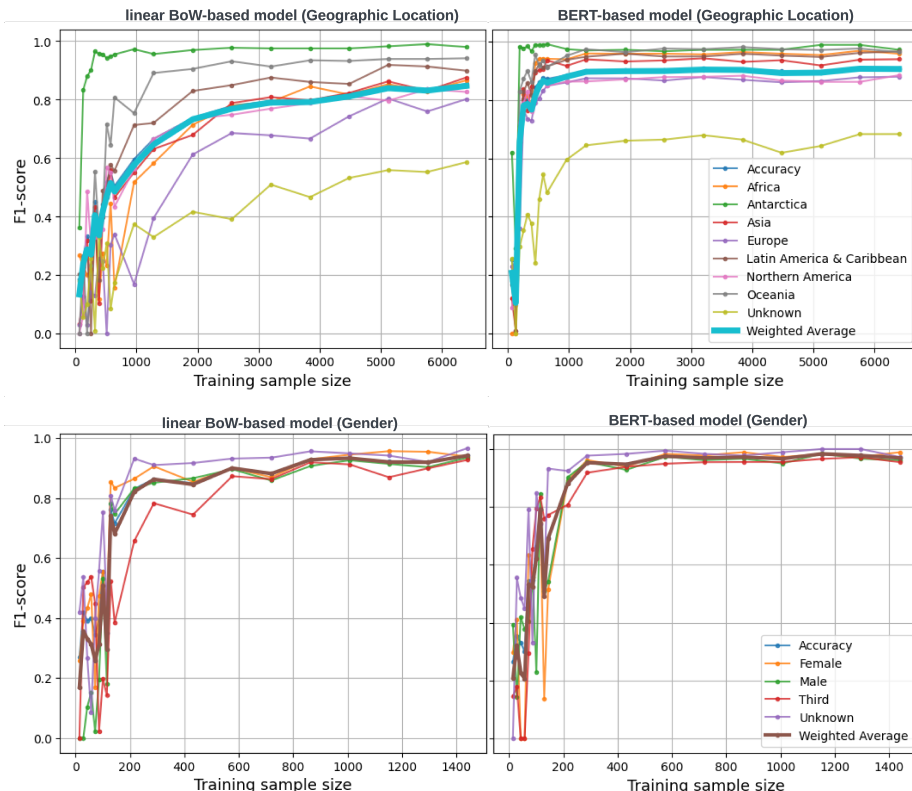


Fig. 4: Classification performance by training sample size (TREC 2021).

standing and capturing linguistic and semantic information compared with the bag-of-words models. For both BoW-based and BERT-based models, training sample size impacts the classification performance. Fig. 4 shows the text classification performance by sample size (reported as F1 scores) using both models to predict GM annotation for the TREC fair ranking track 2021. As can be seen, the linear BoW-based model requires more training samples to converge to the best performance than the BERT-based model, especially when the fairness category contains more subgroups. As shown in Fig. 5, our results regarding geographic location GM also align with previous studies that BERT-based classifiers are less sensitive to imbalanced classes [20]. As a pre-trained model, BERT only needs a few samples to fine-tune. Compared with the size of the entire corpus, we need approximately 1200 training samples when predicting geographic location GM (8 subgroups), and 400 samples when predicting gender GM (4 subgroups) to achieve a reasonable performance using BERT sentence classification. This observation also suggests that more training samples are needed, given a fairness category with more sub-groups. Generally speaking, we recommend using no less than 100-150 training samples per subgroup when training a BERT-based model for GM annotation, depending on the number of subgroups. We also noticed that both BERT and BoW models have difficulties in predicting “unknown” for geo-

graphic location GM, as shown in Fig. 5. This might result from the complexity of “unknown”, which indicates either missing annotation or non-applicable. For instance, annotating geographic locations for a mathematical proof article is not very applicable.

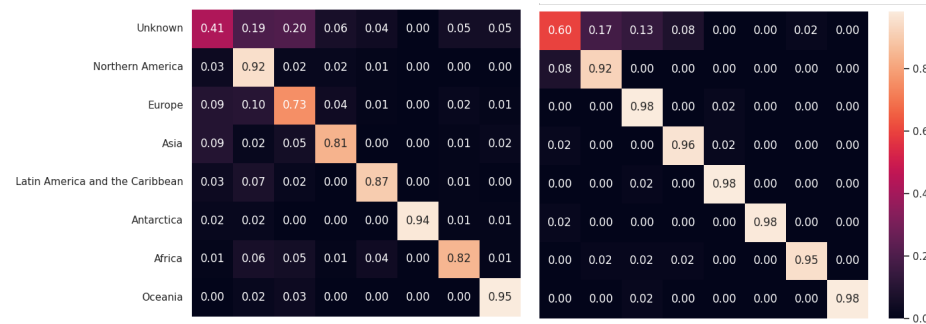


Fig. 5: BERT model (right) outperformed and is less sensitive to imbalanced classes than the bag-of-words model (left) (TREC 2021).

Overall, the BERT sentence classification model is a winner based on the above comparison, given its strong performance, simplicity, and low cost. In the following sections, we explore the impact of using our best classifier to replace human GM annotation.

## 4.2 Group Fairness Evaluation with GM annotations

We showed promising text classification performance when annotating new documents using the BERT sentence classification model. Even though annotation errors still exist, we are curious about the impact of these document-level mistakes and whether these mistakes will be washed out when aggregated into aggregated-level evaluation [1], since fairness evaluation is an aggregated metric. To test the effectiveness of the BERT-based GM annotation, we use the Pearson correlation coefficient test [5] and Spearman’s rank correlation coefficient [28] to test whether evaluation metrics with our GM annotation method can effectively differentiate the fairness quality of different systems (rankings) as the old evaluation metrics can do with human annotation. Specifically, we investigate the correlation between the official evaluation metrics with human GM annotation and those with our BERT-based GM annotation. The investigation is based on all participants’ official submissions to three fair-ranking tasks. Because the official submissions are from multiple groups using different ranking algorithms, we believe the fairness scores of these runs provide the best estimation of the upper and lower bound of fairness performance. There are 13 runs for the TREC fair ranking track 2021, 27 runs for the TREC fair ranking track 2022, and 28 runs for the NTCIR fairweb1 task.

**System-level Evaluation** Our system-level evaluation is based on testing the correlation between metrics using human annotation and metrics using BERT-based annotation to see whether we can effectively replace human annotation

Table 4: Summary of correlation tests between tasks’ official evaluation metrics with human annotation and those with BERT-based GM annotation. \* indicates statistical significance ( $p < 0.05$ ). The “Overall” group for TREC tasks is the intersectional group of Gender and Geographic Location.

	TREC 2021			TREC 2022			NTCIR fairweb1	
	Overall	Gender	Geo	Overall	Gender	Geo	M-Orgin	R-Gender
<b>Pearson</b>	0.9469*	0.9994*	0.9790*	0.9678*	0.9957*	0.9968*	0.9868*	0.9937*
<b>Spearman</b>	0.8187*	0.9945*	0.9231*	0.9609*	0.9670*	0.9976*	0.9189*	0.9688*

while preserving the ability to differentiate ranking fairness. Table 4 reports the correlation between tasks’ official evaluation metrics and our text classification-based evaluation metrics regarding the three fair ranking tasks: TREC fair ranking track 2021, TREC fair ranking track 2022, and NTCIR fairweb1. Based on Pearson correlation and Spearman’s ranked correlation tests, evaluation metrics with our BERT-based GM annotation strongly correlated with the official evaluation metrics with human annotation, and the correlations are statistically significant. This confirms the system-level effectiveness of using BERT classification-based GM annotation in evaluating fairness. Replacing human GM annotation with BERT-based annotation preserves the ability to differentiate fairness among different runs. Even though our text classifier cannot accurately predict some subgroups of some fairness categories (e.g., the group “unknown” of geographic location), when aggregating documents into a system-level evaluation, we can still differentiate rankings’ fairness. Since how to deal with “unknown” is also a challenge for human annotators, this observation also suggests that minimal annotation error will not degrade system-level fairness evaluation, and the BERT-based annotation could be a solution for estimating “unknown”.

**Query-level Robustness.** The query-level evaluation decomposes the system-level evaluation by 50 evaluation queries. In Fig. 6, we show the Pearson correlation coefficient  $r$  between human annotation-based evaluation metrics and those using BERT-based GM annotation by the evaluation query IDs. As can be seen, for most of the queries, the correlation is high and significant. That is, we highly preserve the ability to differentiate fairness when replacing human GM annotation with BERT-based GM annotation, especially for GM of “gender”. The “Overall” group, which is the intersectional product of “geographic location” and “gender” also demonstrates a high correlation between human GM annotation and BERT-based annotation. Recall the Fig. 4, predicting the GM of geographic location is less accurate than predicting the GM of Gender. Therefore, with a higher accuracy of the GM annotation, we observed a more robust query-level correlation. Therefore, to be more confident in replacing human-annotated GM and evaluating at a query level, we need a text classifier that can accurately predict GM at a document level.

### 4.3 Impact of the Annotation Accuracy

So far, we know that to preserve the ability to differentiate the fairness of different systems, we need text classifier to be accurate, and minimal annotation errors will not degrade the ability. However, what are the impacts if the classifiers

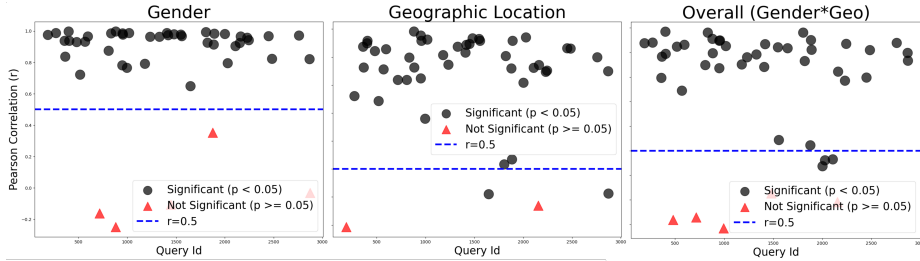


Fig. 6: Query-level robustness (TREC 2022): the correlation between evaluations using human annotation and BERT-based annotation.

are not very precise, and how accurate do we need to be confident when replacing human annotations? To further explore the relationship between annotation accuracy and the corresponding evaluation correlation with the official fairness evaluation metrics, we would like to see the impact of annotation accuracy on the group fairness evaluation metrics. The first step is to obtain annotators with different annotation accuracy. According to Fig. 4, varying training sample size is one of the easiest ways to get different annotation models with different annotation accuracy. Hence, we trained the BERT-based annotation models with varying sample sizes and obtained several annotation models with different accuracy. The relation between the annotation accuracy and the effectiveness (Pearson  $r$ ) of replacing human annotation with BERT-based annotation is plotted in Fig. 7. As can be seen, generally, increasing annotation accuracy can not only improve system-level correlation to the official metrics but also improve query-level robustness. With an annotation accuracy above 0.8, using BERT-based annotation highly preserved the ability to differentiate fairness among different systems. Therefore, if group fairness evaluation focuses on the system level, minimal annotation errors could be ignored to save human efforts.

#### 4.4 Generalizability of System Evaluation

GM annotation models can also be used for other complex evaluation metrics, such as evaluating a sequence of rankings. For example, the second task of the TREC fair ranking track 2021 evaluates fairness of sequence of rankings  $\mathcal{L}_q$  by the expected exposure loss ( $\text{EE-L}(\mathcal{L}_q) = \|\gamma - \gamma^*\|$ ) [7], expected exposure disparity ( $\text{EE-D} = \|\gamma^*\|_2^2$ ), and expected exposure relevance ( $\text{EE-R} = 2\gamma_\pi^T \gamma^*$ ).

With GM annotation obtained from our BERT-based model, we compute the correlations between the TREC fair ranking track 2021 task 2’s official evaluation metrics and those based on our annotation model based on 11 official submitted runs. We achieved Pearson correlation coefficients of 0.75, 0.98, and 0.79 for EE-L, EE-D, and EE-R, respectively, and all coefficient tests are statistically significant. Since the EE-D measures the inequality in exposure distribution across groups, and the BERT-based model has a similar performance across different groups for gender and geographic locations, we achieved a higher correlation with EE-D than the EE-R, which measures the agreement between exposure and relevance. Given the high correlation coefficients, especially for the

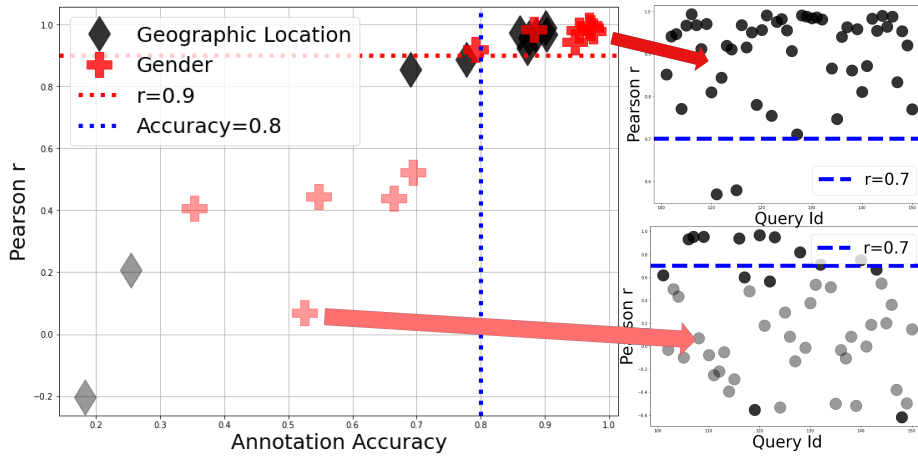


Fig. 7: Annotation accuracy V.S. the correlation between evaluation metrics using human annotation and BERT-based annotation (TREC 2021). Shaded dots indicate  $p > 0.05$ . Two samples from the left plot with different annotation accuracy are selected to show query-level robustness.

pure fairness measure EE-D, our annotation model can also effectively replace human annotation for these fairness evaluation metrics.

## 5 Conclusion

Group membership, as one of the indispensable components in group fairness evaluation, requires massive human efforts to obtain. The sparsity of GM annotations limits the application of fairness evaluation and impedes fair ranking studies on general IR datasets. To overcome this, we compared four different language model-based text classifications for GM annotation. The BERT-based model achieved promising annotation accuracy with small-size training samples and less computational cost. Our query- and system-level evaluations confirmed the effectiveness and robustness of replacing human GM annotation with the BERT-based GM annotation. This opens a new direction to augment existing IR datasets for fairness evaluation and future fair-ranking studies. Even though LLMs have been used for mainstream NLP tasks and achieved impressive performance, they failed to outperform BERT for fairness GM annotation tasks as they were not designed for discriminative tasks. Moreover, according to our impact-oriented evaluation, when replacing human annotation with different annotators that have different annotation accuracy, minimal annotation errors will not degrade the fairness evaluation metrics. In the future, we would like to utilize the new annotation strategy to augment existing IR datasets for fairness studies, including fairness evaluation and fair ranking algorithms.

**Acknowledgments.** This study is supported by the IFSA at the University of Delaware. We would like to thank the reviewers for their invaluable comments and suggestions.

## References

1. Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A.P., Yilmaz, E.: Relevance assessment: are judges exchangeable and does it matter. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 667–674 (2008)
2. Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E.H., et al.: Fairness in recommendation ranking through pairwise comparisons. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2212–2220 (2019)
3. Chae, Y., Davidson, T.: Large language models for text classification: From zero-shot learning to fine-tuning. Open Science Foundation (2023)
4. Chen, F., Fang, H.: Learn to be fair without labels: A distribution-based learning framework for fair ranking. In: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval. pp. 23–32 (2023)
5. Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. Noise reduction in speech processing pp. 1–4 (2009)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Diaz, F., Mitra, B., Ekstrand, M.D., Biega, A.J., Carterette, B.: Evaluating stochastic rankings with expected exposure. In: Proceedings of the 29th ACM international conference on information & knowledge management. pp. 275–284 (2020)
8. Ding, B., Qin, C., Liu, L., Bing, L., Joty, S., Li, B.: Is gpt-3 a good data annotator? arXiv preprint arXiv:2212.10450 (2022)
9. Ekstrand, M.D., Burke, R., Diaz, F.: Fairness and discrimination in retrieval and recommendation. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1403–1404 (2019)
10. Ekstrand, M.D., McDonald, G., Raj, A., Johnson, I.: Overview of the trec 2021 fair ranking track. In: The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings (2022)
11. Ekstrand, M.D., McDonald, G., Raj, A., Johnson, I.: Overview of the trec 2022 fair ranking track. arXiv preprint arXiv:2302.05558 (2023)
12. Gao, R., Ge, Y., Shah, C.: Fair: Fairness-aware information retrieval evaluation. *Journal of the Association for Information Science and Technology* **73**(10), 1461–1473 (2022)
13. Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., Nguyen, L.H., Hao, X., Jaber, B., Reddy, S., Kartha, R., et al.: Llms accelerate annotation for medical information extraction. In: Machine Learning for Health (ML4H). pp. 82–100. PMLR (2023)
14. He, X., Lin, Z., Gong, Y., Jin, A., Zhang, H., Lin, C., Jiao, J., Yiu, S.M., Duan, N., Chen, W., et al.: Annollm: Making large language models to be better crowdsourced annotators. arXiv preprint arXiv:2303.16854 (2023)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
16. Ishita, E., Fukuda, S., Tomiura, Y., Oard, D.W.: Using text classification to improve annotation quality by improving annotator consistency. *Proceedings of the Association for Information Science and Technology* **57**(1), e301 (2020)

17. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
18. Kasthuriarachchy, B., Chetty, M., Shatte, A., Walls, D.: Cost effective annotation framework using zero-shot text classification. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2021)
19. Koroteev, M.: Bert: a review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943 (2021)
20. Laurer, M., van Atteveldt, W., Casas, A., Welbers, K.: Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis* pp. 1–33 (2022)
21. Lucy, L., Bamman, D.: Gender and representation bias in gpt-3 generated stories. In: Proceedings of the Third Workshop on Narrative Understanding. pp. 48–55 (2021)
22. Ma, C., Shen, A., Yoshikawa, H., Iwakura, T., Beck, D., Baldwin, T.: On the effectiveness of images in multi-modal text classification: An annotation study. *ACM Transactions on Asian and Low-Resource Language Information Processing* **22**(3), 1–19 (2023)
23. Narasimhan, H., Cotter, A., Gupta, M., Wang, S.: Pairwise fairness for ranking and regression. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 5248–5255 (2020)
24. Pangakis, N., Wolken, S., Fasching, N.: Automated annotation with generative ai requires validation. arXiv preprint arXiv:2306.00176 (2023)
25. Raj, A., Ekstrand, M.D.: Comparing fair ranking metrics. arXiv preprint arXiv:2009.01311 (2020)
26. Ray, P.P.: Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* (2023)
27. Sapiezynski, P., Zeng, W., E Robertson, R., Mislove, A., Wilson, C.: Quantifying the impact of user attention on fair group representation in ranked lists. In: Companion proceedings of the 2019 world wide web conference. pp. 553–562 (2019)
28. Sedgwick, P.: Spearman’s rank correlation coefficient. *Bmj* **349** (2014)
29. Singh, A., Joachims, T.: Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2219–2228 (2018)
30. Tao, S., Sakai, T., Chen, N., Chu, Z., Arai, H., Soboroff, I., Ferro, N., Maistro, M.: Overview of the ntcir-17 fairweb-1 task. *Proceedings of NTCIR-17*. to appear (2023)
31. Thomas, P., Spielman, S., Craswell, N., Mitra, B.: Large language models can accurately predict searcher preferences. arXiv preprint arXiv:2309.10621 (2023)
32. Zehlike, M., Castillo, C.: Reducing disparate exposure in ranking: A learning to rank approach. In: Proceedings of the web conference 2020. pp. 2849–2855 (2020)
33. Zehlike, M., Yang, K., Stoyanovich, J.: Fairness in ranking: A survey. arXiv preprint arXiv:2103.14000 (2021)
34. Zhang, Y., Wang, M., Ren, C., Li, Q., Tiwari, P., Wang, B., Qin, J.: Pushing the limit of llm capacity for text classification. arXiv preprint arXiv:2402.07470 (2024)
35. Zhang, Y., Jin, R., Zhou, Z.H.: Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics* **1**, 43–52 (2010)