Fairness and Bias in Recommendation Systems

How effective are current fairness intervention methods in addressing unfairness in recommendation systems, and what trade-offs do they introduce in terms of accuracy?

Jiaqing Huang[1]

Supervisor: Masoud Mansoury[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Jiaqing Huang
Final project course: CSE3000 Research Project
Thesis committee: Masoud Mansoury, Nergis Tömen

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

As important tools for information filtering, recommendation systems have greatly improved the efficiency of users' access to information in daily life by providing personalized suggestions. However, as people's reliance on it grows, recent studies have gradually revealed their potential risks of social unfairness, such as gender discrimination that may result from job recommendations. The unfairness not only harms the interests of specific individuals or groups but also threatens the credibility and long-term sustainability of systems. Therefore, building fairness-aware recommendation systems that proactively identify and mitigate unfairness is crucial for achieving responsible recommendation services. This study focuses on systematically evaluating the effectiveness of current fairness intervention strategies. Specifically, preprocessing methods (including data relabeling and resampling) and post-processing methods (including re-ranking, calibration, and equity of attention) are selected and implemented on the two datasets MovieLens-1M and Lastfm-NL, then comprehensively evaluated in terms of two types of metrics: accuracy and fairness. The experimental results show that different methods are effective in improving different fairness targets, with varying degrees of accuracy loss or gain. This paper further explores the trade-offs between maintaining accuracy and improving fairness on intervention methods, and proposes future improvement directions for fairness-aware recommendation systems in light of the experimental results.

1 Introduction

With the rapid expansion of information, people face an overwhelming volume of content, which exceeds their ability to process it effectively, known as information overload [1]. Recommendation systems (RS) have become important tools for handling this issue by helping users navigate the vast digital landscape [2]. From e-commerce and streaming platforms to social networks and personalized services, RS play an important role in many aspects of everyday digital life.

Traditionally, RS mainly focus on utility-based metrics like click-through rate and dwell time. However, this pursuit of optimizing accuracy often raises other issues, such as diversity, privacy, and especially fairness concerns [3]. As systems for allocating attention and exposure, RS essentially influence which users receive access to which content, thus shaping people's digital experience and resource allocation [3]. This has raised concerns about unequal treatment across demographic groups and decreased visibility of niche or underrepresented content[4]. These disparities often stem from biased training data and can be further amplified by recommendation algorithms, leading to reinforced social inequality, decreased user trust, and content diversity [4].

Therefore, the study of fairness in RS is not only a technical challenge but also a social imperative [3]. Fairness has long been regarded as a basic ethical principle and the core of modern legal frameworks, designed to prevent discrimination based on sensitive attributes such as gender, race, and age [5]. From a user's perspective, fair RS ensure that all users have access to high-quality and high-diversity recommendations. From an item's perspective, fair RS can improve the exposure of long-tail content and support niche or minority creators [6]. On a broader level, promoting fairness can improve system sustainability by encouraging a diverse and active user and content base, ultimately realizing long-term sustainability of recommendation platforms [7].

Although multiple fairness-aware recommendation methods have been proposed [3], there are still some open questions about the comparative effectiveness of these methods under different evaluation metrics, and the degree of trade-offs on accuracy and fairness they bring. Based on these gaps, this paper focuses on the following three research questions:

- *RQ1*: How do the current fairness intervention methods affect accuracy and fairness in RS, respectively?
- *RQ2*: What trade-offs exist between accuracy and fairness when applying these methods to real-world datasets?
- *RQ3*: Which type of intervention method achieves the best overall balance between fairness and accuracy?

In this paper, five representative fairness interventions (including two pre-processing and three post-processing methods) are selected and empirically explored using two publicly available real-world datasets: the MovieLens-1M dataset [8] for movie recommendations and the Lastfm-NL dataset [9] for music recommendations. The performance of the methods in terms of the trade-off between accuracy and fairness under multiple metrics is systematically evaluated through a comprehensive quantitative analysis. The metrics used cover user-side and item-side fairness metrics, and accuracy metrics commonly used in RS.

The paper is structured as follows: Chapter 2 introduces the background of fairness in RS; Chapter 3 describes the methodological framework, including the selected intervention methods, evaluation metrics, datasets, and experimental setup; Chapter 4 reports the experimental results; Chapter 5 discusses ethical considerations and reproducibility; Chapter 6 analyzes the main findings and implications; and Chapter 7 presents conclusions and directions for future research.

2 Background

2.1 Recommendation systems

A recommendation system is an information filtering tool that predicts user interest in items based on available data, such as user profiles, item characteristics, and user-item interaction history, and then provides personalized suggestions accordingly [2]. For example, we define a user set $U = \{u_1, u_2, \ldots, u_m\}$ and an item set $V = \{v_1, v_2, \ldots, v_n\}$, where m and n denote the number of users and items respectively. The interaction history between users and items is represented by a binary matrix $H \in \{0,1\}^{m \times n}$, where an entry h_{ij} = 1 means that user u_i interested with item v_j, and otherwise $h_{ij} = 0$. These interactions can be explicit, such as

ratings that directly express user preferences, or implicit, such as purchasing, clicking, or browsing that indirectly indicate user interests. The main task of a recommendation system is to predict a preference score $\hat{h}_{ij}$ for each user-item pair, enabling the system to generate a personalized top-k item list l_{u_i} for each user u_i [10]. To ensure clarity and consistency throughout this paper, a complete list of notations is provided in Appendix A.

RS use different recommendation models to predict user preferences [2]. *Collaborative filtering* recommendation models [11; 12] are mainly based on user-item interaction history to identify user groups with similar preferences or item groups with similar characteristics, and then recommend items that may be of interest to the target user. In contrast, *content-based* recommendation models [13; 14] use item features or metadata, such as textual descriptions or tags, to generate recommendations by matching features of content that the user historical preferred. To overcome the limitations of a single approach, *hybrid* models [15] emerged. Such models improve recommendation effectiveness in practice by integrating collaborative filtering, content-based, or other techniques. These diverse recommendation techniques are widely used in different industries based on their characteristics to improve user experience [2].

2.2 Fairness in recommendation systems

Fairness is the foundation of social construction and a core human value, widely recognized in philosophy, sociology, law and economics [16]. In the context of RS, which functions as a two-sided platform that serves both users and items, fairness refers to equitable treatment of all participants, including users who receive recommendations, and items or providers, whose content is recommended [4; 17].

User fairness refers to the equitable treatment of users, evaluated either at the individual or group level [10]. Individually, users with similar preferences, behaviors, or needs should receive comparable recommendation results. At the group level, users can be grouped based on sensitive attributes(e.g., age, gender, geography) or user behaviors, and different groups should receive comparable recommendation results. This includes consistency in metrics such as accuracy and diversity, to avoid arbitrary differences in user experiences [4]. For example, it is unfair if female or older users receive a poorer quality of recommendations than male or younger users. User attribute bias and user selection bias are common causes of user unfairness [18].

Item fairness, also known as "provider fairness", focuses on the distribution of exposure to items or the entities behind them, evaluated either at the individual or group level [10]. Individually, items of similar quality or relevance should have equal chances of being recommended. At the group level, specific item categories (e.g., genre, types) or provider groups (e.g., independent or minority creators) should not be unreasonably suppressed in terms of exposure [4]. Item fairness is especially important when platforms monetize item exposure or when underexposure leads to negative feedback loops that discourage participation of niche or smaller providers. Exposure bias and popularity bias are common causes of item unfairness [18; 19].

2.3 Intervention Methods

To solve the fairness problem that is gradually becoming apparent in RS, a variety of intervention methods have recently been proposed [3]. These methods can be categorized into three types based on the intervention stage in the recommendation pipeline [18] (as shown in Figure 1):

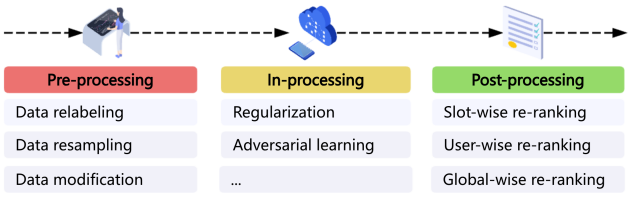


Figure 1: Three fairness intervention stages in RS Pipeline

Pre-processing methods operate at the data level and aim to reduce potential biases before model training [3]. These approaches balance the population distribution or weaken the association of sensitive attributes with predicted outcomes by modifying the training data. Typical techniques include data relabeling [20; 21], which promotes population balance by modifying sample labels; data resampling [22; 23; 24], which achieves balanced distributions by oversampling or undersampling certain groups; and data modification [25; 26], which modifies features to obscure sensitive information or reduce attribute-related disparities. These methods are model-independent and can be integrated with any downstream recommendation algorithm, making them highly adaptable for practical use.

In-processing methods intervene during the training by embedding fairness constraints directly into the models [3]. These often involve fairness-aware regularization terms[27], adversarial learning [28], or reinforcement learning frameworks [29]. Although these methods can offer fine-grained control over trade-offs between fairness and accuracy, they typically require modifying the internal mechanism of specific models, which has the limitation of their applicability across different models or domains.

Post-processing methods adjust the model outputs after training is completed. These methods focus on re-ranking the recommendation results to achieve fairness goals without changing the underlying model [3]. Examples include the FA*IR re-ranking [30], Calibration [31], and exposure-based adjustments [32]. Post-processing is attractive when access to training is restricted, as it allows fairness improvements without retraining or model changes.

This study focuses on pre-processing and post-processing methods because of their widespread applicability, ease of integration, and compatibility with black box recommendation models. The former are suitable for early development stages or when retraining is feasible, while the latter provide practical solutions for retraining existing systems that are costly or impractical. Therefore, these two type interventions provide a flexible and effective toolbox for mitigating unfairness in real-world recommendation settings, allowing interventions both before and after model deployment.

3 Methodology

This section outlines the selected fairness intervention methods, metrics, datasets, and experimental setup to evaluate the intervention effects on fairness and accuracy.

3.1 Fairness Intervention Methods

To mitigate unfairness in recommendation results, two pre-processing and three post-processing intervention methods are used in this study. These methods are selected based on their relevance to the research, effectiveness in improving fairness, and simple applicability to real-world RS.

3.1.1 Data Relabeling

Data relabeling is a pre-processing fairness intervention that modifies training labels to balance the distribution across different groups, thereby reducing bias before model training [3]. The strategy was originally proposed by Kamiran and Calders in classification [21], and then adapted to recommendation tasks. The core idea is to selectively modify ground-truth interactions, such as ratings or clicks, so that the training data does not over- or under-represent the behavior of groups that are either over- or under-represented. This is particularly relevant in datasets where specific user or item groups dominate the distribution of positive interactions.

Objective. Given a user-item ground-truth matrix $H = (h_{ij})^{m\times n}$ and a binary sensitive attribute $A = \{a_1, a_2\}$ that divides users into two disjoint groups (e.g., a_1: protected, a_2: unprotected), the aim is to reduce disparity in positive label distributions between these groups. So, we aim to construct a modified matrix $\tilde{H} = (\tilde{h}_{ij})^{m\times n}$ such that:

$$P(\tilde{h}_{ij} = 1 \mid a_{u_i} = a_1) \approx P(\tilde{h}_{ij} = 1 \mid a_{u_i} = a_2)$$

where a_{u_i} is the sensitive attribute of user u_i. Users from both groups are equally likely to have positive interactions, such as clicks, plays or high ratings represented in the dataset.

To achieve this, relabeling involves flipping some interaction labels, either demoting existing positive interactions or promoting negative ones, based on group-level statistics, to balance the distribution of positive labels across different groups while preserving overall data structure and utility. Thus, the trained model is less likely to replicate or reinforce biases presented in historical interactions.

3.1.2 Data Resampling

Data resampling is a pre-processing fairness intervention that adjusts the interaction frequency or proportion of positive interactions between different groups in the training data. The method was first proposed by Kamiran and Calders in a classification [21], and then applied in RS [22], especially when the distribution of interaction data between groups with different sensitive attributes is imbalanced.

Objective. Data resampling is similar to data relabeling in that they both attempt to reduce the distributional differences of positive interactions between different sensitive attribute groups at the training data level. However, unlike resampling, which directly modifies the labels, resampling preserves the original interaction labels.

There are two main ways of implementation: one is oversampling by replicating samples, and the other is undersampling by randomly selecting a subset based on data statistics. However, in some recommendation models, duplication of interaction data does not affect training results. Therefore, inspired by Rastegarpanah et al. [23], oversampling is achieved by adding *antidote* interactions. For example, to increase the interaction rate of female users, positive interactions between female users and items that other female users have interacted with, as well as negative interactions between female users and items interacted with mainly by male users, are randomly added.

The method is not only applicable to user-side fairness improving, but can also be extended to items to realize fair exposure on content providers.

3.1.3 FA*IR Re-ranking

FA*IR is a fairness-aware re-ranking method proposed by Zehlike et al. [30], which was initially used to process the ordered output of classification models and later adapted to RS. The core idea of it is to ensure that protected group members are proportionally represented within the top-k positions in a user's recommendation list, thus achieving fairness in the sorting result.

The approach balances ranking quality with fairness. FA*IR employs a greedy strategy: when constructing a new list of recommendations, it prioritizes the item with the highest current utility at each step, while dynamically monitoring and satisfying fairness constraints. The approach is useful in contexts where fairness in top-k recommendations is critical, such as news push, recruiting platforms, or resource distribution platforms.

Objective. Given a user u_i, let $l^b_{u_i}$ be the base top-k recommendation list generated from predicted scores $\hat{H}$. The goal is to reorder $l^b_{u_i}$ to obtain a fair list $l^f_{u_i}$ such that the number of protected items among the top-j positions satisfies a minimum fairness threshold:

$$|l^j_{u_i} \cap V_p| \geq \lfloor \alpha \cdot j \rfloor, \quad \forall j \in \{1, 2, \ldots, k\}$$

where: $V_p \subset V$ is the set of all protected items, $\alpha \in [0, 1]$ is a fairness constraint parameter to set the minimum percentage of protected items required for each location j.

This greedy re-ranking strategy incrementally builds a top-k list that ensures the presence of protected items in each prefix while preserving relevance as much as possible. The parameter α controls the strength of the fairness constraints. The higher the value, the stricter the requirement on protected group representation.

3.1.4 Calibration

Calibration-based re-ranking was first proposed by Steck [31], which aims to make the recommendation list better reflect the user's actual preferences across item categories. A calibrated RS ensures that the category distribution in the top-k recommendations closely matches the user's historical interests. This helps avoid overrepresenting or underrepresenting certain types of content, which could otherwise affect user satisfaction and fairness of the experience.

Objective. Let $p(c \mid u_i)$ be the empirical distribution of item categories $c \in C$ based on user u_i's historical interactions, and $q(c \mid u_i)$ be the distribution of item categories in the current top-k recommendation list $l^b_{u_i}$. The goal of calibration is to minimize the divergence between these two distributions, usually measured using the Kullback-Leibler (KL) divergence:

$$\mathrm{Cal}(u_i) = \mathrm{KL}(p \parallel q) = \sum_{c \in C} p(c \mid u_i) \log \frac{p(c \mid u_i)}{q(c \mid u_i)}$$

A lower calibration score indicates that the recommendation list more accurately represents the user's category-level preferences.

To achieve this, a greedy re-ranking strategy is usually used: at each step, it selects the next most relevant item, based on its predictive score $\hat{h}_{ij}$ and minimizes the calibration divergence. In this paper, we group items by gender preference (e.g., female, neutral, male preference) and use the calibration strategy to optimize the distribution of content in each gender group to improve overall fairness.

This method is suitable for reducing category-level bias and promoting a more personalized and equitable user experience.

3.1.5 Equity of Attention

The Equity of Attention framework, introduced by Biega et al. [32], focuses on item-side fairness in RS. The approach emphasizes that the exposure of an item should be proportional to its relevance. In conventional rank-based recommendation settings, this in turn leads to the aggravation of the Matthew effect of the popular items, and makes it difficult for long-tailed content to gain exposure opportunities [33].

Objective. While the original framework defines a global optimization objective to align exposure with relevance across the entire item set, we adopt an efficient greedy re-ranking strategy inspired by the core principle. In the absence of real exposure data (e.g., views or clicks), we approximate an item's exposure e_{v_j} by counting the number of times it appears in the base top-k recommendation lists L^b of all users. We construct the final recommendation list L^f in an iterative manner. At each step, an item $v_j \in l^b_{u_i}$ is selected to have the highest score:

$$\mathrm{score}(v_j \mid u_i) = \hat{h}_{ij} - \lambda \cdot \log(e_{v_j} + 1)$$

where $\hat{h}_{ij}$ is the predicted relevance score and λ is a hyperparameter to control the trade-off between relevance and exposure fairness. Items with high predicted relevance but already high exposure are penalized, encouraging the inclusion of cold items but still acceptable relevance. Each time an item is selected, its exposure count e_{v_j} increases accordingly.

This greedy re-ranking procedure promotes a more balanced exposure distribution while maintaining the quality of personalized recommendations. This approach has good scalability and practicality in offline implementation.

In summary, these five fairness intervention methods adopted in this study have different fairness goals, and we will systematically evaluate their practical effects on fairness and accuracy of recommendation results.

3.2 Evaluation Metrics

Quality evaluation is challenging due to the diverse and subjective ways users perceive recommendation quality. Therefore, existing studies typically rely on statistical indicators rather than user-reported satisfaction [34]. This paper evaluates the effectiveness of the fairness intervention methods from two dimensions: Accuracy and Fairness.

3.2.1 Accuracy Metrics

Accuracy metrics measure the performance of recommendation results. In this paper, we adopt five widely used accuracy metrics from the RecBole framework [35]: Precision, Recall, Hit Ratio, Mean Average Precision, and Normalized Discounted Cumulative Gain. These metrics measure the relevance and ranking quality of the recommendations.

3.2.2 Fairness Metrics

Fairness metrics are linked to how fairness is defined [17]. In RS, fairness is typically evaluated from two perspectives: user-side and item-side. These metrics assess how equitably recommendation results are distributed across different user groups or item categories.

User-side fairness metrics focus on whether users or user groups receive comparable recommendation quality [36]. Recommendation quality here includes both accuracy and diversity. Accuracy measures whether the recommendations match the user's interests, and diversity reflects the exposure to a wider range of content to prevent the formation of information bubbles, and to help users to broaden horizons [36]. We use the User Group Fairness (UGF) metric [36] to capture disparities between groups: UGF-NDCG measures the difference in accuracy (NDCG), while UGF-IC reflects the difference in diversity, measured by Item Coverage.

Item-side fairness evaluates how attention or exposure is distributed across items or item categories. We employ five standard item-side fairness metrics provided by RecBole [35]: Item Coverage, Average Popularity, Shannon Entropy, Gini Index, and Tail Percentage. These metrics reveal whether the recommendation model is biased toward popular items and whether it fairly recommends long-tail content.

These metrics are widely used in fairness-aware recommendation researches [3]. The definitions of each evaluation metric are summarized in Table 1, with detailed descriptions and mathematical formulas provided in Appendix B.

3.3 Datasets

In this paper, two widely used real-world datasets are selected: *MovieLens-1M*(ML-1M) and *Lastfm-NL*. The main reasons for choosing these two datasets include the fact that they both provide user-side demographic information, belong to different interaction domains, and have a broad application base in fairness recommendation research.

The ML-1M dataset [8] contains approximately 1M explicit ratings ranging from 1 to 5, provided by 6,040 users for 3,706 movies. Basic attributes such as gender, age, and occupation of users are provided, and one or more movie genre tags of items are provided.

Table 1: Evaluation metrics used in this study. Arrows (↑, ↓) indicate preference direction.

Metric[1]	Interpretation
Accuracy	
Pre@K ↑	Proportion of recommended items that are relevant
Rec@K ↑	Proportion of relevant items that are retrieved
Hit@K ↑	At least one relevant item appears in top-K
MAP@K ↑	Mean precision over relevant items
NDCG@K ↑	Rank-sensitive relevance evaluation
User Fairness	
UGF-NDCG@K ↓	Accuracy gap (NDCG) between female and male user groups
UGF-IC@K ↓	Diversity gap (IC) between female and male user groups
Item Fairness	
IC@K ↑	Fraction of unique items recommended across all users
AP@K ↓	Mean popularity of recommended items
SE@K ↑	Dispersion of item exposure across users
GI@K ↓	Inequality in item exposure
TP@K ↑	Exposure to long-tail (less popular) items

[1] Metric abbreviations: Pre = Precision, Rec = Recall, Hit = Hit Ratio, MAP = Mean Average Precision, NDCG = Normalized Discounted Cumulative Gain, UGF = User Group Fairness, IC = Item Coverage, AP = Average Popularity, SE = Shannon Entropy, GI = Gini Index, TP = Tail Percentage.

Dataset	Users	Items	Interactions	Sparsity
ML-1M	6,040	3,706	1,000,209	95.532%
Lastfm-NL	8,792	36,077	434,240	99.863%

Table 2: Basic information of the two datasets

The Lastfm-360K dataset [9] records the top 50 most played artists for 359,347 users. It contains about 17.6 million user-artist interactions covering 160,168 unique artists. User attributes such as gender, age, and country are available. Due to its large scale, direct use for training can be computationally expensive and time-consuming. Therefore, we constructed a condensed subset, which keeps only users living in the Netherlands and removes users lacking gender information. The Lastfm-NL contains 8,792 users, 36,077 artists, and 434,240 interaction pairs. The interactions are implicit and represented by play counts ranging from 1 to 56930. The dataset is highly sparse, with a sparsity rate of 99.863%.

These two datasets offer complementary features: ML-1M is more suitable for studying fairness performance in small-scale, structured and explicit feedback scenarios, while Lastfm-NL represents a large-scale, implicit feedback setting with imbalanced user behavior. Using both allows for a comprehensively assess of the applicability and validity of different fairness methods under diverse data featurs.

3.4 Experimental Setup

Baseline Scenario. The baseline represents a standard collaborative filtering setup using Bayesian Personalized Ranking (BPR) without any fairness interventions. This scenario serves as a control to quantify the degree of fairness in a typical recommendation model. We use Bayesian optimization for hyperparameter tuning. The tuning process searches for optimal learning rate, regularization coefficient, batch size, and embedding dimension. The validation set is used for early stopping strategy. The length of the recommendation list is set to $k = 10$.

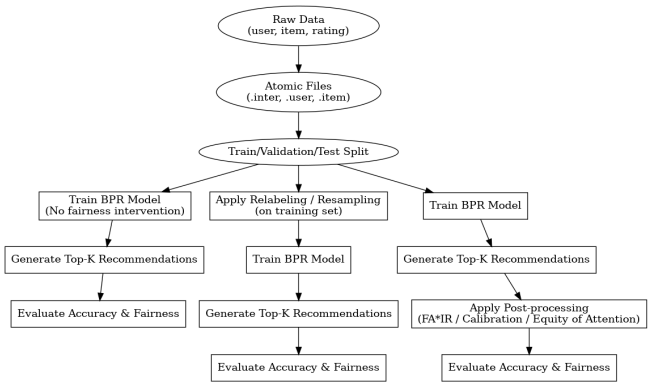


Figure 2: Experimental scenarios overview. All scenarios use the same BPR model configuration and data splits.

Pre-processing Scenario. A fairness-aware data processing strategy is applied before training, including relabeling or resampling, to mitigate bias at the data level.

Post-processing Scenario. A fairness-aware re-ranking strategy is applied to the output of the baseline model, including FA*IR, Calibration, or Equity of exposure, to improve fairness without retraining.

The datasets are split using RecBole's built-in data splitting function, configured as follows: each user's interaction history is randomly sorted and proportionally divided into 80% training, 10% validation and 10% testing set. This process is executed one by one at the user level to ensure that each user has data coverage in the training, validation and testing phases.

Only one fair intervention method is applied per round of experiments to independently evaluate its effectiveness. We measure both accuracy and fairness scores before and after the interventions to assess trade-offs between accuracy and fairness.

4 Experimental Results

This section presents our experimental results in terms of the three research questions (RQs) introduced in the introduction.

4.1 RQ1: Effects of Fairness Interventions on Accuracy and Fairness

Table 3 reports the performance of various fairness intervention methods across accuracy and fairness metrics for the ML-1M and Lastfm-NL datasets. Below, we highlight the best-performing methods for each metric.

4.1.1 Accuracy Metrics

ML-1M: *Undersample* achieves the best performance in Pre@10 (0.0565), Rec@10 (0.0800), Hit@10 (0.4361), MAP@10 (0.0331), and NDCG@10 (0.0770).

Lastfm-NL: *Oversample* shows the best results in Pre@10 (0.0350), Rec@10 (0.0797), and Hit@10 (0.3049). *Undersample* and *Relabel* achieves the best MAP@10 (0.0279). *Undersample* again has the highest NDCG@10 (0.0615).

Method	Accuracy					User Fairness						Item Fairness				
	Pre↑	Rec↑	Hit↑	MAP↑	NDCG↑	NDCG(M)↑	NDCG(F)↑	UGF-NDCG↓	IC(M)↑	IC(F)↑	UGF-IC↓	IC↑	AP↓	SE↑	GI↓	TP↑
ML-1M dataset																
Baseline	0.0547	0.0787	0.4283	0.0326	0.0756	0.0770	0.0721	0.0049	0.3527	0.2737	0.0790	0.3758	1299.8182	0.7930	0.9230	0.0002
Relabel	0.0535	0.0772	0.4210	0.0306	0.0727	0.0739	0.0696	0.0043	0.3641	0.2992	0.0649	0.3986	1276.9728	0.7976	0.9158	0.0001
Oversample	0.0542	0.0759	0.4238	0.0313	0.0736	0.0774	0.0640	0.0134	0.3082	0.2846	**0.0236**	0.3554	1358.2146	0.7800	0.9319	0.0000
Undersample	**0.0565**	**0.0800**	**0.4361**	**0.0331**	**0.0770**	**0.0784**	**0.0734**	0.0050	0.3055	0.2438	0.0616	0.3323	1392.1854	0.7740	0.9377	0.0000
FA*IR	0.0547	0.0787	0.4283	0.0326	0.0756	0.0770	0.0721	0.0049	0.3543	0.2748	0.0796	0.3785	1299.7498	0.7923	0.9230	0.0005
Calibration	0.0529	0.0762	0.4200	0.0320	0.0739	0.0752	0.0706	0.0046	0.3562	0.2740	0.0823	0.3839	1237.6068	0.7992	0.9208	0.0003
Equity	0.0454	0.0660	0.3719	0.0266	0.0626	0.0635	0.0603	**0.0032**	**0.5468**	**0.4600**	0.0869	**0.5778**	**842.5324**	**0.8975**	**0.7863**	**0.0006**
Lastfm-NL dataset																
Baseline	0.0332	0.0753	0.2902	0.0270	0.0594	0.0601	0.0565	0.0036	0.1016	0.0381	0.0635	0.1116	631.4422	0.6938	0.9888	0.0040
Relabel	0.0345	0.0787	0.2981	**0.0279**	0.0614	**0.0624**	0.0572	0.0052	0.1404	0.0490	0.0914	0.1575	599.8631	0.6912	0.9831	0.0080
Oversample	**0.0350**	**0.0797**	**0.3049**	0.0275	0.0614	0.0613	**0.0618**	**0.0005**	0.1650	0.0568	0.1082	0.1860	537.2037	0.7136	0.9780	**0.0112**
Undersample	0.0347	0.0786	0.2974	**0.0279**	**0.0615**	0.0621	0.0591	0.0030	0.1674	0.0577	0.1097	0.1895	553.2270	0.7099	0.9775	0.0108
FA*IR	0.0332	0.0753	0.2902	0.0270	0.0594	0.0601	0.0565	0.0036	0.1042	0.0386	0.0656	0.1147	631.1572	0.6922	0.9886	0.0040
Calibration	0.0312	0.0711	0.2762	0.0263	0.0571	0.0576	0.0554	0.0022	0.0920	0.0379	**0.0541**	0.1011	590.1089	0.7074	0.9893	0.0023
Equity	0.0323	0.0735	0.2804	0.0254	0.0567	0.0573	0.0545	0.0028	**0.1951**	**0.0714**	0.1237	**0.2159**	**487.5218**	**0.7500**	**0.9674**	0.0086

Table 3: Comparison of fairness intervention methods on ML-1M and Lastfm-NL. Accuracy metrics (Pre@10, Rec@10, Hit@10, MAP@10, NDCG@10) and fairness metrics (User: UGF-NDCG@10 and UGF-IC@10; Item: IC@10, AP@10, SE@10, GI@10, TP@10).

4.1.2 User Fairness Metrics

ML-1M: *Equity* has the lowest UGF-NDCG@10 (0.0032). *Oversample* achieves the lowest UGF-IC@10 (0.0236).

Lastfm-NL: *Oversample* achieves the lowest UGF-NDCG@10 (0.0005). *Calibration* has the lowest UGF-IC@10 (0.0541).

4.1.3 Item Fairness Metrics

ML-1M: *Equity* performs best across IC@10 (0.5778), AP@10 (842.5324), SE@10 (0.8975), GI@10 (0.7863), and TP@10 (0.0006).

Lastfm-NL: *Equity* again outperforms other methods in IC@10 (0.2159), AP@10 (487.5218), SE@10 (0.7500), and GI@10 (0.9674). *Oversample* achieves the best TP@10 (0.0112).

Overall, we observe that *Undersample* and *Oversample* perform well on accuracy and user fairness, while *Equity* achieves the strongest improvements in item fairness across both datasets.

4.2 RQ2: Trade-offs of Fairness Interventions Between Accuracy and Fairness

This subsection presents the observed trade-offs between recommendation accuracy and fairness introduced by different intervention methods. We analyze both user fairness and item fairness on the ML-1M and Lastfm-NL datasets.

4.2.1 User Fairness

Figure 3 (left) shows the trade-offs between accuracy (NDCG@10) and user fairness metrics (UGF-NDCG@10, UGF-IC@10). These two metrics reflect the disparity in ranking relevance and diversity between different user groups, respectively.

ML-1M: *Undersample* achieves the highest NDCG@10 while maintaining moderate fairness levels, suggesting a well-balanced outcome. *Equity* achieves the lowest UGF-NDCG@10, suggesting strong relevance fairness. However, it performs poorly in UGF-IC@10 and accuracy. In contrast, *Oversample* significantly reduces UGF-IC@10 but performs poorly in UGF-NDCG@10, though with a small drop in accuracy. *Calibration* and *Relabel* show slight improvements in relevance fairness, while *FA*IR* has similar performance compared to the baseline.

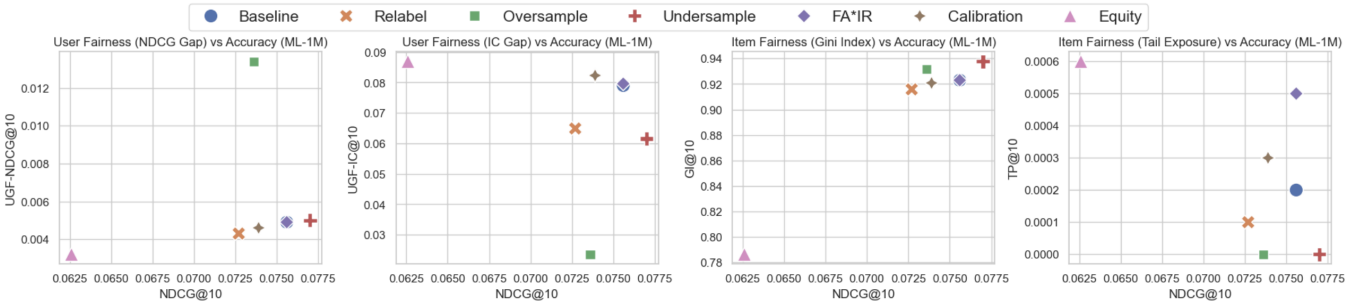
Lastfm-NL: *Oversampling* again performs well, achieving both the lowest UGF-NDCG@10 and competitive NDCG@10. However, its UGF-IC@10 remains relatively high. *Calibration* achieves the best UGF-IC@10 fairness and slightly improves UGF-NDCG@10, with some accuracy degradation. *Equity* shows improvements in relevance fairness but has the worst IC gap and lowest accuracy.

4.2.2 Item Fairness

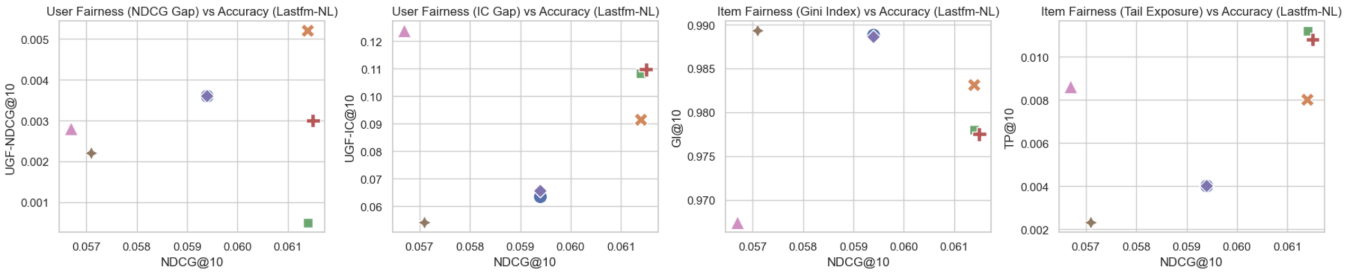
Figure 3 (right) illustrates the trade-offs between accuracy (NDCG@10) and item fairness metrics(GI@10, TP@10). These two metrics reflect the equity of exposure distribution and how well the system promotes long-tail content.

ML-1M: *Equity* achieves the lowest GI@10 and the highest TP@10, indicating highly equitable and diverse item exposure, though it also has the greatest drop in accuracy. *Calibration* and *FA*IR* strike a moderate balance, showing relatively low GI@10 and higher TP@10 values while keeping NDCG@10 close to the base. *Relabel* slightly improves GI@10 but shows limited impact on tail exposure, suggesting it does not effectively promote content diversity. Its accuracy remains below the base level. *Oversample* and *Undersample* achieves the highest GI@10 and lowest TP@10 values among all methods. This indicates their tendency to favor popular items, failing to improve exposure fairness. Although they achieve high accuracy, they do so at the cost of item fairness.

Lastfm-NL: *Equity* again achieves the lowest GI@10, indicating the most equitable item exposure distribution. It also provides one of the highest TP@10 values, showing effectiveness in recommending long-tail items. Despite this, its accuracy remains the lowest among all methods. *Oversample* and *Undersample* show strong performance in TP@10, reaching the highest tail exposure while maintaining the best NDCG@10. *FA*IR* performs similarly to the base model. *Relabel* balances both fairness metrics reasonably well. It achieves relatively high TP@10 and lower GI@10, while maintaining competitive accuracy. *Calibration* shows a decrease in fairness and accuracy, with high GI@10 and the lowest TP@10.



(a) ML-1M: User Fairness vs. Accuracy (left) & Item Fairness vs. Accuracy (right)



(b) Lastfm-NL: User Fairness vs. Accuracy (left) & Item Fairness vs. Accuracy (right)

Figure 3: Fairness(User and Item) vs. Accuracy trade-offs for ML-1M and Lastfm-NL datasets.

4.3 RQ3: Best Overall Balance Between Fairness and Accuracy

Based on the metric scores from Table 3 and the trade-off trends observed in Figure 3, we analyze which methods are most suitable under different optimization goals.

4.3.1 Accuracy-Oriented Ranking

When accuracy is the top priority, *Undersample* consistently achieves the highest NDCG@10 across both datasets. It reaches 0.0770 on ML-1M and 0.0615 on LastFM-NL, with small compromises in fairness. This makes it the most effective method when preserving ranking performance is the main goal.

*FA*IR* also maintains a high accuracy, with a small improvement on fairness, making them practical alternatives in accuracy-constrained environments.

4.3.2 User Fairness-Oriented Ranking

To reduce disparities in recommendation accuracy between user groups, *Oversample* performs best on the user-side fairness metrics. It significantly reduces UGF-IC@10 on ML-1M and UGF-NDCG@10 on Lastfm-NL datasets. These gains come with controllable losses in accuracy for ML-1M, but increase the accuracy for Lastfm-NL surprisingly.

4.3.3 Item Fairness-Oriented Ranking

From the item side, the *Equity* method is effective in improving the fairness of item exposure. In the ML-1M dataset, the method achieves the lowest GI@10 value and the highest TP@10 value, and also demonstrates good long-tail content exposure in the LastFM-NL dataset.

4.3.4 In conclusion:

- **Accuracy-focused systems** should prefer *Undersample* for its higher accuracy than the baseline, but with higher fairness fluctuations; *FA*IR* brings a small fairness improvement while maintaining accuracy.
- **User fairness-sensitive systems** benefit most from *Oversample*, which shows strong improvements in both relevance and diversity fairness.
- **Item fairness-driven systems** should consider *Equity* for effectively improving long-tail exposure and exposure balance, although some accuracy will be sacrificed.

5 Responsible Research

5.1 Ethical Considerations

This study focuses on fairness in RS, exploring five interventions, including relabeling, resampling, and fairness-aware re-ranking, to reduce disparities in recommendation quality across gender populations and in exposure of items. All experiments are conducted using publicly available benchmark datasets: **MovieLens 1M** and **Lastfm 360K**. These datasets do not contain any identifiable personal information. The demographic attribute *gender* is used only for group-level fairness evaluation, not for individual profiling. We did not introduce or manipulate any ethically sensitive or inferred attributes, and all data were used in accordance with responsible research standards.

In addition to their technical effects, fairness interventions may also have long-term implications for the platform ecosystems. For example, a fair exposure mechanism may incentivize more participation, but it may also lead to user attrition or resistance if the intervention strategy is too strict or out of user expectations. In addition, any algorithmic tool can be misused, and fairness mechanisms can even be used to mask or exacerbate bias if they are not regulated. It is

therefore important that such technologies are designed and deployed in a way that is transparent, auditable and subject to strict ethical oversight and governance.

5.2 Reproducibility

To ensure transparency and reproducibility, all aspects of the experimental pipeline are thoroughly documented and will be made publicly available.

Implementation Environment All experiments were conducted on macOS Sequoia machine. The software environment includes:

- Python 3.10
- RecBole 1.2.1
- NumPy, Pandas and Matplotlib

Code and Configuration The complete source code, including data pre-processing scripts, training pipelines, fairness intervention implementations, and evaluation modules, is organized and will be released via a public GitHub repository. The repository contains:

- Atomic file loaders and demographic pre-processing tools
- Implementations of each fairness intervention method
- Fixed YAML configuration files for reproducible training
- Precomputed logs and metric outputs

Reproducibility Support Step-by-step setup instructions are provided, including shell scripts and configuration-driven pipelines for executing all experiments. This ensures that results can be easily replicated by other researchers

These practices ensure that this research remains verifiable, reusable, and aligned with responsible standards in machine learning and AI research.

6 Discussion

This study systematically evaluates a variety of fairness-aware interventions for RS, focusing on analyzing the trade-offs between accuracy and fairness on the user side and the item side.

Balancing Accuracy and Fairness The experimental results clearly reveal that accuracy and fairness are not compatible in RS. If the system prioritizes the balance of user experience, *Oversample* significantly improves the recommendation quality difference among user groups; if the goal is to increase the exposure of long-tail content and the diversity of item distribution, *Equity* significantly reduces the popularity bias. However, both of them reduce the relevance and ranking performance of recommendations to some extent. In contrast, *Undersample* maintains a high recommendation accuracy while sacrificing little fairness. Thus, in practice, the weighting trade-offs need to be clarified according to the system requirements: whether to maximize the overall utility, or to emphasize the fairness goal of the platform, or to find an acceptable compromise between the two.

Comparison to Prior Work The results of this study are basically consistent with existing literature [30; 31; 32], especially in the effectiveness of post-processing methods (such as *Equity* and *Calibration*) in improving fairness. However, our study further highlights the potential of pre-processing methods such as *Oversample* and *Relabel* in controlling for loss of accuracy. Unlike most literatures that only focus on a few metrics or methods, this paper adopts multi-dimensional fairness evaluation and method comparison, thereby providing a more comprehensive evaluation perspective.

Dataset-Specific Observations The experiment also revealed differences in performance of intervention strategies under different datasets. In ML-1M with clear structure and rated scores, item-side interventions (such as *Equity*) perform well in improving exposure diversity; while in Lastfm-NL datasets with sparse interactions and implicit feedback, user-side approaches (such as *Oversample* and *Undersample*) improve more significantly in fairness. Interaction types, thresholds, sparsity, and demographic structure will affect the effectiveness of the intervention strategies. Therefore, in actual deployment, it is necessary to select and optimize the method according to the data characteristics.

Methodological Strengths and Limitations A key strength of our methodology is the comprehensive evaluation framework with multiple fairness and accuracy metrics. This allows insights into the performance of different intervention methods in terms of trade-offs across multiple dimensions. However, the research still has limitations. On the one hand, this paper mainly analyzes the single sensitive attribute of gender, and does not cover more complex cross-groups; on the other hand, the interventions need to be carefully tuned for each dataset application based on its statistics rather than adopting uniform hyperparameter settings.

Unexpected Findings and Open Questions An unexpected result in this study is that the *Undersample* method shows a better balance between accuracy and fairness. The method mitigates the impact of gender bias on model training by controlling the average interaction frequency and the proportion of positive feedback for users of different genders in the training set to make the male and female groups more equal in data distribution. The experimental results show that some user-side fairness metrics are improved along with a significant increase in accuracy, outperforming those intervention methods with stronger expected effects. This finding suggests that in certain scenarios with severe data imbalances, appropriately reducing the dominant group sample can instead contribute to the overall performance and fairness of the model. However, it also leads to new questions: is the strategy applicable to other datasets or have similar effect on other sensitive attributes? Is it equally effective in more complex multi-group or cross-attribute environments? These deserve further exploration in future research.

In general, fair recommendations are essentially a multi-objective optimization problem. Under the needs of different data environments and platforms, the effects and costs of fair intervention strategies vary significantly. Therefore, RS designers need to carefully weigh the accuracy and fairness, and make decisions that match the actual scenario.

7 Conclusion and Future Work

This study systematically evaluates fairness interventions in RS, on both pre-processing and post-processing strategies. By using the RecBole framework, five representative fairness methods: Relabel, Resample, FA*IR, Calibration, and Equity of Attention, are examined for their impacts on accuracy and fairness across two widely-used benchmark datasets: MovieLens 1M (explicit movie ratings) and Lastfm-NL (implicit music interactions).

Through extensive experiments, the primary contributions of this research include:

- A comprehensive comparative evaluation of a variety of fairness intervention methods in RS, from both accuracy and fairness dimensions.
- A unified experimental pipeline integrating pre-processing and post-processing fairness interventions within the RecBole framework.
- Practical insights that there is no universal method that can perform optimally in all scenarios, and the selection of fairness strategies should be weighed and customized with the characteristics of the platform, fairness goals, and the specific needs of the application scenarios.

In summary, this study shows that fairness-aware interventions can be viable, flexible, and effective tools for improving equity in RS. However, no single method consistently outperforms others across all evaluation metrics, which further highlights the inherent complexity and context-dependency of balancing accuracy and fairness.

While this study clarifies several key aspects of trade-offs between fairness and accuracy, several important issues remain open for exploration. First, integrating group-aware personalized objectives directly into model training, i.e., exploring in-processing fairness methods, could achieve more fine-grained fairness optimization, though this typically requires modifications to the model architecture. Second, hybrid approaches that combine pre-processing and post-processing strategies may achieve a better balance between fairness and accuracy. Moreover, extending fairness interventions to domains involving more complex or sensitive user attributes, such as employment, education, or healthcare, could introduce both new opportunities and important ethical considerations. On this basis, further assessment of fairness performance among cross-groups (e.g., combinations of gender and age) can help reveal structural inequalities that are difficult to detect under a single-attribute perspective and promote a more comprehensive understanding of algorithmic bias. Finally, fine-tuning the hyperparameters in fairness intervention methods is also a key direction to enhance their usefulness and stability, especially when facing real systems with different data characteristics and target demands.

Appendix

A Notations

The summary of notations used in the paper:

- $U = \{u_1, u_2, \ldots, u_m\}$: the set of m users.
- $V = \{v_1, v_2, \ldots, v_n\}$: the set of n items to be recommended.
- $H = (h_{ij})^{m \times n}$: the ground-truth interaction matrix, where h_{ij} is the true preference score of user u_i for item v_j.
- $\hat{H} = (\hat{h}_{ij})^{m \times n}$: the predicted interaction matrix, where $\hat{h}_{ij}$ is the score predicted by the recommendation model (e.g., BPR).
- $\tilde{H} = (\tilde{h}_{ij})^{m \times n}$: the interaction matrix after applying a pre-processing method, where $\tilde{h}_{ij}$ is the modified score.
- k: the length of the top-k recommendation list.
- $L^g = \{l^g_{u_1}, l^g_{u_2}, \ldots, l^g_{u_m}\}$: the set of ideal recommendation lists generated from ground-truth preferences.
- $L^b = \{l^b_{u_1}, l^b_{u_2}, \ldots, l^b_{u_m}\}$: the set of base recommendation lists generated from $\hat{H}$.
- $L^f = \{l^f_{u_1}, l^f_{u_2}, \ldots, l^f_{u_m}\}$: the set of fairness-enhanced recommendation lists obtained after applying fairness interventions.
- $l^b_{u_i}[t]$: the item ranked at position t in user u_i's base list $l^b_{u_i}$.
- $l^{b/j}_{u_i}$: the top-j sublist of user u_i's base list $l^b_{u_i}$.
- $A = \{a_1, a_2, \ldots, a_s\}$: the set of values for a protected attribute (e.g., gender or age), where each user u_i has a value $a_{u_i} \in A$.
- $C = \{c_1, c_2, \ldots, c_t\}$: the set of item categories, where each item v_j is assigned to one category $c_{v_j} \in C$.
- s_{u_i}: the performance score of user u_i based on a specified evaluation metric (e.g., NDCG@10).

B Definition of Metrics

B.1 Accuracy Metrics

Precision(Pre) Precision measures the proportion of recommended items in the top-k list that are relevant to the user. For each user u_i, Pre@k is defined as:

$$\text{Pre@}k(u_i) = \frac{|l^k_{u_i} \cap l^g_{u_i}|}{k}$$

Recall(Rec) Recall measures the proportion of relevant items that are accurately recommended in the top-k recommendation list for each user. It reflects the completeness of the recommendations. For each user u_i, Rec@k is defined as:

$$\text{Rec@}k(u_i) = \frac{|l^k_{u_i} \cap l^g_{u_i}|}{|l^g_{u_i}|}$$

Hit Rate(Hit) Hit Rate checks whether at least one relevant item appears in the top-k recommendation list. For each user u_i, Hit@k is defined as:

$$\text{Hit@}k(u_i) = \begin{cases} 1, & \text{if } l^k_{u_i} \cap l^g_{u_i} \notin \emptyset \\ 0, & \text{otherwise} \end{cases}$$

Mean Average Precision (MAP) MAP measures the ranking quality of relevant items in the top-k list. For each user u_i, we sum the precision at each rank where a relevant item appears, and normalize by the number of relevant items (or k). MAP@k is defined as:

$$\text{MAP@}k = \frac{1}{\min(|l^g_{u_i}|, k)} \sum_{t=1}^{k} I[l^b_{u_i}[t] \in l^g_{u_i}] \cdot \text{Pre@}t(u_i)$$

where $\text{Pre@}t(u_i)$ denotes the precision at cut-off t for user u_i.

Normalized Discounted Cumulative Gain(NDCG) NDCG evaluates both the relevance and ranking quality of recommendations. It gives higher weights to items appearing earlier in the ranked list.

For each user $u_i \in U$, let $l^f_{u_i}$ be the final top-k recommendation list, $l^f_{u_i}[t] = v_j$ be the item at position t, and $l^g_{u_i}$ be the ground-truth recommendation list based on ground-truth preference score h_{ij} for user u_i and item v_j. Then the user's utility for this item is defined as $\mu_{u_i}(v_j) = h_{ij}$. The Discounted Cumulative Gain (DCG) for each user u_i is:

$$\text{DCG@}k(u_i) = \sum_{t=1}^{k} \frac{\mu_{u_i}(l^f_{u_i}[t])}{\log_2(t+1)}$$

The Ideal DCG (IDCG) is computed based on $l^g_{u_i}$:

$$\text{IDCG@}k(u_i) = \sum_{t=1}^{\min(k, |l^g_{u_i}|)} \frac{\mu_{u_i}(l^g_{u_i}[t])}{\log_2(t+1)}$$

Finally, the NDCG@k for user u_i is:

$$\text{NDCG@}k(u_i) = \frac{\text{DCG@}k(u_i)}{\text{IDCG@}k(u_i)}$$

B.2 Fairness Metrics

User-oriented Group Fairness(UGF) UGF is defined as the absolute difference in average recommendation quality between the two sensitive groups. Lower UGF values indicate a more equitable distribution of recommendation accuracy or diversity across sensitive user groups. Let $A = \{a_1, a_2\}$ be a binary sensitive attribute set, where each user u_i belongs to exactly one group: either G_{a_1} or G_{a_2}, and s_{u_i} be the recommendation performance score (e.g., NDCG@k) for user u_i. UGF@k is defined as:

$$\text{UGF@k} = \left| \frac{1}{|G_{a_1}|} \sum_{u_i \in G_{a_1}} s_{u_i} - \frac{1}{|G_{a_2}|} \sum_{u_i \in G_{a_2}} s_{u_i} \right|$$

Item Coverage(IC) IC measures the proportion of unique items recommended across all users, reflecting diversity or equitable exposure of recommended items. Higher IC indicates more diverse and equitable recommendations. Let $n = |V|$ be the total number of items, IC@k is defined as:

$$\text{IC@k} = \frac{\left|\bigcup_{u_i \in U} l^f_{u_i}\right|}{n}$$

Average Popularity(AP) AP measures the average popularity of items recommended to users. Lower AP suggests better promotion of long-tail items. Let ϕ_{v_j} represent the interaction number of item v_j in the training set, AP@k is defined as:

$$\text{AP@k} = \frac{1}{|U|} \sum_{u_i \in U} \frac{\sum_{v_j \in l^f_{u_i}} \phi_{v_j}}{k}$$

Shannon Entropy(SE) SE calculates the entropy of the distribution of item occurrences, reflecting the diversity of recommended items. Higher SE indicates greater item diversity. Let ψ_{v_j} represent the occurrence number of item v_j in the top-k recommendation lists L^f of all users, SE@k is defined as:

$$\text{SE@k} = -\sum_{v_j \in V} p(v_j) \log p(v_j), \quad p(v_j) = \frac{\psi_{v_j}}{|U|k}$$

Where $p(v_j)$ is the probability of recommending item v_j in all recommendation lists.

Gini Index(GI) GI measures inequality in the distribution of recommended items. A GI of 0 means perfect equality, all items recommended equally. While a value close to 1 indicates extreme inequality, only one item is recommended. Sort all items by ψ_{v_j} in non-decreasing order, resulting in the sorted sequence $\psi_1 \leq \psi_2 \leq \cdots \leq \psi_n$. GI@$k$ is defined as:

$$\text{GI@}k = \frac{\sum_{j=1}^{n} (2j - n - 1) \cdot \psi_j}{n \sum_{j=1}^{n} \psi_j}$$

Tail Percentage (TP) TP measures the proportion of recommended items that belong to the long-tail set $T \subset V$. A higher TP indicates a greater proportion of long-tail items in the recommendation lists. TP@k is defined as:

$$\text{TP@k} = \frac{1}{|U|} \sum_{u_i \in U} \frac{\sum_{v_j \in l^f_{u_i}} I\{v_j \in T\}}{k}$$

References

[1] J. Jacoby, "Perspectives on information overload," *Journal of Consumer Research*, vol. 10, no. 4, p. 432, 3 1984. [Online]. Available: https://doi.org/10.1086/208981

[2] F. Ricci, L. Rokach, and B. Shapira, *Recommender Systems Handbook*. Springer, 1 2022. [Online]. Available: https://doi.org/10.1007/978-1-0716-2197-4

[3] Y. Wang, W. Ma, M. Zhang, Y. Liu, and S. Ma, "A survey on the fairness of recommender systems," *ACM transactions on office information systems*, vol. 41, no. 3, pp. 1–43, 7 2022. [Online]. Available: https://doi.org/10.1145/3547333

[4] M. D. Ekstrand, A. Das, R. Burke, and F. Diaz, *Fairness in recommender systems*. Springer, 2 2012. [Online]. Available: https://doi.org/10.1007/978-1-0716-2197-4_18

[5] E. Holmes, "Anti-Discrimination rights without equality," *Modern Law Review*, vol. 68, no. 2, pp. 175–194, 2 2005. [Online]. Available: https://doi.org/10.1111/j.1468-2230.2005.00534.x

[6] Z. Liu, Y. Fang, and M. Wu, "Mitigating Popularity Bias for Users and Items with Fairness-centric Adaptive Recommendation," *ACM transactions on office information systems*, vol. 41, no. 3, pp. 1–27, 9 2022. [Online]. Available: https://doi.org/10.1145/3564286

[7] M. Mladenov, E. Creager, O. Ben-Porat, K. Swersky, R. Zemel, and C. Boutilier, "Optimizing long-term social welfare in recommender Systems: a constrained matching approach," *International Conference on Machine Learning*, vol. 1, pp. 6987–6998, 7 2020. [Online]. Available: http://proceedings.mlr.press/v119/mladenov20a/mladenov20a.pdf

[8] F. M. Harper and J. A. Konstan, "The MovieLens datasets," *ACM Transactions on Interactive Intelligent Systems*, vol. 5, no. 4, pp. 1–19, 12 2015. [Online]. Available: https://doi.org/10.1145/2827872

[9] O. Celma, *Music Recommendation and Discovery in the Long Tail*. Springer, 11 2010. [Online]. Available: https://openlibrary.org/books/OL25230501M/Music_recommendation_and_discovery

[10] Y. Li, H. Chen, S. Xu, Y. Ge, J. Tan, S. Liu, and Y. Zhang, "Fairness in Recommendation: Foundations, methods, and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 5, pp. 1–48, 7 2023. [Online]. Available: https://doi.org/10.1145/3610302

[11] M. D. Ekstrand, "Collaborative Filtering recommender systems," *Foundations and Trends® in Human–Computer Interaction*, vol. 4, no. 2, pp. 81–173, 1 2011. [Online]. Available: https://doi.org/10.1561/1100000009

[12] Y. Koren, S. Rendle, and R. Bell, *Advances in collaborative filtering*. Springer, 11 2021. [Online]. Available: https://doi.org/10.1007/978-1-0716-2197-4_3

[13] G. Adomavicius, K. Bauman, A. Tuzhilin, and M. Unger, *Context-Aware Recommender Systems: From foundations to recent developments*. Springer, 11 2021. [Online]. Available: https://doi.org/10.1007/978-1-0716-2197-4_6

[14] C. Musto, M. De Gemmis, P. Lops, F. Narducci, and G. Semeraro, *Semantics and Content-Based recommendations*. Springer, 2 2012. [Online]. Available: https://doi.org/10.1007/978-1-0716-2197-4_7

[15] R. Burke, "Hybrid Recommender Systems: survey and experiments," *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331–370, 1 2002. [Online]. Available: https://doi.org/10.1023/a:1021240730564

[16] Y. Deldjoo, D. Jannach, A. Bellogin, A. Difonzo, and D. Zanzonelli, "Fairness in recommender systems: research landscape and future directions," *User Modeling and User-Adapted Interaction*, vol. 34, no. 1, pp. 59–108, 4 2023. [Online]. Available: https://doi.org/10.1007/s11257-023-09364-z

[17] S. Verma and J. Rubin, "Fairness definitions explained," ser. FairWare '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1–7. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/3194770.3194776

[18] D. Jin, L. Wang, H. Zhang, Y. Zheng, W. Ding, F. Xia, and S. Pan, "A survey on fairness-aware recommender systems," *Information Fusion*, vol. 100, p. 101906, 7 2023. [Online]. Available: https://doi.org/10.1016/j.inffus.2023.101906

[19] M. Mansoury, B. Mobasher, and H. van Hoof, "Mitigating exposure bias in online learning to rank recommendation: A novel reward model for cascading bandits," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, ser. CIKM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 1638–1648. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/3627673.3679763

[20] Z. Wang, Q. Xu, Z. Yang, X. Cao, and Q. Huang, "Implicit Feedbacks are Not Always Favorable: Iterative Relabeled One-Class Collaborative Filtering against Noisy Interactions," *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3070–3078, 10 2021. [Online]. Available: https://doi.org/10.1145/3474085.3475446

[21] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 12 2011. [Online]. Available: https://doi.org/10.1007/s10115-011-0463-8

[22] M. D. Ekstrand, M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, and M. S. Pera, "All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 1 2018, pp. 172–186. [Online]. Available: http://proceedings.mlr.press/v81/ekstrand18b/ekstrand18b.pdf

[23] B. Rastegarpanah, K. P. Gummadi, and M. Crovella, "Fighting Fire with Fire: Using antidote data to improve polarization and fairness of recommender systems," 1 2019, pp. 231–239. [Online]. Available: https://doi.org/10.1145/3289600.3291002

[24] J. Ding, Y. Quan, X. He, Y. Li, and D. Jin, "Reinforced Negative Sampling for Recommendation with Exposure Data," *IJCAI*, pp. 2230–2236, 7 2019. [Online]. Available: https://doi.org/10.24963/ijcai.2019/309

[25] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," 8 2015, pp. 259–268. [Online]. Available: https://doi.org/10.1145/2783258.2783311

[26] Y. Yuan, X. Luo, and M.-S. Shang, "Effects of preprocessing and training biases in latent factor models for recommender systems," *Neurocomputing*, vol. 275, pp. 2019–2030, 11 2017. [Online]. Available: https://doi.org/10.1016/j.neucom.2017.10.040

[27] S. Yao and B. Huang, "Beyond Parity: Fairness Objectives for Collaborative Filtering," *Neural Information Processing Systems*, vol. 30, pp. 2921–2930, 1 2017. [Online]. Available: https://papers.nips.cc/paper/6885-beyond-parity-fairness-objectives-for-collaborative-filtering.pdf

[28] C. Wu, F. Wu, X. Wang, Y. Huang, and X. Xie, "Fairness-aware News Recommendation with Decomposed Adversarial Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 5, pp. 4462–4469, 5 2021. [Online]. Available: https://doi.org/10.1609/aaai.v35i5.16573

[29] W. Liu, F. Liu, R. Tang, B. Liao, G. Chen, and P. A. Heng, "Balancing between accuracy and fairness for interactive recommendation with reinforcement learning," in *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 155–167. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1007/978-3-030-47426-3_13

[30] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates, "Fa* ir: A fair top-k ranking algorithm," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017, pp. 1569–1578.

[31] H. Steck, "Calibrated recommendations," in *Proceedings of the 12th ACM Conference on Recommender Systems*, ser. RecSys '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 154–162. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/3240323.3240372

[32] A. J. Biega, K. P. Gummadi, and G. Weikum, "Equity of attention: Amortizing individual fairness in rankings," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, ser. SIGIR '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 405–414. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/3209978.3210063

[33] H. Guo, "Fairness testing for recommender systems," in *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*, ser. ISSTA 2023. New York, NY, USA: Association for Computing Machinery, 2023, p. 1546–1548. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/3597926.3605235

[34] Y. Wu, J. Cao, and G. Xu, "Fairness in Recommender Systems: evaluation approaches and assurance strategies," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 1, pp. 1–37, 6 2023. [Online]. Available: https://doi.org/10.1145/3604558

[35] L. Xu, Z. Tian, G. Zhang, J. Zhang, L. Wang, B. Zheng, Y. Li, J. Tang, Z. Zhang, Y. Hou, X. Pan, W. X. Zhao, X. Chen, and J. Wen, "Towards a more user-friendly and easy-to-use benchmark library for recommender systems," in *SIGIR*. ACM, 2023, pp. 2837–2847.

[36] Y. Li, H. Chen, Z. Fu, Y. Ge, and Y. Zhang, "User-oriented fairness in recommendation," in *Proceedings of the Web Conference 2021*, ser. WWW '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 624–632. [Online]. Available: https://doi.org/10.1145/3442381.3449866